Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

INGEGNERIA BIOMEDICA, ELETTRICA E DEI SISTEMI

Ciclo 35

Settore Concorsuale: 09/G1 - AUTOMATICA

Settore Scientifico Disciplinare: ING-INF/04 - AUTOMATICA

DISTRIBUTED OPTIMIZATION AND GAMES OVER NETWORKS: A SYSTEM THEORETICAL PERSPECTIVE

Presentata da: Guido Carnevale

Coordinatore Dottorato

Prof. Michele Monaci

Supervisore

Prof. Giuseppe Notarstefano

Co-supervisore

Prof. Lorenzo Marconi

Esame finale anno 2023

Abstract

Several decision and control tasks involve networks of cyber-physical systems that need to be coordinated and controlled according to a fully-distributed paradigm involving only local communications without any central unit. This thesis focuses on distributed optimization and games over networks from a system theoretical perspective. In the addressed frameworks, we consider agents communicating only with neighbors and running distributed algorithms with optimization-oriented goals. The distinctive feature of this thesis is to interpret these algorithms as dynamical systems and, thus, to resort to powerful system theoretical tools for both their analysis and design. We first address the so-called *consensus optimization* setup. In this context, we provide an original system theoretical analysis of the well-known *Gradient Tracking* algorithm in the general case of nonconvex objective functions. Then, inspired by this method, we provide and study a series of extensions to improve the performance and to deal with more challenging settings like, e.g., the derivative-free framework or the online one. Subsequently, we tackle the recently emerged framework named *distributed aggregative optimization*. For this setup, we develop and analyze novel schemes to handle (i) online instances of the problem, (ii) "personalized" optimization frameworks, and (iii) feedback optimization settings. Finally, we adopt a system theoretical approach to address *aggregative games* over networks both in the presence or absence of linear coupling constraints among the decision variables of the players. In this context, we design and inspect novel fully-distributed algorithms, based on tracking mechanisms, that outperform state-of-the-art methods in finding the Nash equilibrium of the game.

Keywords: Distributed Optimization, Consensus Optimization, Online Optimization, Derivative-Free Optimization, Distributed Aggregative Optimization, Personalized Optimization, Distributed Feedback Optimization, Aggregative Games, Distributed Equilibrium Seeking, Singular Perturbations.

Acknowledgment

This thesis work was supported by the projects OPT4SMART n. 638992 funded by the European Research Council and "Distributed Optimization for Cooperative Machine Learning in Complex Networks" n. PGR10067 funded by Ministero degli Affari Esteri e della Cooperazione Internazionale.

Contents

Introduction

Motivation and Challenges

In recent years, the increasing amount of interconnected and embedded engineering systems poses significant challenges for their development [157]. Indeed, several domains have seen an impressive increase in the employment of devices with communication and computation capabilities [101] such as intelligent transportation systems [215], and autonomous mobile robots [4]. For several application tasks, they are organized and controlled as interconnected, complex systems that interact with each other to realize the full potential of the aggregated knowledge and computational power of the entire network. Popular examples can be found in the context of smart grids [49], smart cities [71], or Industry 5.0 [121]. In all these domains, controlling these networks through a *centralized* entity would require a central node that knows some global information, takes all the decisions, and communicates them to all the single entities (or *agents*) that belong to the network. Such an aspect may be undesirable because, e.g., it may represent a barrier to scalability [198], privacy concerns [18], or the development of efficient and flexible swarming systems [44]. For this reason, it is dramatically increasing the attention for approaches relying on the so-called *distributed* paradigm. These approaches take advantage of inter-agent peer-to-peer communication protocols to spread critical information across the entire system and, thus, control the whole network without any centralized unit. In this context, a relevant goal is to formulate a mathematical framework to take decisions that optimize a given performance metric. The network agents can cooperate with the aim of optimizing a common performance index, giving rise to *distributed optimization* scenarios, or can compete with each other with the aim of optimizing individual objective functions, leading to the context of *games* over networks. Indeed, optimization can be used to formalize several challenges stemming from different domains such as, e.g., estimation in sensor networks, cooperative model predictive control, learning in machine learning applications, control of energy networks, and control of networks of mobile robots. Specifically, as for distributed optimization, we focus on two different scenarios, namely *distributed consensus optimization* and *distributed aggregative optimization*. Instead, as regards games, we concentrate on

network aggregative games. Along this thesis, according to the distributed paradigm introduced above, the key assumption is that each agent works with *partial* information about the problem. In particular, each agent is only aware of its own local information (e.g., an associated objective function) and can exchange data according to a sparse communication graph, i.e., only with a subset of the entire set of agents.

The expression distributed consensus (or cost-coupled) optimization refers to optimization problems over networks where the cost is the sum of local functions depending on a common decision variable. Typical tasks that can be posed as consensus optimization problems can be found, e.g., in the context of robust estimation in statistics, support vector machine in machine learning, and signal processing. We refer the reader to the recent surveys [73, 135, 142, 202] for a comprehensive overview of the possible applications of this setting and the existing methods to address it. The complex features of consensus optimization problems arising in, e.g., data analytics and deep learning motivated our interest in addressing (i) nonconvex objective functions, (ii) asynchronous communication, (iii) derivative-free scenarios, and (iv) time-varying settings.

Distributed aggregative optimization, on the other hand, refers to a recently emerged framework in which agents in a network aim to minimize the sum of local objective functions which depend both on local decision variables and a common *aggregative* variable that couples the decisions of all the agents. This challenging setting has been introduced by the pioneering work [104] and suitably models many tasks in the context of cooperative robotics such as, e.g., multi-robot surveillance scenarios.

Instead, in the context of network aggregative games, we still have objective functions both depending on local and aggregative variables but each agent aims at minimizing only its own cost. Many tasks arising in several domains such as smart grids management, economic market analysis, cooperative control of robots, electric vehicles charging, network congestion control, can be formulated as aggregative games. See, e.g., the surveys [12, 87, 152] for a detailed examination. In this context, the goal is to design distributed equilibrium seeking algorithms, i.e., fully-distributed methods able to find a so-called Generalized Nash Equilibrium (GNE) of the considered game.

Recent years have also seen a growing interest in the exploitation of concepts and ideas from system theory in the context of optimization. Such an interest is due to the wide set of well-established tools and concepts provided by system theory that can be leveraged to analyze and design effective optimization methods. Motivated by this, throughout the whole thesis, we take a system theoretical perspective in which our distributed algorithms are interpreted as dynamical systems.

Summary of the Contributions

This thesis contributes to the fields of distributed consensus optimization, distributed aggregative optimization, network aggregative games. In detail, we contribute to all these frameworks by proposing novel schemes under different problem settings and assumptions. A distinctive feature of the thesis is the exploitation of a system theoretical framework both for the analysis and design. The interpretation of the algorithms as dynamical systems allows for a deeper understanding of existing schemes and for the design of novel ones. We resort to Lypaunov-based tools as, e.g., LaSalle's invariance principle, time-scale separation, and averaging theory, to understand and recognize key properties of the algorithms and, thus, to enhance their capabilities. Moreover, a system theoretical perspective naturally leads to a unified description and design of distributed algorithms in different frameworks.

As for distributed consensus optimization, we take into account different problem settings and, in this context, we study and extend the existing Gradient Tracking algorithm. The latter is a popular distributed optimization algorithm for consensus optimization whose main feature consists of combining the gradient descent idea with the so-called trackers, i.e., a set of auxiliary variables that locally compensate for the lack of knowledge of the agents about the gradient of the global objective function. First, we consider the case of nonconvex objective function and analyze the convergence properties of the Gradient Tracking scheme through an elegant system theoretical dissertation. Afterward, we restrict to quadratic programs and propose a control-oriented design to enhance the convergence properties of Gradient Tracking. Subsequently, we derive Continuous Gradient Tracking, i.e., the continuous-time counterpart of Gradient Tracking. Moreover, since such a scheme requires continuous-time communication among the network agents, we also derive two extensions implementing synchronous and asynchronous discrete-time communication, respectively. Then, we tackle a derivative-free scenario, namely the case in which the agents cannot access the gradients of their associated cost functions. To overcome this lack of knowledge, we modify a forward Euler discretization of Continuous Gradient Tracking by taking advantage of a gradient estimation technique based on extremum-seeking concepts. Finally, we consider an online setting, i.e., a scenario in which the objective functions vary over time. In this framework, we propose a novel distributed scheme obtained by combining Gradient Tracking with Adam, i.e., a popular centralized algorithm for stochastic optimization. The effectiveness of the considered algorithms is tested with numerical simulations about typical problems arising in data-analytics and source estimation.

As for the distributed aggregative framework, we start by proposing a novel distributed scheme for constrained online instances of the problem which improves the convergence properties of similar existing schemes. Then, we extend this algorithm

to address the so-called *personalized* framework, i.e., the one in which each objective function has an unknown part that can be accessed by the related agent only in terms of noisy user feedbacks. Specifically, we interlace the standard algorithm with a Recursive Least Squares (RLS) scheme devoted to estimating the unknown part of the cost using the collected user feedbacks. Further, we design a distributed feedback optimization law for the aggregative framework. In particular, we consider a set of systems with continuous-time nonlinear dynamics and aim to design a distributed control law able to steer the network to a steady-state configuration corresponding to a stationary point of an associated aggregative optimization problem with nonconvex cost functions. Notably, in this setting, the agents do not know the objective functions of the problem and can only measure the gradients evaluated in their current configuration. The theoretical results of the chapter are corroborated by numerical simulations involving tasks arising in multi-robot scenarios and opinion dynamics.

As for network aggregative games, the contribution consists of novel distributed equilibrium seeking algorithms for two scenarios, i.e., the one with only local constraints and the one in which also linear coupling constraints are present. The first scenario is tackled through a projected pseudo-gradient scheme combined with a tracking mechanism due to reconstruct the unavailable aggregative variable. As for the second scenario, we adapt a recent augmented primal-dual method for this setting and combine it with average consensus and tracking techniques. Both theoretical and numerical results are provided to show that our schemes outperform the other state-of-the-art distributed methods in terms of convergence rate.

Finally, this thesis contributes with two novel results about the so-called discrete-time singularly perturbed systems. These results extend the existing ones with (i) convergence in a LaSalle sense, and (ii) global exponential convergence. Although we explicitly use them to prove the convergence features of some of the algorithms described above, they represent per se results that can be useful for the analysis of generic dynamical systems.

Organization and Chapter Contributions

The thesis organization follows the contribution scheme outlined in the previous section. We first provide a chapter to formalize the three frameworks addressed in the thesis. Then, for each one of these frameworks, we provide a related chapter containing both the theoretical findings and the numerical simulations that confirm them.

In Chapter 1, we formalize the distributed consensus optimization framework, the distributed aggregative optimization scenario, and the network aggregative games setting.

In Chapter 2, we deeply investigate and extend the existing Gradient Tracking al-

gorithm. First, we provide a clean, elegant system theoretical perspective based on a LaSalle and *singular perturbations* analysis to study the Gradient Tracking convergence properties in the case of nonconvex objective functions. In particular, we perform suitable changes of variables to reformulate the Gradient Tracking algorithm as a singularly perturbed system, namely the interconnection between a slow subsystem, which mimics the gradient descent method, and a fast one describing the dynamics of the consensus error among both the solution estimates and the trackers. We separately study the system theoretical properties of two auxiliary schemes associated to these subsystems and, then, by merging these results through a specific theorem, we assess the asymptotic convergence of the whole interconnection to the set of stationary points of the problem. Afterward, we demonstrate the effectiveness of the system theoretical tools by designing modified Gradient Tracking schemes with sparse matrix gains (rather than the diagonal ones of the standard scheme) for quadratic programs. We numerically show the enhancement in terms of convergence rates by comparing the two versions of the algorithm. Subsequently, by taking inspiration from a branch of research studying the continuous-time counterpart of existing discrete-time optimization algorithms, we develop the continuous-time version of the Gradient Tracking algorithm. Moreover, to avoid the unrealistic implementation of continuous-time inter-agent communication, we also design two additional extensions leveraging synchronous and asynchronous discrete-time communication, respectively. In particular, the asynchronous scheme relies on local triggering conditions that the agents independently check to choose the instants of time in which their own variables must be sent to their neighbors. For all the obtained schemes, we take advantage of a Lyapunov-based analysis to show the exponential convergence to the solution of strongly convex problems. We also show that the asynchronous scheme avoids the so-called Zeno effect, i.e., an infinite number of communications in a finite interval of time. Then, we consider the case in which the agents cannot access the gradients of the objective functions. To overcome this issue, we propose Extremum Tracking Descent, i.e., a novel distributed algorithm obtained by modifying a forward Euler discretization of the Continuous Gradient Tracking policy by replacing the unavailable gradients through a suitable mechanism based on an extremum-seeking technique. The obtained scheme is analyzed by resorting to tools from discrete-time averaging theory. Specifically, we study the average system associated to the original one. For this system, by means of a Lyapunov-based analysis, we find an arbitrarilry small set with semi-global practical stability guarantees. Then, with a suitable choice of the algorithm parameters, we impose the closeness betweeen the original and the average scheme. In this way, we guarantee that, in the case of strongly convex costs, the obtained scheme asymptotically converges to an arbitrarily small neighborhood of the problem solution. Notably, we show that the accuracy can be arbitrarily improved through the amplitude of the so-called dither signals used to

estimate the gradients. Finally, we take into consideration the *online* case, i.e., the one in which the objective functions vary over time. In this setting, we propose GTAdam, i.e., a novel distributed algorithm obtained by combining the Gradient Tracking algorithm with Adam, namely a popular centralized method for stochastic optimization. In detail, inspired by Adam, GTAdam computes the descent direction of each agent through the estimates of the first- and second-order momenta of the trackers. In the case of strongly convex problems, we theoretically (i) provide an upper bound for the dynamic regret achieved by GTAdam and (ii) prove its linear convergence to the optimal solution for the static case. By performing detailed numerical simulations, we show that GTAdam outperforms existing state-of-the-art distributed methods. The results of this chapter are based on [25, 28, 31, 33, 130].

In Chapter 3, we investigate the distributed aggregative optimization framework. In this context, we consider a constrained online version of the problem where cost functions, aggregation rules, and constraints vary over time. To address such a problem, we propose Projected Aggregative Tracking, namely a novel distributed algorithm for constrained online aggregative optimization. Projected Aggregative Tracking combines in each agent a distributed implementation of the gradient descent with two tracking mechanisms devoted to reconstructing, in a distributed manner, both the aggregative variable and the related gradient of the global objective function. Further, we include a convex combination step which turns out to be crucial in improving the existing theoretical results. Indeed, we demonstrate that Projected Aggregative Tracking (i) achieves a dynamic regret with an improved upper bound with respect to the existing one, and (ii) linearly converges to the optimal solution in case of static problems. Then, we consider a personalized scenario, i.e., a setting in which the considered optimization tasks directly involve human end-users. As a consequence, users' dissatisfaction needs to be taken into account together with engineering-oriented goals but, due to the complexity and subjectivity of human preferences, *personalization* may suit better than the exploitation of synthetic models. For this reason, we address this framework assuming that part of each local objective function is unknown and can be accessed only through noisy user feedback of the cost. To overcome this lack of knowledge, we equip each agent with an RLS scheme that, interlaced with Projected Aggregative Tracking, leads to a novel distributed algorithm named RLS Projected Aggregative Tracking. Starting from the existing analysis, we theoretically provide an upper bound for the dynamic regret achieved by RLS Projected Aggregative Tracking. Subsequently, we investigate a feedback optimization setting for the aggregative framework. In this context the challenge is twofold: (i) we investigate the feedback optimization paradigm for nonlinear systems in a distributed framework, and (ii) we consider the aggregative optimization set-up in a nonconvex scenario. Specifically, we propose Aggregative Tracking Feedback, i.e., a novel continuous-time distributed feedback optimization law for aggregative opti-

mization problems. The aim is to steer, in a fully-distributed manner, a network of dynamic agents to a steady-state configuration which is a stationary point of a given aggregative optimization problem with (possibly) nonconvex objective function. Aggregative Tracking Feedback implements a two-step procedure: (i) moves the network along an estimated descent direction of the cost, and (ii) reconstructs in each agent the global information needed for step (i). Step (i) is performed through a distributed implementation of a closed-loop gradient flow. As per step (ii), a consensus-based dynamics is implemented in which two auxiliary states asymptotically compensate for the mismatches between the part of information locally available and the global one. It is worth highlighting that Aggregative Tracking Feedback is a distributed feedback strategy handling at the same time-scale the control and the optimization of a network of nonlinear systems. By resorting to tools from system theory, we guarantee the asymptotic convergence of the network to a steady-state configuration being a stationary point of the optimization problem. Further, we take into account the case with single integrator dynamics and strongly convex objective function. In this case, we adapt Aggregative Tracking Feedback to get a closed loop system that exponentially converges to a configuration corresponding to the optimal solution of the problem. The results of this chapter are based on [26, 29, 30, 32].

Finally, in Chapter 4, we design novel distributed equilibrium seeking algorithms for aggregative games with both local and linear coupling constraints. First, we tackle the case with only local constraints by proposing a scheme that combines a projected pseudo-gradient method with a tracking mechanism. Then, to deal also with coupling constraints, we take inspiration by an existing continuous-time augmented primal-dual scheme for centralized optimization. In detail, we combine a distributed augmented primal-dual scheme with (i) an average consensus step to force agreement among the agents' multipliers and (ii) a tracking-based mechanism to reconstruct in each agent the aggregative variable and the coupling constraint. Both the proposed schemes are analyzed trough the following system theoretical strategy. We reformulate the original scheme as a singularly perturbed system, i.e., as the interconnection between a slow subsystem and a fast one. Then, we provide two Lyapunov functions for two distinct auxiliary schemes related to these two subsystems. Subsequently, by properly merging these two results, we show that our methods linearly converge to the GNE of the problem. Both the theoretical guarantees and the numerical results show that our methods outperform the state-of-the-art distributed algorithms in terms of convergence rate. The results of this chapter are based on [27].

As a complementary part of the thesis, we include Appendix A providing some basic concepts on optimization, Appendix B reporting some auxiliary results that turn out to be useful to prove some intermediate results of the thesis, and Appendix C containing our novel, custom results about singularly perturbed systems.

Notation

The symbols $\mathbb{R}$ and $\mathbb{N}$ denote the set of real and natural numbers, respectively. A matrix $M \in \mathbb{R}^{n\times n}$ is Schur if all its eigenvalues lie in the open unit disc, while is Hurwitz if all its eigenvalues have negative real part. The identity matrix in $\mathbb{R}^{m\times m}$ is I_m, while 0_m is the all-zero matrix in $\mathbb{R}^{m\times m}$. The vector of N ones is denoted by 1_N, while $\mathbf{1}_{N,d} \coloneqq 1_N \otimes I_d$ with $\otimes$ being the Kronecker product. Dimensions are omitted whenever clear from the context. Given two vectors $v_1, v_2 \in \mathbb{R}^n$, their Hadamard product is denoted as $v_1 \odot v_2$. Given a function of two variables $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$, we denote as $\nabla_1 f \in \mathbb{R}^{n_1}$ the gradient of f with respect to its first argument and as $\nabla_2 f \in \mathbb{R}^{n_2}$ the gradient of f with respect to the second one. The vertical concatenation of column vectors $v_1\mathbb{R}^{n_1}, \dots, v_N \in \mathbb{R}^{n_N}$ is $\textsc{col}(v_1, \dots, v_N) \in \mathbb{R}^{\sum_{i=1}^N n_i}$. $\mathbb{R}^n_+$ identifies the positive orthant in $\mathbb{R}^n$. We denote with $\mathrm{diag}(v_1, \dots, v_n)$ the diagonal matrix whose i-th diagonal element is given by v_i, and with $\mathrm{blkdiag}(M_1, \dots, M_N)$ the block diagonal matrix whose i-th block is $M_i \in \mathbb{R}^{n_i\times n_i}$. Given a vector $v \in \mathbb{R}^n$ and a set $\S \subseteq \mathbb{R}^n$, $P_S x$ denotes the projection of v on S, i.e., $P_S[v] \coloneqq \arg\min x \in S \, \|v - x\|$, while we use $\mathrm{dist}(v, S)$ to denote its distance from the set, namely $\mathrm{dist}(v, S) = \min_{x\in S} \|v - x\|$. For a finite set S, we denote by $|S|$ its cardinality. Given $v \in \mathbb{R}^n$, we use $[v]^+$ to denote $\max\{0, v\}$ in a component-wise sense. Given a square matrix $M \in \mathbb{R}^{n\times n}$, a set $S \subset \mathbb{R}^n$ is said to be M-invariant if for all $v \in S$ it holds $Mv \in S$. Given $v \in \mathbb{R}^n$ and a symmetric, positive definite matrix $M \in \mathbb{R}^{n\times n}$, $\|v\|_M = \sqrt{v^\top M v}$. As for the euclidean norm, we omit the subscript, namely $\|v\| = \sqrt{v^\top v}$. Let $M \in \mathbb{R}^{n\times n}$, then we denote as $\rho_{\max}(M)$ its spectral radius. Given a vector v and a matrix M, we denote as $[v]_j$ and $[M]_j$ the j-th component of v and the j-th row of M, respectively. Given $c \in \mathbb{R}$, $b \in \mathbb{R}^n$, and $M \in \mathbb{R}^{n\times n}$, let $v \coloneqq \textsc{col}(c, b, [M]_1, \dots, [M]_n) \in \mathbb{R}^{1+n+n^2}$, then we define the operator $\textsc{unpack}(v)$ so that $(M, b, c) = \textsc{unpack}(v)$. Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, we define $\ker\{f(\cdot)\} \coloneqq \{x \in \mathbb{R}^n \mid f(x) = 0\}$. Given $c > 0$, we use the symbol $\mathcal{B}_c$ to denote the sphere with radius c, namely $\mathcal{B}_c \coloneqq \{v \in \mathbb{R}^n | \, \|v\| \leq c\}$.

Chapter 1

Distributed Optimization and Aggregative Games over Networks

In this chapter, we introduce the frameworks addressed in this thesis. We first review some basic concepts of graph theory and introduce the distributed computation model. Then, we formalize the three frameworks studied throughout the thesis and, for each of them, we present a practical application.

1.1 Graph Theory and Distributed Communication Model

In a distributed context, there are N units, called *agents* or *players*, that have both communication and computation capabilities. We assume that the agents can exchange information with each other by sending and receiving packets of information.

This communication model is formalized by resorting to graph theory. Formally, we define a graph $\mathcal{G}$ as the ordered pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, N\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. We call nodes (or vertices) the elements in $\mathcal{V}$, while we call edges the ones in $\mathcal{E}$ which are of the type (i, j) with $i, j \in \mathcal{V}$. The nodes represent the set of agents of the network, while the edges represent the communication links among them. If the edges are not oriented, i.e., $(i, j) \in \mathcal{E}$ iff $(j, i) \in \mathcal{E}$, we say that the graph is *undirected*, otherwise we call it *directed*. An example of a directed and of an undirected graph is depicted in Figure 1.1.

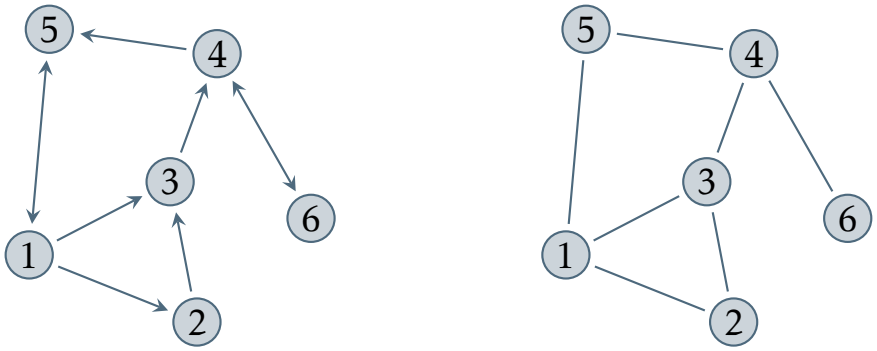


Figure 1.1: A directed (left) and an undirected (right) graph of $N = 6$ nodes.

Further, we also associate to the graph $\mathcal{G}$ a *weighted adjacency matrix* $\mathcal{W}_{\mathcal{G}} \in \mathbb{R}^{N\times N}$ matching the graph sparsity in the sense that its (i,j)-th entry $w_{ij} > 0$ if $(j,i) \in \mathcal{E}$ and 0 otherwise. Given an edge $(i,j) \in \mathcal{E}$, i is called *in-neighbor* of j and j is an *out-neighbor* of i. For each agent i, we define the in-neighbor set as $\mathcal{N}_i = \{j \in \mathcal{V} : (j,i) \in \mathcal{E}\}$ and the out-neighbor set as $\mathcal{N}_i^{\text{out}} = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$. If the graph is undirected, we simply say that i is a *neighbor* of j (and viceversa), and denote $\mathcal{N}_i$ (wich coincides with $\mathcal{N}_i^{\text{out}}$) as the *neighbor* set. Analogously, we define the weighted in-degree of the agent i as $d_i^{\text{in}} := \sum_{j\in\mathcal{N}_i} w_{ij}$ and its out-degree as $d_i^{\text{out}} := \sum_{j\in\mathcal{N}_i^{\text{out}}} w_{ji}$. If it holds $d_i^{\text{in}} = d_i^{\text{out}}$ for all $i \in \{1,\dots,N\}$, we say that the graph is *weight-balanced*. In this connection, we define the in-degree matrix $\mathcal{D}^{\text{in}} \in \mathbb{R}^{N\times N}$ and the out-degree matrix $\mathcal{D}^{\text{out}} \in \mathbb{R}^{N\times N}$ that are the diagonal matrices whose i-th blocks are given by d_i^{in} and d_i^{out}, respectively. Also, we define the *Laplacian matrix* $\mathcal{L} \in \mathbb{R}^{N\times N}$, where each entry (i,j) is equal to

$$L_{ij} = \begin{cases} d_i^{\text{in}} & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j \end{cases}. \tag{1.1}$$

Thus, it can be equivalently written $\mathcal{L} = \mathcal{D}^{\text{in}} - \mathcal{W}_{\mathcal{G}}$.

In a distributed computation context, each agent in the network is associated to a fixed identifier $i \in \mathcal{V}$ from the set of nodes, while, if the edge (i,j) belongs to the communication graph, then agent i can send information to agent j. In this context, each agent typically maintains some local states that change according to the specific algorithm in execution. In particular, if each agent uses only information received by its in-neighbors, then we refer the algorithm as distributed. As one may expect, the connectivity among the agents plays a key role in determining the effectiveness of the algorithm. In detail, most distributed algorithms need to be executed over strongly connected graphs which are the ones satisfying the following connectivity property.

Definition 1.1 (Connectivity [21])**.** *A directed graph $\mathcal{G}$ is said to be* strongly connected *if for every pair of nodes (i,j) there exists a path of directed edges that goes from i to j. If $\mathcal{G}$ is undirected, we say that $\mathcal{G}$ is* connected. △

In this setting, it is customary to employ suitable weighted adjacency matrices matching the following definition.

Definition 1.2 (Stochastic matrices [21])**.** *The matrix $A \in \mathbb{R}^{N\times N}$ is said to be* row stochastic *if it holds*

$$A1_N = 1_N.$$

Analogously, A is said to be column stochastic *if it holds*

$$1_N^\top A = 1_N^\top.$$

If A is both row and column stochastic, then it is said to be doubly stochastic. △

1.2 Consensus Optimization

In this section, we formalize the distributed consensus optimization setup and present a related application example. This framework is widely investigated in Chapter 2 with different problem settings and assumptions.

1.2.1 Problem Description

We deal with a network of $N \in \mathbb{N}$ agents that must solve a *consensus* (or *cost-coupled*) optimization problem, which can be stated as

$$\min_{x\in X} \sum_{i=1}^{N} f_i(x), \tag{1.2}$$

where $x \in \mathbb{R}^n$ is the (common) decision variable of the objective functions $f_i : \mathbb{R}^n \to \mathbb{R}$, and $X \subseteq \mathbb{R}^n$ is the feasible set. In our distributed context, each agent $i \in \{1, \dots, N\}$ is only aware of its own associated function f_i and maintains a local estimate about the solution of problem (1.2). In this connection, we term such a problem as "consensus optimization" because the agents aim to asymptotically reach a consensus among their estimates in a point that coincides with a solution of (1.2). Indeed, if the decision variable had not been common, then the problem would split into N independent problems of the form

$$\min_{x_i\in X} f_i(x_i), \quad i \in \{1, \dots, N\}.$$

1.2.2 Application Example: Classification using Logistic Regression

Consider a network of agents that want to cooperatively train a linear classifier for a set of points in a given feature space. Each agent i is equipped with $m_i \in \mathbb{N}$ points $p_{i,1}, \dots, p_{i,m_i} \in \mathbb{R}^{d-1}$ with binary labels $l_{i,q} \in \{-1, 1\}$ for all $q \in \{1, \dots, m_i\}$. The problem consists of building a linear classification model from the given points, also called training samples. In particular, we look for a pair $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ defining a separating hyperplane described by $\{p \in \mathbb{R}^{n-1} \mid w^\top p + b = 0\}$. The aim is to find the ideal hyperplane $(w_\star, b_\star)$ that separates all points with $l_{i,q} = -1$ from all the ones with $l_{i,q} = 1$, namely to find $(w^\star, b_\star)$ so that

$$\begin{aligned} w_\star^\top p_{i,q} + b_\star \geq 0 \quad \forall (i,q) \quad \text{such that} \quad l_{i,q} = 1 \\ w_\star^\top p_{i,q} + b_\star < 0 \quad \forall (i,q) \quad \text{such that} \quad l_{i,q} = -1. \end{aligned}$$

Figure 1.2 depicts the scenario described above. We use different colors to denote the points belonging to different agents, while we distinguish among the two possible class of points by using circles and triangles to label them. Figure 1.2 clearly highlights that a single agent may have only points belonging to the same class and, therefore, would have no way to build classifier without cooperation with other agents in the network.

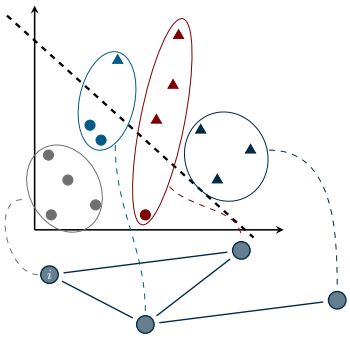


Figure 1.2: Illustration of the classification in a network of 4 agents. Each agent has private points with two possible labels, i.e., circles or triangles. The black, dotted line perfectly separates the two classes of points.

The classification problem can be posed as a minimization problem described by

$$\min_{w \in \mathbb{R}^{n-1}, b \in \mathbb{R}} \sum_{i=1}^{N} \sum_{q=1}^{m_i} \log\left(1 + e^{-l_{i,q}(w^\top p_{i,q} + b)}\right) + \frac{C}{2}\left(\|w\|^2 + b^2\right), \tag{1.3}$$

where $C > 0$ is the so-called regularization parameter. The optimization problem formalized in (1.3) is an unconstrained instance of (1.2), i.e., $X \equiv \mathbb{R}^n$. Indeed, the (common) decision variable is $x = \text{col}(w, b)$, while the local objective function of agent

i is given by

$$f_i(\text{col}(w,b)) = \sum_{q=1}^{m_i} \log\left(1 + e^{-l_{i,q}(w^\top p_{i,q}+b)}\right) + \frac{C}{2N}\left(\|w\|^2 + b^2\right).$$

1.3 Distributed Aggregative Optimization

This section is devoted to formalize the distributed aggregative optimization setup and present a related application example. In Chapter 3, we address this kind of problems and related variants.

1.3.1 Problem Description

We consider an optimization problem written in the form

$$\min_{(x_1,\dots,x_N)\in X} \sum_{i=1}^{N} f_i(x_i, \sigma(x)), \tag{1.4}$$

in which $x := \text{col}(x_1,\dots,x_N) \in \mathbb{R}^n$ is the global decision vector, with each $x_i \in \mathbb{R}^{n_i}$ and $n = \sum_{i=1}^N n_i$. The global decision vector is constrained to belong to a set $X \subseteq \mathbb{R}^n$ that can be written as $X = \prod_{i=1}^N X_i$, where each $X_i \subseteq \mathbb{R}^{n_i}$. The functions $f_i : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}$ represent the local objective functions, while the *aggregative* variable $\sigma(x)$ has the form

$$\sigma(x) := \frac{\sum_{i=1}^N \phi_i(x_i)}{N}, \tag{1.5}$$

where each $\phi : \mathbb{R}^{n_i} \to \mathbb{R}^d$ is the i-th contribution to the aggregative variable. We compactly denote the cost function of problem (1.4) through $f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ defined as $f(x, \sigma(x)) := \sum_{i=1}^N f_i(x_i, \sigma(x))$. Each agent of the network has only partial information about problem (1.4). In particular, each agent $i \in \{1,\dots,N\}$ can only privately access f_i, X_i, and ϕ_i and, thus, needs to exchange information with the other agents of the network to find the i-th component of an optimal solution of problem (1.4). Here, the coupling among the agents is due to the fact that each objective function f_i depends on the aggregative variable $\sigma(x)$.

1.3.2 Application Example: Multi-Robot Surveillance

Consider a network of N mobile robots, whose position is $x_i \in \mathbb{R}^2$, that aim to protect a common target, located at $b \in \mathbb{R}^2$, from a collection of N intruders, see Figure 1.3. In particular, each robot i of the surveillance team is associated to an intruder located at $p_i \in \mathbb{R}^2$. Therefore, the protection strategy of the team consists of a trade-off between two

competing objectives: (i) each robot tries to stay close to the intruder, (ii) the surveillance team tries to keep its barycenter close to the target.

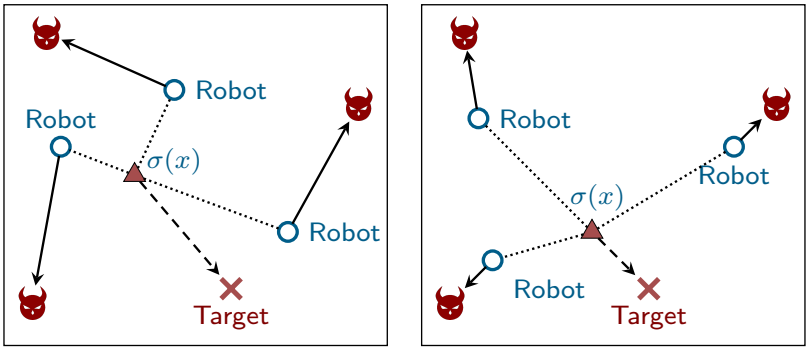
The scenario described above can be suitably captured by means of the distributed aggregative optimization framework described in the previous section. Specifically, in problem (1.4) we can set

$$f_i(x_i, \sigma(x)) = \frac{1}{2}\|x_i - p_i\|^2 + \frac{\gamma}{2N}\|\sigma(x) - b\|^2 \tag{1.6}$$

with $\gamma > 0$, and aggregative variable denoting the weighted center of mass of the defending team

$$\sigma(x) = \frac{1}{N}\sum_{i=1}^{N} \beta_i x_i, \tag{1.7}$$

for some weights $\beta_i > 0$. We notice that if $\beta_i = 1$ for all $i \in \{1, \dots, N\}$, then the standard center of mass is recovered. An illustrative concept of this framework is provided in Figure 1.3, where an initial configuration and the (unique) optimal one are shown.



(a) Initial configuration. (b) Optimal configuration.

Figure 1.3: Multi-robot surveillance scenario.

1.4 Distributed Equilibrium Seeking in Aggregative Games

In this section, we introduce aggregative games over networks and present a related application example. As we will see in the sequel, this framework has many similarities with the distributed aggregative optimization setup presented in Section 1.3. However, the key difference is that, here, the goal is to compute a (generalized) Nash equilibrium rather than an optimal solution cooperatively. Chapter 4 is devoted to providing distributed algorithms able to find such an equilibrium.

1.4.1 Problem Description

We consider a population of $N \in \mathbb{N}$ agents who, given all other agents' strategies, aim at finding a local strategy solving the optimization problem:

$$\forall i \in \{1, \dots, N\} : \begin{cases} \min\limits_{x_i \in X_i} & J_i(x_i, \sigma(x)) \\ \text{s.t.} & \sum\limits_{j=1}^{N} A_j x_j \leq \sum\limits_{j=1}^{N} b_j, \end{cases} \tag{1.8}$$

where $X_i \subseteq \mathbb{R}^{n_i}$, $A_i \in \mathbb{R}^{m \times n_i}$, and $b_i \in \mathbb{R}^m$ model the feasible strategy set for agent i, while the cost function $J_i : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}$ depends on the i-th individual strategy $x_i \in \mathbb{R}^{n_i}$, as well as on the aggregative variable $\sigma(x) \in \mathbb{R}^d$, with $x \coloneqq \text{col}(x_1, \dots, x_N) \in \mathbb{R}^n$, $n \coloneqq \sum_{i=1}^{N} n_i$. We consider $m \leq n$. As in Section 1.3, the aggregative variable $\sigma(\cdot)$ formally reads as

$$\sigma(x) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \phi_i(x_i), \tag{1.9}$$

where each aggregation rule $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}^d$ models the contribution of the corresponding strategy x_i to the aggregate $\sigma(x)$. We define the constraint functions $c_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$, $c_{-i} : \mathbb{R}^{n-n_i} \to \mathbb{R}^m$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ as follows:

$$c_i(x_i) = A_i x_i - b_i, \tag{1.10a}$$

$$c_{-i}(x_{-i}) = \sum_{j \in \{1,\dots,N\} \setminus \{i\}} (A_j x_j - b_j), \tag{1.10b}$$

$$c(x) = c_i(x_i) + c_{-i}(x_{-i}) = Ax - b, \tag{1.10c}$$

where $x_{-i} \coloneqq \text{col}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N) \in \mathbb{R}^{n-n_i}$, $A \coloneqq [A_1 \ \dots \ A_N] \in \mathbb{R}^{m \times n}$, and $b \coloneqq \sum_{i=1}^{N} b_i$. Then, the collective vector of strategies x belongs to the feasible set $\mathcal{C} \coloneqq \{\text{x} \in X \mid c(x) \leq 0\} \subseteq \mathbb{R}^n$, where $X \coloneqq \prod_{i=1}^{N} X_i \subseteq \mathbb{R}^n$.

We refer to any equilibrium solution to the collection of inter-dependent optimization problems (1.8) as aggregative Generalized Nash Equilibrium (GNE) [59] (or simply GNE), and to the problem of finding such an equilibrium as GNE problem (GNEP) in aggregative form – as opposed to a Nahs Equilibrium problem (NEP) which is characterized by local constraints only. We will design distributed algorithms to find aggregative GNEs, which formally correspond to the following definition:

Definition 1.3 (Generalized Nash Equilibrium [59])**.** *A collective vector of strategies $x^\star \in \mathcal{C}$*

is a GNE of (4.1) *if, for all* $i \in \{1, \dots, N\}$*, we have:*

$$J_i(x_i^\star, \sigma(x^\star)) \leq \min_{x_i \in \mathcal{C}_i(x_{-i}^\star)} J_i\left(x_i, \tfrac{1}{N}\phi_i(x_i) + \sigma_{-i}(x_{-i}^\star)\right),$$

with $\mathcal{C}_i(x_{-i}) \coloneqq \{x_i \in X_i \mid A_i x_i \leq b_i - c_{-i}(x_{-i})\}$. △

We note that the definition of Nash Equilibrium (NE) follows directly from the above by replacing $\mathcal{C}_i(x_{-i}^\star)$ simply with X_i.

An equivalent definition of GNE requires one to find a fixed-point of the *best response* mapping $x_{i,\text{br}} : \mathbb{R}^{n-n_i} \to \mathbb{R}^{n_i}$ of each agent, which is formally defined as:

$$\begin{aligned} x_{i,\text{br}}(x_{-i}) &\in \operatorname*{arg\,min}_{x_i \in \mathcal{C}_i(x_{-i})} \; J_i\left(x_i, \sigma(x)\right) \\ &= \operatorname*{arg\,min}_{x_i \in \mathcal{C}_i(x_{-i})} \; J_i\left(x_i, \tfrac{1}{N}\phi_i(x_i) + \sigma_{-i}(x_{-i})\right), \end{aligned}$$

In fact, a collective vector of strategies $x^\star$ is a GNE if, for all $i \in \{1, \dots, N\}$, $\mathrm{x}_i^\star = x_{i,\text{br}}(x_{-i}^\star)$.

Also in this setting, we want to develop methods that work in a distributed fashion. Similarly to the distributed aggregative optimization setup (see Section 1.3), we assume that agent i only knows J_i, ϕ_i, X_i, A_i, and b_i. Hence, also in this setting the local lack of knowledge must be compensated by leveraging inter-agents communication.

1.4.2 Application Example: Nash-Cournot Game

In this section, we show a case study from [11] that can be formalized as an instance of problem (1.8), i.e., a Nash-Cournot game. In this connection, consider N firms that compete over n_m markets. In particular, for each market $\tau \in \mathcal{M} \coloneqq \{1, \dots, n_m\}$, firm i is characterized by a production $g_{i,\tau} \geq 0$ and sales $s_{i,\tau} \geq 0$. For each $i \in \{1, \dots, N\}$ and $\tau \in \mathcal{M}$, the cost of production amounts to

$$f_{i,\tau}(g_{i,\tau}) = q_{i,\tau} g_{i,\tau}^2 + c_{i,\tau} g_{i,\tau}.$$

The revenue of firm i at market τ is modelled as $(d_\tau - \bar{s}_\tau)s_{i,\tau}$, where $d_\tau > 0$ is the total demand for location τ, and $\bar{s}_\tau \coloneqq \sum_{i \in \{1,\dots,N\}} s_{i,\tau}$ represents the aggregate sales at location τ. For all firms $i \in \{1, \dots, N\}$ and markets $\tau \in \mathcal{M}$, we assume a production limitation $u_{i,\tau}$. Moreover, in each market τ, the total production $\sum_{i \in \{1,\dots,N\}} g_{i,\tau}$ must cover the demand d_τ without exceeding a maximum capacity r_q. We can thus cast this setting as an instance of the GNEP in (1.8) with each strategy vector given by $\mathrm{x}_i \coloneqq \textsc{col}(g_{i,1}, \dots, g_{i,n_m}, s_{i,1}, \dots, s_{i,n_m}) \in \mathbb{R}^{2n_m}$, and cost function

$$J_i(x_i, \sigma(x)) = x_i^\top Q_i x_i + \ell_i^\top x_i + (\Delta\sigma(x))^\top x_i,$$

where we introduce the symbols $Q_i \coloneqq \mathrm{diag}(q_{i,1}, \dots, q_{i,n_m}, 0, \dots, 0) \in \mathbb{R}^{2n_m \times 2n_m}$, $\ell_i \coloneqq \mathrm{col}(c_{i,1}, \dots, c_{i,n_m}, -d_1, \dots, -d_{n_m}) \in \mathbb{R}^{2n_m}$, $\Delta = \mathrm{blkdiag}(0_{n_m}, NI_{n_m})$, and set the aggregation rule as $\phi_i(x_i) = x_i$ for all $i \in \{1, \dots, N\}$. As for the constraints, for all $i \in \{1, \dots, N\}$, we have the local constraint set

$$X_i \coloneqq \left\{ x_i \in \mathbb{R}^{2n_m} \mid \begin{bmatrix} -1_{2n_m}^\top & 1_{2n_m}^\top \end{bmatrix} x_i \leq 0, 0 \leq g_{i,\tau} \leq u_{i,\tau}, 0 \leq s_{i,\tau}, \tau = 1, \dots, n_m \right\},$$

while the coupling constraints are defined by

$$A_i \coloneqq \begin{bmatrix} I_{n_m} & 0_{n_m} \\ -I_{n_m} & 0_{n_m} \end{bmatrix}, \quad b_i \coloneqq \frac{1}{N} \begin{bmatrix} r_1 & \dots & r_{n_m} & -d_1 & \dots & -d_{n_m} \end{bmatrix}^\top.$$

Chapter 2

Gradient Tracking Algorithms: System Theoretical Perspective and Algorithm Extensions for Asynchronous, Derivative-Free, and Online Scenarios

In this chapter, we focus on the Gradient Tracking algorithm, i.e., a distributed method widely employed to solve over a network of N agents distributed consensus optimization problems (see Section 1.2) in the form

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{N} f_i(x), \tag{2.1}$$

with each function $f_i : \mathbb{R}^n \to \mathbb{R}$ known to agent i only. In Section 2.2, we begin by providing a system theoretical analysis to study the convergence properties of the Gradient Tracking in the case of nonconvex objective functions. The analysis takes on a singular perturbation perspective and LaSalle-based arguments. Afterward, In Section 2.3, we focus on quadratic programs and perform a control-based approach to design a sparse gain matrix enhancing the convergence properties of the standard Gradient Tracking. Subsequently, in Section 2.4, we devise the continuous-time counterpart of the Gradient Tracking and two additional versions implementing synchronous and asynchronous inter-agent communication, respectively. A Lyapunov-based analysis assesses that all these schemes exponentially converge to the optimal solution of the problem. Then, in Section 2.5, by extending the forward Euler discretization of the continuous-time scheme introduced in Section 2.4, we develop a distributed algorithm for derivative-free

scenarios, i.e., the ones in which the agents cannot access the gradients of their cost functions. We perform a system theoretical analysis to guarantee that the scheme converges to a neighborhood of the solution. Finally, in Section 2.6, we focus on the time-varying setting and we address it by proposing GTAdam, i.e., a novel distributed method obtained by combining the Gradient Tracking method with Adam. In this connection, we provide an upper bound for the dynamic regret achieved by the proposed method. The results of this chapter are based on [25, 28, 31, 33, 130].

2.1 Literature Review

In distributed consensus optimization, a network of agents wants to minimize the sum of local functions depending on the same decision variable. Early attempts to address this kind of problem consist in combining the gradient method with average consensus steps to enforce agreement [88, 137, 138]. When constant step-sizes are employed, these methods exhibit fast convergence rates but cannot achieve the exact solution due to the partial knowledge of the gradient of the global objective function. Exact convergence is guaranteed by the so-called *Gradient Tracking* algorithm. This feature is achieved by resorting to a "tracking action" based on dynamic average consensus (see [93, 216]) to reconstruct the gradient of the global objective function gradient in each agent. The convergence properties of Gradient Tracking have been studied under different problem assumptions, see [54, 136, 158, 168, 186, 196, 199, 200]. In [50], the authors analyze the Gradient Tracking in the case of nonconvex objective function over digraphs, while, still in this framework, the authors of [181] study the perturbed push-sum algorithm with diminishing step-size. The first part of this chapter investigates this nonconvex setting but, differently from [50, 181], takes on a system theoretical perspective. Other works have shown the advantages of system theoretical tools for distributed optimization in the convex case as, e.g., [16, 37, 79, 102, 177, 189, 190]. Inspired by these works, we focus on quadratic programs and modify the Gradient Tracking from a control-oriented perspective by designing sparse gain matrices. We mention existing approaches to sparse gain design for dynamical systems because distributed solutions are associated to sparse system matrices. In [112], a Lyapunov-based technique for optimal sparse state-feedback design is proposed leveraging the Alternating Direction Method of Multipliers (ADMM). A similar problem is considered in [111] in which, instead an augmented Lagrangian approach is employed while in [62] the authors used sequential convex programming to accomplish sparse design. In [6], an algorithm for a sparse gain synthesis based on ADMM and regularization is proposed. In [100] a unified design strategy for decentralized linear quadratic state-feedback controllers is proposed to account also for delays. A strategy based on the so-called Projection Lemma is proposed in [67] for the design of structured stabilizers for linear systems.

Continuous-time counterpart of discrete-time algorithms

In this section, we develop the continuous-time counterpart of the Gradient Tracking algorithm. Indeed, numerous applications falling within different domains can be addressed through continuous-time optimization algorithms. In this context, the work [72] proposes a distributed continuous-time optimization algorithm to solve a consensus convex optimization problem over a directed network. A constrained convex problem is solved in [115] for a network of agents having local, second-order dynamics. In [210], a distributed continuous-time projected algorithm is proposed to tackle nonsmooth convex optimization problems with local constraints. Authors in [113] consider continuous-time multi-agent systems with single-integrator dynamics to solve a distributed optimization problem. A proportional-integral protocol is designed in [201] to solve distributed problems with equality and inequality constraints. In [108], an adaptive continuous-time method based is designed to deal with distributed optimization problems with nonconvex objective functions. In [79], the convergence of distributed continuous-time schemes is ensured through passivity-based arguments for both unconstrained and constrained scenarios, also in presence of communication delays. Authors in [103] assess the exponential convergence of their algorithm by decomposing it as a set of interacting input feed-forward passive systems. Paper [133] proposes a continuous-time optimization algorithm inspired by the existing discrete-time algorithm named Newton-Raphson method. Indeed, a branch of research is recently trying to study the convergence properties of dynamic systems representing the continuous counterparts of existing iterative optimization schemes. This line of research starts in [176], where the authors analyze a second-order differential equation associated to the Nesterov's accelerated gradient method. The authors of [193] establish a systematic way to develop discrete-time accelerated algorithms starting from continuous-time differential equations generated by a Lagrangian functional. In [194], the so-called estimating sequence analysis (typically adopted for algorithms with momenta) is connected with a Lyapunov approach to analyze accelerated optimization methods. In [170], the continuous-time counterpart of the Nesterov algorithm and heavy ball algorithm are studied by means of high-resolution ordinary differential equations. Paper [55] studies the first-order mirror descent algorithm by deriving ordinary differential equations from duality gaps.

Distributed continuous-time schemes rely on continuous-time communication among the network agents. In order to avoid the (not implementable) continuous-time communication required by distributed continuous-time schemes, the design of continuous-time distributed algorithms with discrete-time communication has gained attention. In this regard, [92] addresses a consensus optimization problem by proposing a continuous-time optimization algorithm and two related variants characterized by periodic and event-triggered communication among the agents of the network. The same setting

is tackled in [116], where a distributed event-triggered scheme based on is proposed to deal also with quantized communication. Authors in [52] leverage internal model concepts to extend the event-triggered scheme by [92] to reject external disturbances. Paper [89] combines an event-triggered communication policy and a distributed subgradient method. In [207] a continuous-time distributed optimization algorithm with second order dynamics both with continuous communication and event-triggered communication between agents is proposed. A quadratic problem is considered in [213] and a continuous-time algorithm with event-triggered communication is proposed to solve it. An event-triggered implementation of the distributed gradient descent is given in [1] to deal with nonconvex optimization problems.

Derivative-free distributed optimization

Recent years have seen increasing attention in derivative-free scenarios, namely in the case in which neither gradients nor other derivatives of the objective functions are available to agents in the network, see [43] for an overview.

In this context, the key idea consists of the approximation of the local gradients through a finite set of (possibly random) evaluations of the cost functions. The work [179] modifies the distributed gradient method by approximating the gradients through a two-point estimator. Authors in [119] propose a continuous-time gradient-free algorithm based on a distributed gradient algorithm. The work [114] proposes the discrete-time version of the algorithm given in [119]. In [209], random gradient-free oracles are used within a continuous-time distributed algorithm. This kind of estimation technique is used also in [149] and [41]. In [56], a distributed gradient-free algorithm is designed to deal also with quantized inter-agent communication.

The work in [148] instead develops a "directed-distributed projected pseudo-gradient" descent method for directed graphs. In[187], the gradient-free strategy of [209] has been combined with a saddle-point algorithm. Authors in [191] address an online constrained optimization problem by proposing a distributed algorithm relying on the Kiefer-Wolfowitz method to approximate the gradients. In [15], the estimation of the gradient is performed by combining a "simultaneous perturbation stochastic approximation" technique with the so-called matrix exponential learning optimization method. Authors in [162] propose distributed algorithms based on a Frank–Wolfe update.

In this section, the derivative-free setting is addressed through a scheme whose algorithmic structure is inspired by Gradient Tracking, and the unavailable gradients are replaced by resorting to an equilibrium seeking mechanism. In [154], an extremum seeking control, based on classical evolutionary game-theory ideas, is designed to perform distributed real-time resource allocation. The work [155] proposes a distributed continuous-time extremum-seeking scheme using sign-based consensus. An equilibrium seeking technique is used in [127] in a quadratic distributed consensus optimization

framework. Authors in [203] propose a continuous-time distributed equilibrium seeking algorithm based on saddle point dynamics. A distributed proportional-integral equilibrium seeking design technique is proposed in [77]. In [57], a network of agents combines a distributed consensus method with an equilibrium seeking technique. In [164], a distributed optimization method based on sliding mode and extremum seeking is proposed. More recently, [109] tackles in a distributed fashion a stochastic source localization problem by using a method based on equilibrium seeking. Authors in [76] propose a distributed method that is inspired by the Newton method and uses equilibrium seeking to approximate both the gradients and hessian functions. As for constraint-couopled setting, in [129], a Lie bracket approximation technique is exploited to implement the extremum seeking strategy for problems with linear constraints. In [188], resource allocation problems are addressed by an extremum seeking algorithm which is shown to be semi-globally practically stable.

It is worth mentioning that our scheme, together with the ones in [99,185], is the only distributed extremum-seeking scheme proposed in discrete-time with the following distinctive features. The work [185] (i) does not address a consensus optimization problem and (ii) relies on consensus dynamics estimating the global cost, while ours estimates the global gradient. Instead, in [99], (i) the addressed consensus optimization problems have scalar decision variables and (ii) an extremum-seeking technique is combined with a distributed gradient algorithm, i.e., without a tracking mechanism.

Distributed Online Consensus Optimization

Since stationary optimization problems are of limited use in a multitude of practical applications in dynamic environments, online optimization is gaining increasing popularity, see, e.g., [65,174] where, in a centralized setting, the authors consider problems with time-varying costs. As for distributed online optimization, see the recent survey [105] for an overview of the algorithms and the applications arising in this context. An online distributed subgradient scheme is proposed in [34]. Variations over time of both the cost and constraint functions are handled in [184] in a distributed fashion by using an adaptive diffusive algorithm. In [122], a class of coordination algorithms that generalize distributed online subgradient descent and saddle-point dynamics is proposed to tackle online problems. Authors in [2] combine a subgradient flow with a push-sum consensus for online settings in which also the graph varies over time. A distributed algorithm based on dual subgradient averaging is proposed in [86] to address cost uncertainties and switching communication topologies. A distributed online algorithm inspired by the mirror descent algorithm is proposed in [169]. In [3], it is proposed a distributed online scheme based on the alternating direction method of multipliers, while [206] takes into account also time-varying inequality constraints. Online optimization is strictly related to stochastic optimization, see [64,156,161] for

distributed approaches to this kind of problem. As for the Gradient Tracking scheme, in [211], it is used for online optimization problems, while, in [140], it is combined with a recursive least squares scheme to address in a distributed way the so-called personalized optimization framework (see [171] that introduces this problem in a centralized setting).

2.2 Nonconvex Distributed Optimization via LaSalle and Singular Perturbations

In this section, we analyze the Gradient Tracking algorithm in the case of nonconvex objective functions. In particular, we provide a system theoretical analysis based singular perturbation and LaSalle argument to assess that the solution estimates asymptotically converge to a stationary point of the problem in a consensual manner.

As formalized in the next assumption, we do not require the convexity of f_i, thus making this work attractive for complex settings as, e.g., the ones involving big-data and deep learning (where nonconvex cost functions are often used).

Assumption 2.1 (Objective function)**.** *For all $i \in \{1, \dots, N\}$, $f_i : \mathbb{R}^n \to \mathbb{R}$ is of class $\mathcal{C}^1$ and has $\overline{L}$-Lipschitz continuous gradient, for some $\beta > 0$. Moreover, $f(x) := \sum_{i=1}^N f_i(x)$ is radially unbounded.* △

In this context, we model the inter-agent communication through a graph $\mathcal{G}$ and an associated adjacency matrix $\mathcal{W}_\mathcal{G}$ (see Section 1.1 for further details). The next assumption formalizes the class of networks considered in this section.

Assumption 2.2 (Network)**.** *The directed graph $\mathcal{G}$ is strongly connected and the associated weighted adjacency matrix $\mathcal{W}_\mathcal{G}$ is doubly stochastic.* △

The Gradient Tracking algorithm is a popular distributed method to solve instances of problem (2.1). We provide as follows the idea besides the design of this method. An effective strategy to solve problem (2.1) is the gradient descent method. The latter is the iterative procedure in which each agent $i \in \{1, \dots, N\}$, at each iteration $k \in \mathbb{N}$, maintains an estimate $\mathrm{x}_i^k \in \mathbb{R}^n$ of a solution of problem (2.1) and updates such an estimate by using the gradient of the objective function. Indeed, when applied to problem (2.1), the gradient descent update of agent i reads as

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k - \gamma \nabla f(\mathrm{x}_i^k), \tag{2.2}$$

where $\gamma > 0$ is a step-size. However, in our distributed setting, agent i can only access its local gradient $\nabla f_i(\mathrm{x}_i^k)$ and, thus, the global gradient $\nabla f(\mathrm{x}_i^k) = \sum_{j=1}^N \nabla f_j(\mathrm{x}_i^k)$ is not locally available. To overcome this issue, the Gradient Tracking algorithm uses (i) an auxiliary state $\mathrm{s}_i^k \in \mathbb{R}^n$ called tracker, and (ii) an average consensus step to enforce

consensus among the local solution estimates x_i^k of the agents. Thus, update (2.2) is modified as

$$\mathrm{x}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{x}_j^k - \gamma \mathrm{s}_i^k,$$

where each $w_{ij} \geq 0$ is the (i,j)-th entry of the adjacency matrix $\mathcal{W}_\mathcal{G}$. The goal of the trackers s_i^k is to reconstruct, in each agent $i \in \{1,\dots,N\}$, the global vector $\sum_{j=1}^N \nabla f_j(\mathrm{x}_j^k)$. To this end, each agent updates s_i^k according to the following perturbed average consensus scheme

$$\mathrm{s}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{s}_j^k + \nabla f_i(\mathrm{x}_i^{k+1}) - \nabla f_i(\mathrm{x}_i^k).$$

Thus, the whole update of the Gradient Tracking scheme from the local perspective of agent i reads as

$$\mathrm{x}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{x}_j^k - \gamma \mathrm{s}_i^k \tag{2.3a}$$

$$\mathrm{s}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{s}_j^k + \nabla f_i(\mathrm{x}_i^{k+1}) - \nabla f_i(\mathrm{x}_i^k). \tag{2.3b}$$

The initialization $\mathrm{s}_i^0 = \nabla f_i(\mathrm{x}_i^0)$ is required for all $i \in \{1,\dots,N\}$. We notice that (2.3b) is not causal in the sense that s_i^{k+1} depends on x_i^{k+1}.

In order to recover a causal (still distributed) version of (2.3), following [16], we define $\mathrm{z}_i^k = \gamma(\mathrm{s}_i^k - \nabla f_i(\mathrm{x}_i^k))$ and accordingly rewrite the scheme dynamics as

$$\mathrm{x}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{x}_j^k - \mathrm{z}_i^k - \gamma\nabla f_i(\mathrm{x}_i^k) \tag{2.4a}$$

$$\mathrm{z}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{z}_j^k - \gamma\nabla f_i(\mathrm{x}_i^k) + \gamma\sum_{j\in\mathcal{N}_i} w_{ij}\nabla f_j(\mathrm{x}_j^k). \tag{2.4b}$$

In an aggregate form Algorithm (2.4) reads as

$$\mathrm{x}^{k+1} = \mathcal{W}\mathrm{x}^k - \mathrm{z}^k - \gamma G(\mathrm{x}^k) \tag{2.5a}$$

$$\mathrm{z}^{k+1} = \mathcal{W}\mathrm{z}^k - \gamma(I_{Nn} - \mathcal{W})G(\mathrm{x}^k), \tag{2.5b}$$

where $\mathcal{W} := \mathcal{W}_\mathcal{G} \otimes I_n$ and

$$\mathrm{x}^k := \begin{bmatrix} \mathrm{x}_1^k \\ \vdots \\ \mathrm{x}_N^k \end{bmatrix}, \quad \mathrm{z}^k := \begin{bmatrix} \mathrm{z}_1^k \\ \vdots \\ \mathrm{z}_N^k \end{bmatrix}, \quad G(\mathrm{x}^k) := \begin{bmatrix} \nabla f_1(\mathrm{x}_1^k) \\ \vdots \\ \nabla f_N(\mathrm{x}_N^k) \end{bmatrix}.$$

As shown in [16], in these new coordinates, the initialization of the auxiliary state reads as $\mathbf{1}_{N,n}^\top \mathrm{z}^0 = 0$.

2.2.1 Gradient Tracking as a Singularly Perturbed System

Here, we provide an equivalent reformulation of the Gradient Tracking which take advantages on the properties of the doubly stochastic matrix $\mathcal{W}_\mathcal{G}$ to give insights on the dynamics of the scheme. Let $R \in \mathbb{R}^{Nn\times(N-1)n}$ be such that $R^\top R = I$ and $R^\top \mathbf{1}_{N,n} = 0$, and define

$$\begin{bmatrix} \bar{\mathrm{x}}^k \\ \mathrm{x}_\perp^k \end{bmatrix} := \begin{bmatrix} \frac{\mathbf{1}_{N,n}^\top}{N} \\ R^\top \end{bmatrix} \mathrm{x}^k, \quad \begin{bmatrix} \bar{\mathrm{z}}^k \\ \mathrm{z}_\perp^k \end{bmatrix} := \begin{bmatrix} \frac{\mathbf{1}_{N,n}^\top}{N} \\ R^\top \end{bmatrix} z^k. \tag{2.6}$$

Thus, we can rewrite (2.5) as

$$\begin{aligned} \begin{bmatrix} \bar{\mathrm{x}}^{k+1} \\ \mathrm{x}_\perp^{k+1} \\ \bar{\mathrm{z}}^{k+1} \\ \mathrm{z}_\perp^{k+1} \end{bmatrix} = \begin{bmatrix} I_n & 0 & \frac{\mathbf{1}_{N,n}}{N} & 0 \\ 0 & R^\top \mathcal{W} R & 0 & -I_{(N-1)n} \\ 0 & 0 & I_n & 0 \\ 0 & 0 & R^\top \mathcal{W} \mathbf{1}_{N,n} & R^\top \mathcal{W} R \end{bmatrix} \begin{bmatrix} \bar{\mathrm{x}}^k \\ \mathrm{x}_\perp^k \\ \bar{\mathrm{z}}^k \\ \mathrm{z}_\perp^k \end{bmatrix} \\ + \gamma \begin{bmatrix} -\frac{\mathbf{1}_{N,n}^\top}{N} \\ -R^\top \\ R^\top(\mathcal{W} - I_{Nn}) \\ 0 \end{bmatrix} G(\mathbf{1}_{N,n}\bar{\mathrm{x}}^k + R\mathrm{x}_\perp^k). \end{aligned} \tag{2.7}$$

The initialization $\mathbf{1}_{N,n}^\top \mathrm{z}^0 = 0$ guarantees that $\bar{\mathrm{z}}^k \equiv 0$ for all $k \geq 0$ so that we can neglect the dynamics of $\bar{\mathrm{z}}$ and study

$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k - \frac{\gamma}{N} \mathbf{1}_{N,n}^\top G\left(\mathbf{1}_{N,n}\bar{\mathrm{x}}^k + R\mathrm{x}_\perp^k\right) \tag{2.8a}$$

$$\begin{bmatrix} \mathrm{x}_\perp^{k+1} \\ \mathrm{z}_\perp^{k+1} \end{bmatrix} = A \begin{bmatrix} \mathrm{x}_\perp^k \\ \mathrm{z}_\perp^k \end{bmatrix} + \gamma B G\left(\mathbf{1}_{N,n}\bar{\mathrm{x}}^k + R\mathrm{x}_\perp^k\right), \tag{2.8b}$$

with

$$A := \begin{bmatrix} R^\top \mathcal{W} R & -I_{(N-1)n} \\ 0 & R^\top \mathcal{W} R \end{bmatrix}, \; B := \begin{bmatrix} -R^\top \\ R^\top(\mathcal{W} - I_{Nn}) \end{bmatrix}. \tag{2.9}$$

System (2.8) fits the class of singularly perturbed systems (see, e.g., [19]) given by the interconnection between a slow dynamics, which in our case is (2.8a), and a fast one represented by (2.8b) (see Figure 2.1 for a schematic representation). Appendix C is devoted to providing results for this kind of systems. We point out that, numerically, it

clearly appears that the convergence of subsystem (2.8b) is faster than the one of (2.8a).

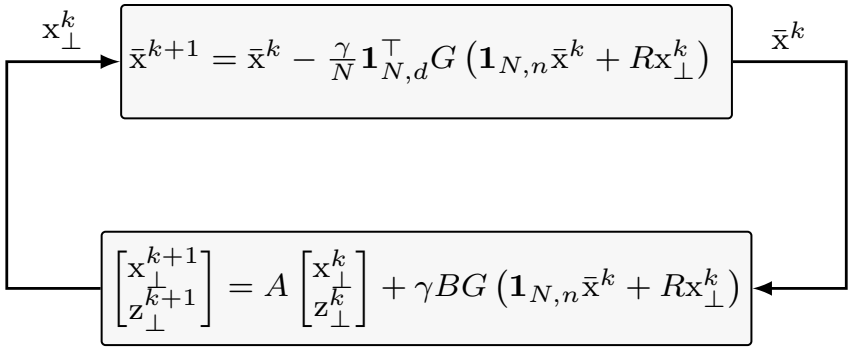


Figure 2.1: Schematic representation of system (2.8).

Now, we provide some notation of singularly perturbed systems that will be used in the analysis. We denote boundary layer system the one obtained by "freezing" $\bar{\mathrm{x}} \in \mathbb{R}^n$ within (2.8b). As we will see later, such a system exhibits an equilibrium parametrized in $\bar{\mathrm{x}}$, say $h(\bar{\mathrm{x}}, \gamma)$. We denote as reduced system the one obtained by considering (2.8a) with $\mathrm{col}(\mathrm{x}_\perp^k, \mathrm{z}_\perp^k) = h(\bar{\mathrm{x}}^k, \gamma)$ for any $k \geq 0$.

Once the Gradient Tracking has been posed in the singularly perturbed form (2.8), we can separately study two auxiliary schemes associated to the subsystems (2.8a) and (2.8b). The *boundary layer system* is obtained by freezing the state of the slow dynamics in the fast one (2.8b). Notice that

$$h(\bar{\mathrm{x}}, \gamma) := \gamma \begin{bmatrix} 0 \\ -R^\top G(\mathbf{1}_{N,n}\bar{\mathrm{x}}) \end{bmatrix} \tag{2.10}$$

is an equilibrium of system (2.8b) for any "frozen" $\bar{\mathrm{x}}$. Now, we introduce the error coordinates of the fast dynamics with respect to $h(\bar{\mathrm{x}}, \gamma)$. Let $\psi^k := \mathrm{col}(\mathrm{x}_\perp^k, \mathrm{z}_\perp^k) - h(\bar{\mathrm{x}}, \gamma)$, we can write the boundary layer system associated to (2.8a) as

$$\psi^{k+1} = A\psi^k + \gamma B u_{\mathrm{bl}}^k(\bar{\mathrm{x}}), \tag{2.11}$$

where

$$u_{\mathrm{bl}}^k(\bar{\mathrm{x}}) := G\left(\mathbf{1}_{N,n}\bar{\mathrm{x}} + \mathcal{R}_1\psi^k\right) - G\left(\mathbf{1}_{N,n}\bar{\mathrm{x}}\right). \tag{2.12}$$

Remark 2.1. From Assumption 2.2 the matrix $\mathcal{W}$ has 1 as an eigenvalue with multiplicity d, while all the remaining ones lie within the open unit circle. The left and right eigenvectors associated to 1 belong to the span of $\mathbf{1}_{N,n}^\top$ and $\mathbf{1}_{N,n}$, respectively. Thus, $R^\top \mathcal{W} R$ is Schur. Then, the matrix A, being up-triangular with two Schur matrices on the diagonal blocks, is Schur too. See [16] for a detailed discussion. △

Exponential stability of (2.11) uniformly in $\bar{\mathrm{x}}$ is given now.

Lemma 2.1. *Consider system* (2.11). *Let* $m := 2(N-1)n$. *Then, there exists* $\bar{\gamma}_1 > 0$ *and a Lyapunov function* $W : \mathbb{R}^m \to \mathbb{R}$ *such that, for any* $\gamma \in (0, \bar{\gamma}_1)$, *it holds*

$$b_1 \|\psi\|^2 \leq W(\psi) \leq b_2 \|\psi\|^2 \tag{2.13a}$$

$$W(A\psi + \gamma u_{bl}^k(\bar{\mathrm{x}})) - W(\psi) \leq -b_3 \|\psi\|^2 \tag{2.13b}$$

$$|W(\psi_1) - W(\psi_2)| \leq b_4 \|\psi_1 - \psi_2\| (\|\psi_1\| + \|\psi_2\|), \tag{2.13c}$$

for any $\psi, \psi_1, \psi_2 \in \mathbb{R}^m$, $\bar{\mathrm{x}} \in \mathbb{R}^n$, *and some* $b_1, b_2, b_3, b_4 > 0$.

Proof. Pick any $Q \in \mathbb{R}^{m\times m}$, $Q = Q^\top > 0$. Being A Schur (see Remark 2.1), there exists $P = P^\top > 0$ so that

$$A^\top P A - P = -Q. \tag{2.14}$$

Pick such P to define $W : \mathbb{R}^m \to \mathbb{R}$ as

$$W(\psi^k) := (\psi^k)^\top P \psi^k,$$

which clearly satisfies (2.13a) and (2.13c). Further, along (2.11), $\Delta W(\psi^k) := W(\psi^{k+1}) - W(\psi^k)$ is given by

$$\begin{aligned}\Delta W(\psi^k) &= (\psi^k)^\top (A^\top P A - P)\psi^k + 2\gamma (\psi^k)^\top A^\top P B u_{\mathrm{bl}}^k(\bar{\mathrm{x}}) + \gamma^2 (u_{\mathrm{bl}}^k(\bar{\mathrm{x}}))^\top B^\top P B u_{\mathrm{bl}}^k(\bar{\mathrm{x}}) \\ &\overset{(a)}{\leq} -(\psi^k)^\top Q \psi^k + 2\gamma \left\|A^\top P B\right\| \left\|\psi^k\right\| \left\|u_{\mathrm{bl}}^k(\bar{\mathrm{x}})\right\| + \gamma^2 \left\|B^\top P B\right\| \left\|u_{\mathrm{bl}}^k(\bar{\mathrm{x}})\right\|^2,\end{aligned} \tag{2.15}$$

where in (a) we have used the result (2.14) and the Cauchy-Schwarz inequality. Being each ∇f_i Lipschitz continuous (cf. Assumption 2.1), we bound $u_{\mathrm{bl}}^k(\bar{\mathrm{x}})$ (defined in (2.12)) as

$$\left\|u_{\mathrm{bl}}^k(\bar{\mathrm{x}})\right\| \leq \overline{L} \|\mathcal{R}_1\| \left\|\psi^k\right\|, \tag{2.16}$$

for any $\bar{\mathrm{x}} \in \mathbb{R}^n$. Thus, we can use (2.16) to bound (2.15) as

$$\Delta W(\psi^k) \leq -(q - \gamma c_1 - \gamma^2 c_2) \left\|\psi^k\right\|^2, \tag{2.17}$$

where q is the (positive) smallest eigenvalue of Q, while $c_1 := 2\overline{L} \left\|A^\top P B\right\| \|\mathcal{R}_1\|$, $c_2 := \overline{L}^2 \left\|B^\top P B\right\| \|\mathcal{R}_1\|^2$. Thus, there exists $\bar{\gamma}_1 > 0$ such that $q - \gamma c_1 - \gamma^2 c_2 > 0$ for any $\gamma \in (0, \bar{\gamma}_1)$ so that (2.13b) holds and the proof follows. ■

Now, we consider $\mathrm{col}(\mathrm{x}_\perp^k, \mathrm{z}_\perp^k) = h(\bar{\mathrm{x}}^k, \gamma)$ for all $k \geq 0$ within (2.8a) obtaining the

so-called reduced system as

$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k - \frac{\gamma}{N}\mathbf{1}_{N,n}^\top G\left(\mathbf{1}_{N,n}\bar{\mathrm{x}}^k + \mathcal{R}_1 h(\bar{\mathrm{x}}^k, \gamma)\right). \tag{2.18}$$

By exploiting h (cf. (2.10)) and G, we write system (2.18) as

$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k - \frac{\gamma}{N}\nabla f(\bar{\mathrm{x}}^k). \tag{2.19}$$

It is worth noting that the reduced system has recovered the desired update (2.2), i.e., the one given by applying the (centralized) gradient descent method to solve problem (2.1). The next lemma uses the radially unbounded (see Definition A.2 in Appendix A) function f to show the convergence of system (2.19) to the set $X^\star := \{\bar{\mathrm{x}} \in \mathbb{R}^n \mid \nabla f(\bar{\mathrm{x}}) = 0\}$ of stationary points of (2.1).

Lemma 2.2. *Consider system* (2.19) *Then, there exists $\bar{\gamma}_2 > 0$ such that, for any $\gamma \in (0, \bar{\gamma}_2)$, it holds*

$$f\left(\bar{\mathrm{x}} - \frac{\gamma}{N}\nabla f(\bar{\mathrm{x}})\right) - f(\bar{\mathrm{x}}) \leq -\gamma d_1 \left\|\nabla f(\bar{\mathrm{x}})\right\|^2 \tag{2.20a}$$

$$f(\mathrm{x}_1 + \mathrm{x}_2) - f(\mathrm{x}_1 + \mathrm{x}_3) \leq d_2 \left\|\nabla f(\bar{\mathrm{x}}_1)\right\| \left\|\mathrm{x}_2 - \mathrm{x}_3\right\| + d_3\left(\left\|\mathrm{x}_2\right\|^2 + \left\|\mathrm{x}_3\right\|^2\right), \tag{2.20b}$$

for any $\mathrm{x}, \mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3 \in \mathbb{R}^n$, and some $d_1, d_2, d_3 > 0$.

Proof. In light of Assumption 2.1, f has Lipschitz continuous gradient. Thus, we apply the Descent Lemma (cf. [14, Proposition 6.1.2]) to write

$$f(\bar{\mathrm{x}}^{k+1}) - f(\bar{\mathrm{x}}^k) \leq -\frac{\gamma}{N}\left\|\nabla f(\bar{\mathrm{x}}^k)\right\|^2 + \gamma^2\frac{\overline{L}}{2N^2}\left\|\nabla f(\bar{\mathrm{x}}^k)\right\|^2. \tag{2.21}$$

Choose any $d_1 \in (0, 1/N)$. Then, for any $\gamma \in (0, \bar{\gamma}_2)$ with $\bar{\gamma}_2 := \frac{2N(1-Nd_1)}{\overline{L}} > 0$, the inequality (2.21) ensures that (2.20a) is satisfied. With same arguments, also the inequality (2.20b) with $d_2 = 1$ and $d_3 = \frac{\overline{L}}{2}$ can be shown. ■

Once these preliminary results have been provided, we can use them in the next theorem to state the convergence properties of the Gradient Tracking distributed algorithm.

Theorem 2.1. *Consider the Gradient Tracking given in* (2.5)*. Let Assumptions 2.1 and 2.2 hold. Then, for any initial condition $(\mathrm{x}^0, \mathrm{z}^0) \in \mathbb{R}^{2Nn}$ such that $\mathbf{1}_{N,n}^\top \mathrm{z}^0 = 0$, there exists $\bar{\gamma} > 0$ such that, for any $\gamma \in (0, \bar{\gamma})$, it holds*

$$\lim_{k\to\infty} \inf_{\xi \in X^\star} \left\|\mathrm{x}_i^k - \xi\right\| = 0, \quad \forall i \in \{1, \ldots, N\}.$$

Proof. The proof relies on Theorem C.1 in Appendix C. Indeed, Lemma 2.1 and Lemma 2.2 provide the functions W and $U \equiv f$ satisfying conditions (C.4) and (C.5),

respectively. Further, Assumption 2.1 guarantees the radial unboundedness of U and the regularity properties required for ϕ, g, and h. Hence, by Theorem C.1, there exists $\bar{\gamma} > 0$ such that, for any $\gamma \in (0, \bar{\gamma})$, any trajectory of system (2.8) satisfies

$$\lim_{k\to\infty} \inf_{\xi\in\mathcal{M}'} \left\| \begin{bmatrix} \bar{\mathrm{x}}^k \\ \mathrm{x}_\perp^k \\ \mathrm{z}_\perp^k \end{bmatrix} - \begin{bmatrix} \xi \\ h(\xi,\gamma) \end{bmatrix} \right\|, \tag{2.22}$$

where $\mathcal{M}' \subseteq \ker\{\nabla f(\cdot)\} \subseteq \mathbb{R}^n$ denotes the largest invariant set for system (2.19) contained within $\ker\{\nabla f(\cdot)\}$. The proof follows by noting that $\mathcal{M}' \equiv \ker\{\nabla f(\cdot)\} \equiv X^\star$. ■

2.2.2 Numerical Simulations

This section validates our theoretical findings with a numerical simulation about the target localization problem given in [54, Section IV.A]. A network of $N = 10$ agents aims to locate a common target through some distance measurements. Each agent i is located at $\omega_i \in \mathbb{R}^n$ and has a noisy measurement $\phi_i > 0$ of the target squared distance. The target position is estimated by solving the nonconvex distributed consensus optimization problem

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{N} \left(\phi_i - \|x - \omega_i\|^2\right)^2.$$

We set $n = 3$ and consider an Erdős-Rényi directed graph with parameter 0.6. We uniformly randomly set each component of the target location within the interval $[0, 1]$. As for the parameters ϕ_i, we generate them adding Gaussian noise to the target location. We uniformly randomly generate the parameters ω_i within the interval $[0, 1]$ for all $i \in \{1, \dots, N\}$. We randomly generate the initial conditions x_i^0 choosing them according to 3-dimensional Gaussian distributions with unitary variance centered in the target location. The step-size parameter is empirically tuned as $\gamma = 0.01$ to guarantee the algorithm effectiveness. Figure 2.2 separately shows the convergence of the fast and slow subsystems identified in (2.8) and graphically highlights their different rates.

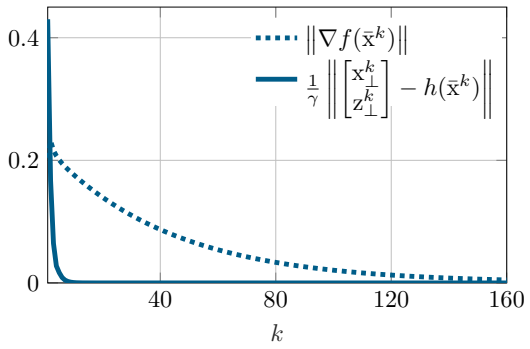


Figure 2.2: Convergence of the fast and slow subsystem.

2.3 Revisited Gradient Tracking Algorithms for Distributed Quadratic Optimization via Sparse Gain Design

In this section, we consider a quadratic instance of problem (2.1), i.e., the optimization problem described by

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{N} f_i(x), \tag{2.23}$$

in which, for each $i \in \{1,\dots,N\}$, $f_i : \mathbb{R}^n \to \mathbb{R}$ has the following quadratic form

$$f_i(x) = \frac{1}{2}(x - \Gamma_i\theta_0)^\top C_i(x - \Gamma_i\theta_0), \tag{2.24}$$

with $C_i \in \mathbb{R}^{n\times n}$ symmetric and positive definite, $\Gamma_i \in \mathbb{R}^{n\times p}$, and $\theta_0 \in \mathbb{R}^p$. It is easy to show that (2.23) admits a unique optimal solution $x^\star \in \mathbb{R}^n$ given by

$$x^\star = \Sigma\theta_0, \tag{2.25}$$

with

$$\Sigma := \left(\sum_{i=1}^{N} C_i\right)^{-1} \sum_{i=1}^{N} C_i\Gamma_i.$$

Since problem (2.23) has quadratic costs (cf. (2.24)), each gradient ∇f_i has the linear form

$$\nabla f_i(x_i) = C_i x_i + Q_i\theta_0, \tag{2.26}$$

in which $Q_i = -C_i\Gamma_i$. As before, also in this section we want to solve problem (2.23) over a strongly connected graph. Further, we assume that the agents can communicate using the weighted adjacency matrices $\mathcal{W}_\mathcal{G} \in \mathbb{R}^{N\times N}$ and $\tilde{\mathcal{W}}_\mathcal{G} \in \mathbb{R}^{N\times N}$ which respectively are a row and a column stochastic matrix matching the graph $\mathcal{G}$. Indeed, we want to solve

problem (2.23) by using a slightly modified version of the Gradient Tracking algorithm as we (cf. (2.5)). First, we use $\mathcal{W}_{\mathcal{G}}$ for the average consensus step of the solution estimates, and $\tilde{\mathcal{W}}$ for the one of the auxiliary variables. Hence, by exploiting the linearity of the gradients (cf. (2.26)), we get a linear time invariant system described by

$$\begin{bmatrix} \mathrm{x}^{k+1} \\ \mathrm{z}^{k+1} \end{bmatrix} = F_\gamma \begin{bmatrix} \mathrm{x}^{k} \\ \mathrm{z}^{k} \end{bmatrix} + G_\gamma \theta_0, \tag{2.27}$$

in which $\mathrm{x}^k, \mathrm{z}^k \in \mathbb{R}^{Nn}$ have the same meaning as in (2.5) and we introduced

$$F_\gamma := \begin{bmatrix} \mathcal{W} - \gamma C & -\gamma I \\ (\tilde{\mathcal{W}} - I)C & \tilde{\mathcal{W}} \end{bmatrix}, \quad G_\gamma := \begin{bmatrix} -\gamma Q \\ (\tilde{\mathcal{W}} - I)Q \end{bmatrix},$$

where $\mathcal{W} := \mathcal{W}_{\mathcal{G}} \otimes I_n$ and $\tilde{\mathcal{W}} := \tilde{\mathcal{W}}_{\mathcal{G}} \otimes I_n$, and

$$C := \begin{bmatrix} C_1 & & & \\ & C_2 & & \\ & & \ddots & \\ & & & C_N \end{bmatrix}, \quad Q := \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_N \end{bmatrix}.$$

If $\gamma = 0$, the state matrix F_γ has an eigenvalue at 1 with multiplicity $2n$ and, therefore, it is not Schur. For γ positive and sufficiently small, instead, the eigenvalues of F_γ move inside the unit circle, and F_γ becomes Schur, see [16]. In this sense, the term $-\gamma(z + y)$ can be interpreted as a *stabilizing output-feedback control action*, conferring asymptotic stability on (2.27) with $\theta_0 = 0$. The main idea of this section is that we may substitute the "control gain" $-\gamma I$ with a general matrix $K \in \mathbb{R}^{Nn \times Nn}$, in this way obtaining a variation of the Gradient Tracking algorithm that we denote as Revisited Gradient Tracking and is described by

$$\begin{bmatrix} \mathrm{x}^{k+1} \\ \mathrm{z}^{k+1} \end{bmatrix} = F_K \begin{bmatrix} \mathrm{x}^{k} \\ \mathrm{z}^{k} \end{bmatrix} + G_K \theta_0, \tag{2.28}$$

in which

$$F_K := \begin{bmatrix} \mathcal{W} + KC & K \\ (\tilde{\mathcal{W}} - I)C & \tilde{\mathcal{W}} \end{bmatrix}, \quad G_K := \begin{bmatrix} KQ \\ (\tilde{\mathcal{W}} - I)Q \end{bmatrix}.$$

By following [16], we introduce some definitions to formally establish sufficient conditions to solve problem 2.23. Define $m := 2Nn$ and, with $m_\nu \leq m$, consider an m_ν-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^m$, and let $T \in \mathbb{R}^{m \times m}$ be an orthonormal matrix of the form $T = [T_1, T_2]$, with $T_1 \in \mathbb{R}^{m \times m_\nu}$ and $T_2 \in \mathbb{R}^{m \times (m - m_\nu)}$ satisfying

$$\mathrm{Im}(T_1) = \mathcal{V} \quad , \quad \mathrm{Im}(T_2) = \mathcal{V}^\perp.$$

Then, $\mathcal{V}$ is F-invariant if and only if

$$T^\top F T = \begin{bmatrix} F_I & F_J \\ 0 & F_E \end{bmatrix},$$

for some $F_I \in \mathbb{R}^{m_\nu \times m_\nu}$, $F_J \in R^{m_\nu \times (m-m_\nu)}$ and $F_E \in \mathbb{R}^{(m-m_\nu)\times(m-m_\nu)}$.

Definition 2.1 (Internal stability and external anti-stability [16])**.** *The subspace $\mathcal{V}$ is said to be:*

- *internally stable if F_I is Schur;*
- *externally anti-stable if F_E has no eigenvalues inside the open unit disc.* △

Let $\mathcal{O}$ be an affine subspace of $\mathbb{R}^m$ of the form

$$\mathcal{O} := \mathcal{V} + U\theta_0, \tag{2.29}$$

for some linear subspace $\mathcal{V}$ of $\mathbb{R}^m$ of dimension m_ν and for some matrix $U \in \mathbb{R}^{(m-m_\nu)\times p}$ satisfying $\mathrm{Im}(U) \subset \mathcal{V}^\perp$.

Definition 2.2 (Admissible initialization [16])**.** *Consider system* (2.28)*. A set $\mathcal{O}$ of the form* (2.29) *is said to be an admissible initialization set if $\mathcal{V}$ is F_K-invariant and externally anti-stable.* △

With the above definitions at reach, the results of [16] are summarized within the following theorem.

Theorem 2.2. *Consider system* (2.28) *and suppose that $(x(0), z(0)) \in \mathcal{O}$, in which $\mathcal{O}$ is an admissible initialization set of the form* (2.29)*. If*

- *$\mathcal{V}$ is internally stable;*
- *$U = T_2(T_2^\top T_2)^{-1}T_2^\top \Pi$, with $\Pi = \mathrm{col}(\mathbf{1}_{N,n}\Sigma, -C\mathbf{1}_{N,n}\Sigma - Q)$,*

then it holds

$$\lim_{k\to\infty} \left\| \mathrm{x}^k - \mathbf{1}_{N,n}\theta^\star \right\|,$$

namely all the estimates $\mathrm{x}_1^k, \dots, \mathrm{x}_N^k$ asymptotically converge to the optimal solution of (2.23)*.* △

Theorem 2.2 suggests that the matrix F_K in (2.28) needs to be designed to possess an internally stable subspace that we can use to define an admissible initialization set. We underline, however, that not every subspace fits our purposes. In fact, if the matrix U in (2.29) is not zero, then the admissible initialization of the algorithm would

depend on the unknown variable θ_0 and, as such, it would not be implementable. In the following, we construct a matrix K ensuring that the corresponding matrix F_K possesses an invariant subspace $\mathcal{V}$ with the desired properties, whose corresponding matrix U is zero. The search of such K is approached as a stabilization problem. Consider the transformation matrix $T := [T_1, T_2]$ with

$$T_1 := \begin{bmatrix} I & 0 \\ 0 & R \end{bmatrix}, \quad T_2 := \tfrac{1}{\sqrt{N}} \begin{bmatrix} 0 \\ \mathbf{1}_{N,n} \end{bmatrix},$$

in which $R \in \mathbb{R}^{Nd \times N(d-1)}$ is such that $RR^\top = I$ and $R^\top I_n = 0$. Then, it holds $T^{-1} = T^\top$, and T transforms F_K to

$$T^\top F_K T = \begin{bmatrix} F_{K_I} & F_{K_J} \\ 0 & F_{K_E} \end{bmatrix},$$

in which

$$\begin{aligned} F_{K_I} &:= \begin{bmatrix} \mathcal{W} + KC & KR \\ R^T(\tilde{\mathcal{W}} - I)C & R^T\tilde{\mathcal{W}}R \end{bmatrix} \in \mathbb{R}^{(m-n)\times(m-n)}, \\ F_{K_J} &:= \tfrac{1}{\sqrt{N}} \begin{bmatrix} K\mathbf{1}_{N,n} \\ R^\top\tilde{\mathcal{W}}\mathbf{1}_{N,n} \end{bmatrix} \in \mathbb{R}^{(m-n)\times n}, \\ F_{K_E} &:= I \in \mathbb{R}^{n\times n}. \end{aligned} \tag{2.30}$$

The structure of the matrix $T^\top F_K T$ implies that there exists an $(m-n)$-dimensional subspace $\mathcal{V}$ that is F_K-invariant. This subspace is given by vectors $w \in \mathbb{R}^m$ such that $Tw = \text{col}(\tilde{w}_1, 0)$, with $\tilde{w}_1 \in \mathbb{R}^{m-n}$. Equivalently we can say that $\mathcal{V} = \{\text{col}(x, z) \in \mathbb{R}^m | z := (z_1; z_2; \dots; z_N) \in \mathbb{R}^{Nn}, \sum_{i=1}^N z_i = 0\}$. We point out that, using the definition in Theorem 2.2, the choice for T_2 implies that $U = 0$. We stress that, according to (2.29), having $U = 0$ ensures that the algorithm can be properly initialized without relying on any unknown quantity (in this case x^0 is arbitrary, while z^0 is only constrained to have a zero mean). We also remark that this is consistent with (and actually slightly milder than) the usual initialization of gradient tracking algorithms (see [16]). It remains to show that K can be chosen to guarantee that $\mathcal{V}$ is also internally stable, in this way completing the design in view of Theorem 2.2. According to Definition 2.1, we have that $\mathcal{V}$ is internally stable if and only if F_{K_I} is Schur. Moreover, the matrix F_{K_I} can be further decomposed as

$$F_{K_I} = F_{K_{I0}} + B_{I0}KH,$$

in which

$$F_{K_{I0}} := \begin{bmatrix} \mathcal{W} & 0 \\ R^\top(\tilde{\mathcal{W}} - I)C & R^\top\tilde{\mathcal{W}}R \end{bmatrix}, \quad B_{I0} := \begin{bmatrix} I \\ 0 \end{bmatrix},$$

with $H := [C, R]$. The design of a matrix K such that F_{K_I} is Schur, on the other hand,

can be cast as the stabilization of the following linear system

$$x_I^+ = F_{K_{I0}} x_I + B_{I0} u_0, \tag{2.31a}$$

$$u_0 = KHx_I. \tag{2.31b}$$

From (2.31), it can be seen that the matrix F_{K_I} is the closed-loop matrix obtained by choosing a feedback control law. This shows that the design of the gradient tracking algorithm can be posed as a feedback stabilization problem. Notice that, however, in order to preserve the distributed nature of the optimization algorithm, we need to add an additional sparsity requirement on the gain, which will be discussed in the following section.

2.3.1 Revisited Gradient Tracking: Sparse Gain Design

In this section we present our algorithmic strategy to design a sparse gain K for (2.28) in order to solve problem (2.23).

To this end, we derive a Linear Matrix Inequality (LMI) to obtain a stabilizing gain K for system (2.31). The approach relies on a discrete-time version of the Lyapunov-based approach presented in [20] for continuous-time systems. We consider the closed-loop system obtained by substituting the feedback control (2.31b) in (2.31a)

$$\mathrm{x}_I^k = (F_{K_{I0}} + B_{I0}KH)\mathrm{x}_I^k. \tag{2.32}$$

For the sake of readability, from now on, we drop the subscripts in (2.32) and write

$$\mathrm{x}^k = (F + BKH)\mathrm{x}^k. \tag{2.33}$$

The linear time-invariant system (2.33) is asymptotically stable if and only if there exist $Q = Q^\top \in \mathbb{R}^{m_\nu \times m_\nu}$ and $K \in \mathbb{R}^{Nn \times Nn}$ satisfying

$$\begin{cases} Q > 0 \\ Q - (F + BKH)^\top Q (F + BKH) > 0. \end{cases} \tag{2.34}$$

Notice that (2.34) is not linear in the unknown (Q, K). However, it can be equivalently written as

$$\begin{cases} Q > 0 \\ Q - (Q(F + BKH))^\top Q^{-1} (Q(F + BKH)) > 0. \end{cases} \tag{2.35}$$

Using the Schur complement lemma (cf. [20]), we can write (2.35) as

$$\begin{bmatrix} Q & Q(F+BKH) \\ (Q(F+BKH))^\top & Q \end{bmatrix} > 0 \tag{2.36}$$

that is still meant to be solved in the unknowns Q and K.

Let $P := Q^{-1}$. Since Q is symmetric, then also P is symmetric. By pre- and post-multiplying (2.36) by the following symmetric and positive definite matrix

$$\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix},$$

we obtain the equivalent inequality

$$\begin{bmatrix} P & (F+BKH)P \\ P(F+BKH)^\top & P \end{bmatrix} > 0, \tag{2.37}$$

which is still not linear, because of the product between the unknowns P and K. We thus introduce a further matrix $L \in \mathbb{R}^{Nn \times 2Nn}$ defined as $L := KHP$, and rewrite (2.37) as

$$\begin{cases} \begin{bmatrix} P & FP+BL \\ PF^\top + L^\top B^\top & P \end{bmatrix} > 0 \\ L - KHP = 0. \end{cases} \tag{2.38}$$

Although (2.38) is linear in both the unknowns P and L, it is still not sufficient to provide a *distributed* solution, since the feedback control law (2.31b) would not be implementable by a network of agents. In fact, the resulting matrix K obtained from (2.38) need not be sparse (and typically it will not), as no *sparsity constraints* are imposed in (2.38). In the next subsection we show how sparsity constraints in K can be included in the solution of (2.38), and we develop an algorithmic procedure to solve the resulting problem.

Encoding Sparsity of the Gain Matrix

In this subsection we add a set of constraints imposing a sparsity pattern to the gain K in order to match the network topology. Formally, $K \in \mathbb{R}^{Nn \times Nn}$ must be such that its (i,j)-th is zero whenever $(i,j) \notin \mathcal{E}$.

For each pair $(i,j) \in \mathcal{E}$, let $M^{ij} \in \mathbb{R}^{Nn \times Nn}$ be the matrix having zeros everywhere except for the (i,j)-entry which is equal to 1. Then, a matrix K satisfying the sparsity constraint of the network can be expressed as a linear combination of the matrices

$M^{ij} \otimes I_n$, i.e.,

$$K = \sum_{(i,j)\in\mathcal{E}} k_{ij}(M^{ij} \otimes I_n), \tag{2.39}$$

with $k_{ij} \in \mathbb{R}$ for all $(i,j) \in \mathcal{E}$. For notational convenience, we let $\boldsymbol{k} \in \mathbb{R}^{|\mathcal{E}|}$ collect all the coefficients k_{ij} in a single vector with an arbitrary ordering of the edges. The expansion (2.39) can be used to encode in (2.38) the desired sparsity constraints. In particular, by substituting (2.39) into the second condition of (2.38), we obtain the following constraints

$$\begin{cases} \begin{bmatrix} P & FP + BL \\ PF^\top + L^\top B^\top & P \end{bmatrix} > 0 \\ L - \sum_{(i,j)\in\mathcal{E}} k_{ij} M^{ij} HP = 0, \end{cases} \tag{2.40}$$

in the unknowns P, L and $\boldsymbol{k}$.

We stress that including the sparsity constraints (2.39) makes (2.40) a nonlinear problem because of the product between P and $\boldsymbol{k}$ in the second condition of (2.40). Unfortunately, no general procedure exists to solve nonlinear problems of this form. Therefore, in the following we propose an iterative procedure to tackle such nonlinear problem.

In the proposed procedure, at each iteration τ, an *approximate* version of (2.40) is solved, in which the decision variable P in the equality constraint of the second condition is substituted with a fixed value, denoted by $\widehat{P}_\tau$, which coincides with the solution found in the previous iteration. In this way, at each iteration τ, we obtain a *linear* matrix inequality in the variables P, L and $\boldsymbol{k}$, given by

$$\begin{bmatrix} P & FP + BL \\ PF^\top + L^\top B^\top & P \end{bmatrix} > 0 \tag{2.41}$$

$$L - \sum_{(i,j)\in\mathcal{E}} k_{ij} M^{ij} H\widehat{P}_\tau = 0, \tag{2.42}$$

which can be efficiently solved using numerical routines. Once a solution $(P_\tau, L_\tau, \boldsymbol{k}_\tau)$ to (2.41) is obtained, the matrix P_τ serves as new value for $\widehat{P}_{\tau+1}$ in the next iteration $\tau + 1$. The procedure starts with an arbitrary initialization $\widehat{P}_0$ and is repeated until some convergence criterion is satisfied, e.g., until $\|P_{\tau+1} - P_\tau\|$ falls below a given threshold $\epsilon > 0$. Notice that the described procedure has no theoretical convergence guarantees. However, in Section 2.3.2 we show the effectiveness of the proposed scheme (2.41) for the design of sparse feedback through simulations.

Remark 2.2. We underline that (2.41) is a feasibility problem. Then, once its constraints

are fulfilled, we can also include an optimality criterion in its selection. For this reason, in our numerical experiments we also add a cost function in order to optimize the convergence rate of the resulting optimization algorithm. △

The following Algorithm 1 summarizes the described iterative procedure.

Algorithm 1 Iterative Procedure for Sparse Gain Design

given tolerance $\epsilon > 0$
initialize $\widehat{P}_0$
for $\tau = 0, 1, 2, \dots$ **do**
  obtain $(P_\tau, L_\tau, \boldsymbol{k}_\tau)$ as a solution to (2.41)
  if $\|P_\tau - \widehat{P}_\tau\| < \epsilon$ **then**
    set $\boldsymbol{k}^\star = \boldsymbol{k}_\tau$
    break
  else
    $\widehat{P}_{\tau+1} = P_\tau$
  end if
end for
retrieve the sparse gain $K^\star = \sum_{(i,j)\in\mathcal{E}} k_{ij}^\star M^{ij}$.

2.3.2 Numerical Simulations

In this section we propose a numerical study to show the effectiveness of the proposed design strategy. In particular, we compare the convergence behavior of the gradient tracking including the sparse (possibly non-diagonal) gain K (cf. (2.28)) with its basic version with diagonal gains (cf. (2.27)).

As mentioned in Remark 2.2, we include in the solution of each problem (2.41) the minimization of the objective

$$\|F + BKH\| + \beta \left\| P - \widehat{P}_\tau \right\|,$$

in which $\beta > 0$ represents a trade-off parameter in the following sense. Minimizing the term $\|F + BKH\|$ reflects in maximizing the convergence rate of the resulting distributed optimization algorithm. Indeed, $\|F + BKH\|$ is directly related to the maximum singular value of the closed-loop matrix $F + BKH$. The term $\|P - \widehat{P}_\tau\|$ is, instead, a "regularization" introduced to foster the convergence of the iterative procedure in Algorithm 1. The design parameter β can be thus chosen to privilege one of the two terms as desired.

We term "Basic GT" the standard Gradient Tracking with diagonal gains, while we term "Rev. GT" the algorithm that implements the sparse gain K designed using our procedure described in Section 2.3.1. In order to to choose the step-size γ in the Basic

GT, we resort to Algorithm 1. Indeed, the case of diagonal K is a special case obtained by imposing a graph structure with only self-loops.

In the following, we present simulations obtained for different numbers of agents N and for different "graph density" $d_A := |\mathcal{E}|/N^2$. We set $\widehat{P}_0 = I$ and $\beta = 1$. In the figures below we plot the norm of the difference between the mean vector and the optimal solution of (2.23), namely $\left\|\bar{\mathrm{x}}^k - x^\star\right\| = \left\|\frac{1}{N}\sum_{i=1}^N \mathrm{x}_i^k - x^\star\right\|$.

In Figure 2.3, three networks with respectively 5, 10 and 15 agents are considered for comparison. In all the cases we consider $d = 2$ and $d_A = 0.7$.

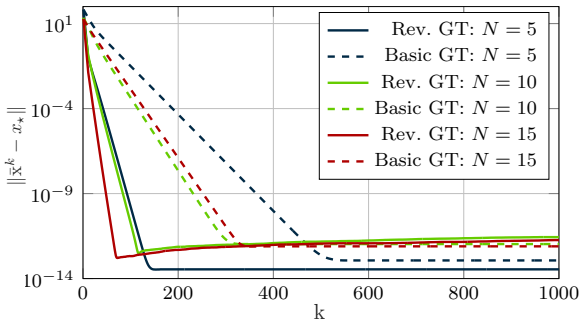


Figure 2.3: Evolution of the optimality error for different values of N.

Figure 2.3 shows that in all cases the Rev. GT has a faster convergence rate than the Basic GT. The converge rate enhances as the number of agents increases. This behavior can be explained by noticing that a larger number of agents implies a "larger space" in which the sparse matrix K is searched.

In Figure 2.4, four networks with density d_A equal to 0.3, 0.6, 0.75, and 0.9 are considered. In all the cases we consider $n = 2$ and $N = 10$.

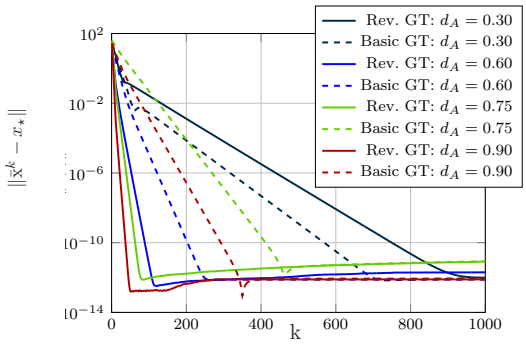


Figure 2.4: Evolution of the optimality error for different values of d_A.

Figure 2.4 shows that only in one case the Basic GT is faster than the Rev. GT. Specifically, it occurs when graph density is really low, namely $d_{\mathcal{W}} = 0.3$. This confirms our interpretation that the Basic GT is actually obtained by limiting the gain structure choice to diagonal matrices. We point out that, although the Basic GT is faster, both gains

are associated to the same cost value $\|F+BKH\|_2+\beta\|P-\widehat{P}_\tau\|_2$. This is reasonable since the cost function does not encode directly an information related to the convergence rate of the closed-loop system. Moreover, also the regularization term plays a role in the optimization procedure. Consistently to the previous case, we observe that the algorithm performance enhances as the graph density d_A increases.

2.4 Asynchronous Distributed Consensus Optimization

In this section, we devise the continuous-time version of the Gradient Tracking algorithm and, then, two additional versions replacing continuous-time inter-agent communication with discrete-time synchronous and asynchronous communication, respectively. Along this section, these schemes will be analyzed under the following assumptions.

Assumption 2.3. *For all i, the function $f_i : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with coefficient $\mu > 0$.* △

Assumption 2.4. *For all i, the function $f_i : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient with constant $\overline{L} > 0$.* △

We recall that Assumption 2.3 ensures that problem (2.1) has a unique optimal solution, denoted by $x^\star \in \mathbb{R}^n$ (cf. Proposition A.2 in Appendix A).

Assumption 2.5. *$\mathcal{G}$ is undirected and connected. Moreover, the associated weighted adjacency matrix $\mathcal{W}_\mathcal{G}$ is symmetric.* △

From Discrete to Continuous

In this section, we will derive the continuous-time counterpart of the Gradient Tracking algorithm. For the sake of readability, we recall its local causal form (2.4)

$$\mathrm{x}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{x}_j^k - \mathrm{z}_i^k - \gamma\nabla f_i(\mathrm{x}_i^k) \tag{2.43a}$$

$$\mathrm{z}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{z}_j^k - \gamma\nabla f_i(\mathrm{x}_i^k) + \gamma\sum_{j\in\mathcal{N}_i} w_{ij}\nabla f_j(\mathrm{x}_j^k), \tag{2.43b}$$

and its aggregate formulation (2.5)

$$\mathrm{x}^{k+1} = \mathcal{W}\mathrm{x}^k - \mathrm{z}^k - \gamma G(\mathrm{x}^k) \tag{2.44a}$$

$$\mathrm{z}^{k+1} = \mathcal{W}\mathrm{z}^k - \gamma(I_{Nn}-\mathcal{W})G(\mathrm{x}^k), \tag{2.44b}$$

Following the arguments proposed in [176] for a centralized optimization algorithm, we interpret the sequences x^k and z_k of (2.44) as sampled versions of two continuous-time signals $x(t)$ and $z(t)$. These signals are assumed to be smooth. According to this

interpretation, the step-size $\gamma > 0$ can be then seen as the sampling time characterizing a discretization procedure. Figure 2.5 shows a graphical representation of the discretization of the continuous signal $x(t)$ resulting in the discrete sequence x_k.

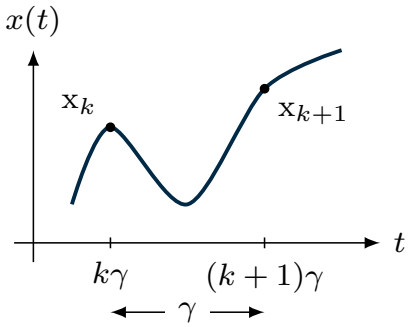


Figure 2.5: x_k as sampled version of $x(t)$.

Informally, we start from the intuition

$$\mathrm{x}_k \approx x(t)\Big|_{t=k\gamma}, \qquad \mathrm{z}_k \approx z(t)\Big|_{t=k\gamma},$$

in which the iteration index k is obtained by setting $k := t/\gamma$ with t being the continuous time. For any fixed t, by choosing an arbitrarily small step-size γ, we can consider the following approximations

$$x(t) \approx \mathrm{x}_{t/\gamma} = \mathrm{x}_k, \quad x(t+\gamma) \approx \mathrm{x}_{(t+\gamma)/\gamma} = \mathrm{x}_{k+1}.$$

The same clearly holds also for the sequence z_k. With these approximations in mind, we can write the following Taylor expansions

$$\mathrm{x}_{k+1} = x(t)\Big|_{t=(k+1)\gamma} = x(t)\Big|_{t=k\gamma} + \gamma\dot{x}(t)\Big|_{t=k\gamma} + o(\gamma) \tag{2.45a}$$

$$\mathrm{z}_{k+1} = z(t)\Big|_{t=(k+1)\gamma} = z(t)\Big|_{t=k\gamma} + \gamma\dot{z}(t)\Big|_{t=k\gamma} + o(\gamma), \tag{2.45b}$$

where $o(\gamma)$ collects the higher order terms of the expansions. As γ goes to zero, the higher order terms in (2.45) can be neglected, leading to

$$\dot{x}(t) = \frac{1}{\gamma}(\mathrm{x}_{k+1} - \mathrm{x}_k) = \frac{\mathcal{W} - I_{Nn}}{\gamma}x(t) - z(t) - G(x(t))$$
$$\dot{z}(t) = \frac{1}{\gamma}(\mathrm{z}_{k+1} - \mathrm{z}_k) = \frac{\mathcal{W} - I_{Nn}}{\gamma}z(t) - \frac{\mathcal{W} - I_{Nn}}{\gamma}G(x(t)),$$

which can be rewritten as

$$\dot{x}(t) = -L_\gamma x(t) - z(t) - G(x(t)) \tag{2.46a}$$

$$\dot{z}(t) = -L_\gamma z(t) - L_\gamma G(x(t)). \tag{2.46b}$$

where $L_\gamma := (I_{Nn} - \mathcal{W})/\gamma$.

2.4.1 Continuous Gradient Tracking: Algorithm Description and Analysis

The Ordinary Differential Equation (ODE) in (2.46) involves matrices whose structure depends on the preceding derivation. However, one may consider any weighted Laplacian matrix $\mathcal{L}$ associated to $\mathcal{G}$. Therefore, we define the Continuous Gradient Tracking as

$$\begin{bmatrix}\dot{x}(t)\\ \dot{z}(t)\end{bmatrix} = \begin{bmatrix}-L & -I_{Nn}\\ 0 & -L\end{bmatrix}\begin{bmatrix}x(t)\\ z(t)\end{bmatrix} - \begin{bmatrix}I_{Nn}\\ L\end{bmatrix} G(x(t)), \tag{2.47}$$

where $L \in \mathbb{R}^{Nn\times Nn}$ is given by $L := \mathcal{L}\otimes I_n$.

It is useful to also provide a local view of algorithm (2.47), i.e., from the perspective of a generic agent i. The i-th block-components of $x(t)$ and $z(t)$ corresponds, respectively, to the local states $x_i(t)$ and $z_i(t)$ of agent i. The state $x_i(t)$ represents the local estimate at time t of the optimal solution of problem (2.1) while $z_i(t) \in \mathbb{R}^n$ is an auxiliary state. Set $\mathcal{W}_\mathcal{G} := \mathcal{D} - \mathcal{L}$ (with $\mathcal{D}$ being the degree matrix of $\mathcal{G}$) and let w_{ij} be its (i,j)-th entry. Exploiting the sparsity in $\mathcal{W}_\mathcal{G}$, the i-th block-components of (2.47) can be then written as

$$\dot{x}_i(t) = -\sum_{j\in\mathcal{N}_i} w_{ij}\big(x_i(t) - x_j(t)\big) - z_i(t) - \nabla f_i(x_i(t)) \tag{2.48a}$$

$$\dot{z}_i(t) = -\sum_{j\in\mathcal{N}_i} w_{ij}\big(z_i(t) - z_j(t)\big) - \sum_{j\in\mathcal{N}_i} w_{ij}\big(\nabla f_i(x_i(t)) - \nabla f_j(x_j(t))\big). \tag{2.48b}$$

In the next, we will analyze Continuous Gradient Tracking through a Lyapunov approach relying on the feedback structure of (2.47) as represented in Figure 2.6.

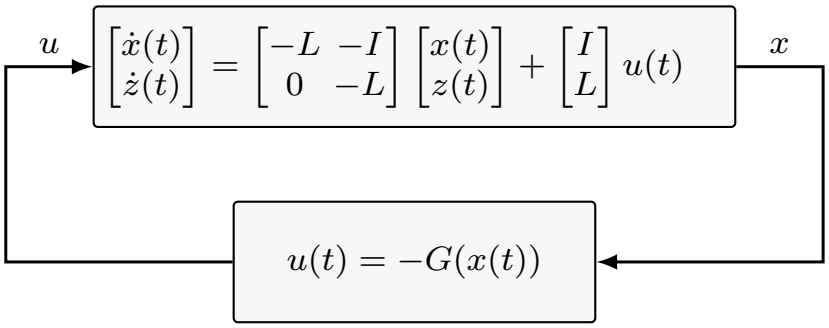


Figure 2.6: Block diagram representation of system (2.47).

Specifically, noticing that $\text{col}(\mathbf{1}_{N,n}x^\star, -G(\mathbf{1}_{N,n}x^\star))$ is the unique equilibrium point

for system (2.47) and by exploiting the initialization, it is possible to perform a sequence of coordinate changes to obtain an equivalent, reduced system formulation. The equivalent system is characterized by a (marginally stable) linear part, associated to the consensus mechanism, which is perturbed by a nonlinear (feedback) term depending on the gradient G. The next theorem proves the exponential stability of the equilibrium by designing a quadratic Lyapunov function based on the linear part only and bounding the nonlinear gradient term using the strong convexity and the Lipschitz continuity. We point out that the initialization $\mathbf{1}_{N,n}^\top z(0) = 0$ can be obtained in a fully distributed way by simply setting each $z_i(0) = 0$, for all $i \in \{1, \dots, N\}$.

Theorem 2.3. *Consider the Continuous Gradient Tracking distributed algorithm described by* (2.47)*. Let Assumptions 2.3, 2.4, 2.5 hold and pick any* $\textsc{col}(x(0), z(0))$ *such that* $\mathbf{1}_{N,n}^\top z(0) = 0$*. Then, there exist* $a_1 > 0$ *and* $a_2 > 0$ *such that*

$$\|x_i(t) - x^\star\| \leq a_1 \exp(-a_2 t), \quad \forall i \in \{1, \dots, N\}. \qquad \triangle$$

Proof. We first observe that, in light of Assumption 2.4, the ODE in (2.47) is well posed for any $t \geq 0$ and admits a unique solution, see [91, Theorem 3.2]. By inspecting system (2.47), we can assert that it has a unique equilibrium point at

$$\begin{bmatrix} \mathrm{x}_{\mathrm{eq}} \\ \mathrm{z}_{\mathrm{eq}} \end{bmatrix} := \begin{bmatrix} \mathbf{1}_{N,n} x^\star \\ -G(\mathbf{1}_{N,n} x^\star) \end{bmatrix},$$

which represents the situation in which the N agents have a consensual solution estimate equal to the optimal solution $x^\star$ of the optimization problem (2.1). In order to use a Lyapunov approach, we put the system in error coordinates. Let

$$\begin{bmatrix} x \\ z \end{bmatrix} \longmapsto \begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} := \begin{bmatrix} x \\ z \end{bmatrix} - \begin{bmatrix} \mathrm{x}_{\mathrm{eq}} \\ \mathrm{z}_{\mathrm{eq}} \end{bmatrix}. \tag{2.49}$$

Then, system (2.47) can be rewritten as

$$\begin{bmatrix} \dot{\tilde{x}} \\ \dot{\tilde{z}} \end{bmatrix} = \begin{bmatrix} -L & -I \\ 0 & -L \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} + \begin{bmatrix} I \\ L \end{bmatrix} u(\tilde{x}), \tag{2.50}$$

where the role played by the "input" term $u(\tilde{x}) := G(\mathbf{1}_{N,n} x^\star) - G(\tilde{x} + \mathbf{1}_{N,n} x^\star)$ has been highlighted. Indeed, it can be interpreted as a nonlinear feedback of the output $\tilde{y} = \tilde{x}$ and suggests to introduce a further change of coordinates given by

$$\begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} \longmapsto \begin{bmatrix} \tilde{y} \\ \tilde{\eta} \end{bmatrix} := \underbrace{\begin{bmatrix} I & 0 \\ L & -I \end{bmatrix}}_{T_1} \begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix}. \tag{2.51}$$

Since T_1 is an involutory matrix (i.e., it coincides with its inverse), the change of coordinates (2.51) transforms (2.50) in

$$\begin{bmatrix}\dot{\tilde{y}}\\ \dot{\tilde{\eta}}\end{bmatrix} = \begin{bmatrix}-2L & I\\ -L^2 & 0\end{bmatrix}\begin{bmatrix}\tilde{y}\\ \tilde{\eta}\end{bmatrix} + \begin{bmatrix}I\\ 0\end{bmatrix} u(\tilde{y}). \tag{2.52}$$

Before studying the stability of the origin for (2.52), the effect of the initialization $\mathbf{1}_{N,n}^\top z(0) = 0$ in the new coordinates $(\tilde{y}, \tilde{\eta})$ is investigated. We observe that the subspace

$$\mathcal{S} := \{(\tilde{y}, \tilde{\eta}) \mid \mathbf{1}_{N,n}^\top \tilde{\eta} = 0\}$$

is invariant for (2.52). In light of (2.51), it holds

$$0 = \mathbf{1}_{N,n}^\top \tilde{\eta} = \mathbf{1}_{N,n}^\top (L\tilde{x} - \tilde{z}) = \mathbf{1}_{N,n}^\top z,$$

where the last equality holds in light of (2.49) and since $\mathbf{1}_{N,n}^\top G(\mathbf{1}_{N,n} x^\star) = 0$ and $\mathbf{1}_{N,n}^\top L = 0$ (cf. Assumption 2.5). Therefore the initialization of $z(0)$ guarantees that $\tilde{\eta}(0) \in \mathcal{S}$. Hence, we can perform a final change of coordinates to isolate the invariant state to further restrict the dynamics. Let

$$\begin{bmatrix}\tilde{y}\\ \tilde{\eta}\end{bmatrix} \longmapsto \begin{bmatrix}\tilde{y}\\ \tilde{\psi}\\ \tilde{\eta}_{\text{avg}}\end{bmatrix} := T_2 \begin{bmatrix}\tilde{y}\\ \tilde{\eta}\end{bmatrix}, \tag{2.53}$$

in which

$$T_2 := \begin{bmatrix}T_{\tilde{y}}\\ T_{\tilde{\eta}}\end{bmatrix},\ T_{\tilde{y}} := \begin{bmatrix}I & 0\\ 0 & R^\top\end{bmatrix},\ T_{\tilde{\eta}} := \begin{bmatrix}0 & \frac{1}{\sqrt{N}}\mathbf{1}_{N,n}^\top\end{bmatrix}, \tag{2.54}$$

with $R \in \mathbb{R}^{Nn\times(N-1)n}$ such that $R^\top R = I$, $R^\top \mathbf{1}_{N,n} = 0$ and $\|R\| = 1$. The following useful relations holds true

$$RR^\top = I - \frac{1}{N}\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top. \tag{2.55}$$

It is easy to check that $T_2^{-1} = T_2^\top$, thus (2.52) can be rewritten as

$$\begin{bmatrix}\dot{\tilde{y}}\\ \dot{\tilde{\psi}}\\ \dot{\tilde{\eta}}_{\text{avg}}\end{bmatrix} = \begin{bmatrix}-2L & R & \frac{\mathbf{1}_{N,n}}{\sqrt{N}}\\ -R^\top L^2 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\tilde{y}\\ \tilde{\psi}\\ \tilde{\eta}_{\text{avg}}\end{bmatrix} + \begin{bmatrix}I\\ 0\\ 0\end{bmatrix} u(\tilde{y}). \tag{2.56}$$

In light of the invariance of $\mathcal{S}$, it holds $\tilde{\eta}_{\text{avg}}(t) \equiv 0$. Then we can consider only $\zeta :=$

$\text{col}(\tilde{y}, \tilde{\psi}) \in \mathbb{R}^d$, with $d := (2N-1)n$. The dynamics (2.56) can be written as

$$\dot{\zeta} = A\zeta + Bu(\tilde{y}), \tag{2.57a}$$

with

$$A := \begin{bmatrix} -2L & R \\ -R^\top L^2 & 0 \end{bmatrix}, \quad B := \begin{bmatrix} I \\ 0 \end{bmatrix}. \tag{2.58}$$

Next, consider a quadratic, candidate Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}$ given by

$$V(\zeta) := \zeta^\top P \zeta, \tag{2.59}$$

with $P \in \mathbb{R}^{d\times d}$ such that $P = P^\top > 0$ and arranged in blocks as

$$P := \begin{bmatrix} P_1 & P_2 \\ P_2^\top & P_3 \end{bmatrix}, \tag{2.60}$$

where $P_1 \in \mathbb{R}^{Nn\times Nn}$, $P_2 \in \mathbb{R}^{Nn\times n}$, and $P_3 \in \mathbb{R}^{(Nn-n)\times(Nn-n)}$. Next, it is shown how to choose P in order to prove global exponential stability of the origin of (2.57). Let $m > 0$ and set

$$P_1 = mI, \quad P_2 = -R, \quad P_3 = mR^\top (L^2)^\dagger R, \tag{2.61}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse. By the Schur complement lemma, $P > 0$ imposes that m must satisfy

$$\begin{cases} mI > 0 \\ mR^\top (L^2)^\dagger R - \frac{1}{m} I > 0 \end{cases} \implies m > \frac{1}{\sqrt{\min\{\sigma(R^\top (L^2)^\dagger R)\}}}. \tag{2.62}$$

The time-derivative of V along trajectories of (2.57) is

$$\dot{V}(\zeta) = \zeta^\top \underbrace{(A^\top P + PA)}_{-Q} \zeta + 2\zeta^\top PBu. \tag{2.63}$$

The choices (2.61) yield to

$$Q = \begin{bmatrix} 4mL - 2L^2 & 2LR \\ 2R^\top L & 2I \end{bmatrix}, \quad PB = \begin{bmatrix} mI \\ -R^\top \end{bmatrix}. \tag{2.64}$$

We separately study the quadratic term $-\zeta^\top Q \zeta$ and the cross term $2\zeta^\top PBu$ as a function of m to show that $\dot{V}(\zeta)$ can be made negative definite for a sufficiently large m.

As for the first term in (2.63), we observe that Q is a solution to a Lyapunov equation

associated to a marginally stable matrix. Therefore, it can only be positive semidefinite. Indeed, the upper-left block within the expression (2.64) has the kernel spanned by $\mathbf{1}_{N,n}$ for any choice of m. By the Schur complement lemma, imposing $Q \geq 0$ is equivalent to

$$4mL - 2L^2 - 2LRR^\top L \geq 0. \tag{2.65}$$

In light of (2.55) and since $L\mathbf{1}_{N,n} = 0$, condition (2.65) reduces to $4mL - 4L^2 \geq 0$. Since L and L^2 have the same kernel, the latter condition is fulfilled by any m such that

$$m \geq \frac{\max\{\sigma(L^2)\}}{\min\{\sigma(L) \setminus \{0\}\}}. \tag{2.66}$$

Moreover, L positive semidefinite, condition (2.66) can be satisfied with $m > 0$.

Next, the second term in (2.63) is considered to show $\dot{V} < 0$. In light of (2.64), it holds

$$2\zeta^\top PBu = 2m\tilde{y}^\top u - 2\tilde{\psi}^\top R^\top u \overset{(a)}{\leq} -2m\mu \left\|\tilde{y}\right\|^2 - 2\tilde{\psi}^\top R^\top u, \tag{2.67}$$

where in (a) we use the strong convexity of the cost functions (cf. Assumption 2.3). Using the Cauchy-Schwarz inequality, condition (2.67) can be manipulated as

$$\begin{aligned} 2\zeta^\top PBu &\leq -2m\mu \left\|\tilde{y}\right\|^2 + 2\left\|R\right\| \left\|\tilde{\psi}\right\| \left\|u\right\| \\ &\overset{(a)}{\leq} -2m\mu \left\|\tilde{y}\right\|^2 + 2\overline{L} \left\|\tilde{\psi}\right\| \left\|\tilde{y}\right\| \\ &\overset{(b)}{\leq} -2m\mu \left\|\tilde{y}\right\|^2 + \frac{\overline{L}}{\epsilon} \left\|\tilde{y}\right\|^2 + \overline{L}\epsilon \left\|\tilde{\psi}\right\|^2 \\ &\overset{(c)}{=} \zeta^\top \underbrace{\begin{bmatrix} \left(-2m\mu + \frac{\overline{L}}{\epsilon}\right) I & 0 \\ 0 & \overline{L}\epsilon I \end{bmatrix}}_{Q_0} \zeta, \end{aligned} \tag{2.68}$$

where in (a) we use the Lipschitz continuity of the gradient of the cost functions (cf. Assumption 2.4) and the fact that $\|R\| = 1$, while in (b) we use the Young's inequality with $\epsilon > 0$, and in (c) we introduce the matrix Q_0. Indeed, we want to show that the zero eigenvalues of Q can be moved inside the open left-half plane through Q_0. Thus, by plugging (2.68) in (2.63), it holds

$$\dot{V}(\zeta) \leq -\zeta^\top \tilde{Q} \zeta, \tag{2.69}$$

where $\tilde{Q} := Q - Q_0$, i.e., it holds

$$\tilde{Q} := \begin{bmatrix} 4mL - 2L^2 + \left(2m\mu - \frac{\overline{L}}{\epsilon}\right) I & 2LR \\ 2R^\top L & (2 - \overline{L}\epsilon) I \end{bmatrix}. \tag{2.70}$$

By the Schur complement lemma, $\tilde{Q} > 0$ is equivalent to

$$\begin{cases} 2 - \overline{L}\epsilon > 0 \\ 4mL - 2\left(\frac{2}{2-\overline{L}\epsilon} + 1\right) L^2 + \frac{2m\mu\epsilon - \overline{L}}{\epsilon} I > 0, \end{cases}$$

which is verified for every $\epsilon < \frac{2}{\overline{L}}$ and

$$m > \max\left\{ \frac{\left(1 + \frac{1}{2-\overline{L}\epsilon}\right) \max\{\sigma(L^2)\}}{2\min\{\sigma(L) \setminus \{0\}\}}, \frac{\overline{L}}{2\mu\epsilon} \right\}. \tag{2.71}$$

Therefore, we can conclude that $\dot{V}(\zeta) < -\min\{\sigma(\tilde{Q})\} \|\zeta\|^2$ which implies that the origin is globally exponentially stable for system (2.57) (cf. [91, Theorem 4.10]). Specifically, there exist $a_7, a_2 > 0$ such that

$$\|\zeta(t)\| \leq a_7 \|\zeta(0)\| \exp(-a_2 t), \tag{2.72}$$

for any $\zeta(0) \in \mathbb{R}^d$. By noticing that $\|x_i(t) - x^\star\| \leq \|x(t) - \mathbf{1}_{N,n} x^\star\| = \|y(t)\| \leq \|\zeta(t)\|$, the proof follows by (2.72) by setting $a_1 = a_7 \|\zeta(0)\|$. ■

We underline that both a_1 and a_2 in Theorem 2.3 depend on *(i)* the distance between the initial conditions and the system equilibrium and *(ii)* the problem properties as, e.g., the network connectivity, the strong convexity parameter of the cost and the Lipschitz constants of the cost gradients. The same observation consistently applies to the subsequent results.

Remark 2.3. The expression of the discrete-time dynamics as in (2.44) turns out to be crucial in the derivation of its continuous-time version. In fact, one can check that

$$\begin{aligned} G(\mathrm{x}_{k+1}) &= G\left(x(t)\Big|_{t=(k+1)\gamma} \right) \\ &= G\left(x(t)\Big|_{t=k\gamma} + \gamma \dot{x}(t)\Big|_{t=k\gamma} + o(\gamma) \right) \\ &= G\left(x(t)\Big|_{t=k\gamma} \right) + \gamma \nabla G\left(x(t)\Big|_{t=k\gamma} \right) \dot{x}(t)\Big|_{t=k\gamma} + o(\gamma). \end{aligned}$$

Thus, the arguments presented to derive (2.47) applied to the algorithm in its original

coordinates (2.3) results in

$$\begin{bmatrix}\dot{x}(t)\\ \dot{s}(t)\end{bmatrix} = \begin{bmatrix}-L_\gamma & -I_{Nn}\\ -\nabla G(x(t))L_\gamma & -(L_\gamma + \nabla G(x(t)))\end{bmatrix}\begin{bmatrix}x(t)\\ s(t)\end{bmatrix},$$

where $s(t)$ is the continuous counterpart of $\mathrm{s}_k := \mathrm{col}(\mathrm{s}_1^k, \dots, \mathrm{s}_N^k)$ while $\nabla G(x(t)) := \mathrm{blkdiag}(\nabla^2 f_1(x_1(t)), \dots, \nabla^2 f_N(x_N(t))) \in \mathbb{R}^{Nn\times Nn}$. Notice that these coordinates involve the second-order matrix $\nabla G(\cdot)$ which usually requires a non-negligible computational complexity and, in certain applications, is not even known. △

2.4.2 Triggered Gradient Tracking: Algorithms Description and Analysis

The Continuous Gradient Tracking would require communication among agents at all $t \geq 0$. Clearly, this prevents its practical implementation on real devices that require time-slotted communication. This issue is addressed next by proposing two alternative schemes in which inter-agent communication is triggered synchronously and asynchronously, respectively.

Let $\{t_i^{k_i}\}_{k_i\in\mathbb{N}}$, be the sequence of time instants at which agent i sends its states (x_i, z_i) and ∇f_i to its neighbors $j \in \mathcal{N}_i$. Consistently, at time $t_j^{k_j}$, agent i receives the updated variables from its neighbor $j \in \mathcal{N}_i$. Let $\{\tilde{t}^k\}_{k\geq 0}$ be the ordered sequence of all the triggering times that occurred in the network. Then, given any $t \in [\tilde{t}^k, \tilde{t}^{k+1})$, let us introduce, for all $i \in \{1, \dots, N\}$, the shorthands

$$\begin{aligned}\hat{x}_i^k &:= x_i(t)\Big|_{t=\inf_{k_i\in\mathbb{N}}\left\{t_i^{k_i}\geq\tilde{t}^k\right\}}\\ \hat{z}_i^k &:= z_i(t)\Big|_{t=\inf_{k_i\in\mathbb{N}}\left\{t_i^{k_i}\geq\tilde{t}^k\right\}}\\ \nabla f_i^k &:= \nabla f_i(x_i(t))\Big|_{t=\inf_{k_i\in\mathbb{N}}\left\{t_i^{k_i}\geq\tilde{t}^k\right\}}.\end{aligned} \tag{2.73}$$

Quantities in (2.73) represent the most updated values in the network within the considered time interval. Under the described communication paradigm, we propose to modify the local dynamics in (2.48) as follows

$$\dot{x}_i(t) = -\sum_{j\in\mathcal{N}_i} w_{ij}\left(\hat{x}_i^k - \hat{x}_j^k\right) - z_i(t) - \nabla f_i(x_i(t)) \tag{2.74a}$$

$$\dot{z}_i(t) = -\sum_{j\in\mathcal{N}_i} w_{ij}\left(\hat{z}_i^k - \hat{z}_j^k\right) - \sum_{j\in\mathcal{N}_i} w_{ij}\left(\nabla f_i^k - \nabla f_j^k\right), \tag{2.74b}$$

for all $t \in [\tilde{t}^k, \tilde{t}^{k+1})$. Within the k-th period, the variable z_i behaves as an integrator. As for the variable x_i, it is a local gradient flow compensated with an integral action z_i and a constant consensus-error-like term. Agent i does not use its own variables $x_i(t)$,

$z_i(t)$, and $\nabla f_i(x_i(t))$ in the consensus mixing terms, but it rather uses their sampled version. This fact allows one to preserve the theoretical consensus properties of the original scheme (2.48).

As one can expect, the specific rule to choose the triggering time $t_i^{k_i}$ will play a crucial role in the convergence properties of the resulting algorithms.

Synchronous Triggered Gradient Tracking

We start by presenting Synchronous Triggered Gradient Tracking, obtained by imposing a *synchronous communication* among agents. Specifically, in this protocol each agent $i \in \{1, \dots, N\}$ sends its local variables to its neighbors at common instants of time chosen according to

$$t_i^{k_i+1} := t_i^{k_i} + \Delta, \tag{2.75}$$

for some common $\Delta > 0$ and with $t^{0_i} = t_0$ for all $i \in \{1, \dots, N\}$. Intuitively, the greater Δ, the more inter-agent communication reduces. On the other hand, the greater Δ, the more the triggered algorithmic evolution moves away from the behavior of Continuous Gradient Tracking.

For all $t \in [\tilde{t}^k, \tilde{t}^{k+1})$, the aggregate form of (2.74) reads as

$$\begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = H \begin{bmatrix} x(t) \\ z(t) \end{bmatrix} + B_1 G(x(t)) + B_2 \begin{bmatrix} \hat{x}^k \\ \hat{z}^k \\ G^k \end{bmatrix}, \tag{2.76}$$

where $\hat{x}^k := \text{col}(\hat{x}_1^k, \dots, \hat{x}_N^k)$, $\hat{z}^k := \text{col}(\hat{z}_1^k, \dots, \hat{z}_N^k)$, $G^k := \text{col}(\nabla f_1^k, \dots, \nabla f_N^k)$, and

$$H := \begin{bmatrix} 0 & -I \\ 0 & 0 \end{bmatrix}, \quad B_1 := \begin{bmatrix} -I \\ 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} -L & 0 & 0 \\ 0 & -L & -L \end{bmatrix}.$$

The convergence properties of (2.76) can be shown by reformulating it as a perturbed instance of the Continuous Gradient Tracking system (2.47). In particular, it is possible to show that the periodic triggering law (2.75) gives rise to a perturbation term that vanishes at the equilibrium point (see, e.g., [91, Chapter 9] for the notion of vanishing perturbation) and that can be arbitrarily bounded through the parameter Δ. Based on this observation, the next thereom considers the same Lyapunov function V used in the proof of Theorem 2.3 and shows that, with a sufficiently small Δ, the perturbation does not alter the sign of the derivative of V and, hence, the exponential convergence is preserved. The next theorem formalizes this result.

Theorem 2.4. *Consider the algorithm in* (2.74) *with the synchronous communication protocol given by* (2.75)*. Let Assumptions 2.3, 2.4, 2.5 hold and pick any* $\text{col}(x(0), z(0))$ *such that*

$\mathbf{1}_{N,n}^\top z(0) = 0$. *Then, there exist* $\Delta^\star > 0$, $a_3 > 0$, *and* $a_4 > 0$ *such that for any* $\Delta \in (0, \Delta^\star)$, *it holds*

$$\|x_i(t) - x^\star\| \leq a_3 \exp(-a_4 t), \quad \forall i \in \{1, \dots, N\}. \qquad \triangle$$

Proof. The dynamics (2.74) of Synchronous Triggered Gradient Tracking can be reformulated as a perturbed instance of the nominal dynamics of Continuous Gradient Tracking described by (2.47). Clearly, the perturbation expresses the impact of the triggering mechanism on the algorithmic evolution. Thus, by adding and subtracting the term $B_2 \textsc{col}(x(t), z(t), G(x(t)))$ in the dynamics (2.76), we get

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} -L & -I \\ 0 & -L \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} -I \\ -L \end{bmatrix} G(x) + B_2 e, \tag{2.77}$$

where e has the same meaning as in (2.114). By performing the same changes of coordinates defined in (2.49), (2.51), and (2.53), the dynamics (2.77) can be equivalently reformulated as the following (restricted) dynamics

$$\dot{\zeta} = A\zeta + Bu + E e_{\zeta,\nabla}, \tag{2.78}$$

where the vectors $\zeta \in \mathbb{R}^d$, $u \in \mathbb{R}^{Nn}$ and the matrices $A \in \mathbb{R}^{d \times d}$ and $B \in \mathbb{R}^{d \times Nn}$ are as in (2.57), while the quantities associated to the perturbation are

$$E := T_{\tilde{y}}^\top T_1 B_2 T_1 T_{\tilde{y}} = \begin{bmatrix} -L & 0 & 0 \\ 0 & -R^\top L R & R^\top L \end{bmatrix}, \tag{2.79a}$$

$$e_{\zeta,\nabla} := \textsc{col}(\hat{\tilde{y}} - \tilde{y}, \hat{\tilde{\psi}} - \tilde{\psi}, e_\nabla) := \textsc{col}(\hat{\tilde{y}} - \tilde{y}, \hat{\tilde{\psi}} - \tilde{\psi}, G^k - G(\tilde{y} + x^\star)) \tag{2.79b}$$

with T_1 and $T_{\tilde{y}}$ defined in (2.51) and (2.54), respectively. Let us consider a quadratic, candidate Lyapunov function $V(\zeta) = \zeta^\top P \zeta$ as in (2.59) with the blocks of P set as in (2.61). The time-derivative of V along the trajectories of (2.77) satisfies

$$\begin{aligned} \dot{V}(\zeta) &= \zeta^\top (A^\top P + PA)\zeta + 2\zeta^\top PBu + 2\zeta^\top P E e_{\zeta,\nabla} \\ &\leq -\zeta^\top \tilde{Q} \zeta + 2\zeta^\top P E e_{\zeta,\nabla}, \end{aligned} \tag{2.80}$$

where $\tilde{Q}$ is as in (2.70) so that the inequality holds in light the previous proof of Theorem 2.3 (cf. (2.69)). By using the Young's inequality with $\epsilon > 0$, we can further upper bound (2.80) as

$$\begin{aligned} \dot{V}(\zeta) &\leq -\zeta^\top \tilde{Q} \zeta + \epsilon \zeta^\top P P \zeta + \frac{1}{\epsilon} e_{\zeta,\nabla}^\top E^\top E e_{\zeta,\nabla} \\ &\overset{(a)}{=} -\zeta^\top \Big(\tilde{Q} - \epsilon P^2 \Big) \zeta + \frac{1}{\epsilon} e_{\zeta,\nabla}^\top E^\top E e_{\zeta,\nabla}, \end{aligned} \tag{2.81}$$

where in (a) the terms in ζ have been grouped. In light of the sufficient condition in (2.71) to get a positive definite $\tilde{Q}$, we can always take ϵ such that

$$0 < \epsilon < \frac{\min\{\sigma(\tilde{Q})\}}{\max\{\sigma(P^2)\}}$$

in order to impose also $\tilde{Q} - \epsilon P^2$ positive definite. Thus, by denoting as $q > 0$ the smallest eigenvalue of the matrix $\tilde{Q} - \epsilon P^2$ and by applying the Cauchy-Schwarz inequality to the quadratic term in $e_{\zeta,\nabla}$ of (2.81), we bound (2.81) as

$$\begin{aligned}
\dot{V}(\zeta) &\leq -q \left\|\zeta\right\|^2 + \frac{1}{\epsilon} \left\|E^\top E\right\| \left\|e_{\zeta,\nabla}\right\|^2 \\
&\overset{(a)}{=} -q \left\|\zeta\right\|^2 + \frac{1}{\epsilon} \left\|E^\top E\right\| \left(\left\|e_\zeta\right\|^2 + \left\|e_\nabla\right\|^2\right) \\
&\overset{(b)}{\leq} -q \left\|\zeta\right\|^2 + \frac{1}{\epsilon} \left\|E^\top E\right\| \left(\left\|e_\zeta\right\|^2 + \overline{L}^2 \left\|\hat{\tilde{y}} - \tilde{y}\right\|^2 \right) \\
&\overset{(c)}{\leq} -q \left\|\zeta\right\|^2 + \underbrace{\frac{1}{\epsilon} \left\|E^\top E\right\| (1 + \overline{L}^2)}_{c_1} \left\|e_\zeta\right\|^2 ,
\end{aligned} \tag{2.82}$$

where in (a) we introduce $e_\zeta := \text{col}(\hat{\tilde{y}} - \tilde{y}, \hat{\tilde{\psi}} - \tilde{\psi})$ to write $\left\|e_{\zeta,\nabla}\right\|^2 = \left\|e_\zeta\right\|^2 + \left\|e_\nabla\right\|^2$, in (b) we use the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.4) to bound $\left\|e_\nabla\right\|^2 \leq \overline{L}^2 \|\hat{\tilde{y}} - \tilde{y}\|^2$, and in (c) we rely on the fact that $\hat{\tilde{y}} - \tilde{y}$ is a component of e_ζ.

The proof continues by deriving an upper bound for $\left\|e_\zeta\right\|^2$ in (2.82). We start by defining

$$r := \frac{\left\|e_\zeta\right\|}{\left\|\zeta\right\|}. \tag{2.83}$$

Moreover, recall that in each interval $[\tilde{t}^k, \tilde{t}^{k+1})$, the error e_ζ is set to zero at $\tilde{t}^k$ and grows until $\tilde{t}^{k+1}$ when it is reset again to zero. Hence, the goal is to establish a lower bound on the needed time for $r(t)$ to reach $\sqrt{q/c_1}$. By computing the time derivative of (2.83), it follows

$$\dot{r} = \frac{e_\zeta^\top \dot{e}_\zeta}{\left\|e_\zeta\right\| \left\|\zeta\right\|} - \frac{\left\|e_\zeta\right\| \zeta^\top \dot{\zeta}}{\left\|\zeta\right\|^3}. \tag{2.84}$$

Using the Cauchy-Schwarz inequality, we bound $\dot{r}$ as

$$\begin{aligned}
\dot{r} &\leq \frac{\left\|e_\zeta\right\| \left\|\dot{e}_\zeta\right\|}{\left\|e_\zeta\right\| \left\|\zeta\right\|} + \frac{\left\|e_\zeta\right\| \left\|\zeta\right\| \left\|\dot{\zeta}\right\|}{\left\|\zeta\right\|^3} \\
&\overset{(a)}{\leq} \frac{\left\|\dot{\zeta}\right\|}{\left\|\zeta\right\|} + \frac{\left\|e_\zeta\right\| \left\|\dot{\zeta}\right\|}{\left\|\zeta\right\|^2} \overset{(b)}{=} (1 + r) \frac{\left\|\dot{\zeta}\right\|}{\left\|\zeta\right\|}
\end{aligned} \tag{2.85}$$

where in (a) we use the identity $\dot{e}_\zeta = -\dot{\zeta}$ while in (b) we exploit the definition of r in (2.83). Then, in light of the dynamics of ζ in (2.78), it holds

$$\begin{aligned}\dot{r} &\leq (1+r)\frac{\|A\zeta + Bu + Ee_{\zeta,\nabla}\|}{\|\zeta\|} \\ &\overset{(a)}{\leq} (1+r)\frac{\|A\|\,\|\zeta\| + \|u\| + \|E\|\,\|e_\zeta\| + \|E\|\,\|e_\nabla\|}{\|\zeta\|},\end{aligned} \tag{2.86}$$

where in (a) we use the triangle and the Cauchy-Schwarz inequalities combined with $\|Bu\| = \|u\|$. Next, by using the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.4), we have $\|u\| \leq \overline{L}\,\|\zeta\|$ and $\|e_\nabla\| \leq \overline{L}\,\|e_\zeta\|$. Thus, (2.86) becomes

$$\begin{aligned}\dot{r} &\leq (1+r)\frac{(\|A\| + \overline{L})\,\|\zeta\| + (1+\overline{L})\,\|E\|\,\|e_\zeta\|}{\|\zeta\|} \\ &\overset{(a)}{=} (1+r)\frac{\overline{L}\,\|\zeta\|}{\|\zeta\|} + (1+r)\frac{\|A\|\,\|\zeta\| + (1+\overline{L})\,\|E\|\,\|e_\zeta\|}{\|\zeta\|} \\ &\overset{(b)}{=} \overline{L}(1+r) + (1+r)\frac{c_2\,\|\zeta\| + c_2\,\|e_\zeta\|}{\|\zeta\|} \\ &\overset{(c)}{=} \overline{L}(1+r) + c_2(1+r)^2,\end{aligned} \tag{2.87}$$

where in (a) we simply rearrange the terms, in (b) we introduce $c_2 := \max\{\|A\|, (1+\overline{L})\,\|E\|\}$, and in (c) we use the definition of r in (2.83).

Using the Comparison Lemma (see [91, Lemma 3.4]) the bound (2.87) translates in the following inequality

$$r(t, r(\tilde{t}^k)) \leq \bar{r}(t, \bar{r}(\tilde{t}^k)), \tag{2.88}$$

where $r(t, r(\tilde{t}^k))$ denotes the solution of (2.84) with initial condition at $t = t_k$ given by $r(t_k)$ while $\bar{r}(t, \bar{r}(\tilde{t}^k))$ denotes the solution of

$$\dot{\bar{r}}(t) = \overline{L}(1+\bar{r}(t)) + c_2(1+\bar{r}(t))^2, \tag{2.89}$$

for some initial condition initial condition at $t = \tilde{t}^k$ given by $\bar{r}(\tilde{t}^k)$ such that $r(\tilde{t}^k) \leq \bar{r}(\tilde{t}^k)$. Recalling that the protocol (2.75) imposes $r(\tilde{t}^k) = 0$ at the beginning of each time interval $[\tilde{t}^k, \tilde{t}^{k+1})$, then we select $\bar{r}(\tilde{t}^k) = 0$. The solution of (2.89) can be shown to be (cf. [92])

$$\bar{r}(t,0) = \frac{(\overline{L}+c_2)(\exp(\overline{L}(t-\tilde{t}^k)) - 1)}{-c_2\exp(\overline{L}(t-\tilde{t}^k)) + \overline{L} + c_2}. \tag{2.90}$$

Notice that $\bar{r}(t,0)$ starts from 0 at $t = \tilde{t}^k$ and monotonically increases within the interval

$\left[0, t_k + \ln\left(\frac{\overline{L}+c_2}{c_2}\right)/\overline{L}\right)$. Thus, we can always find a triggering value $t = \Delta^\star > 0$ such that $\bar{r}(\Delta^\star, 0) = \sqrt{q/c_1}$. Hence, by choosing any $\Delta \in (0, \Delta^\star)$ in (2.75), the inequality (2.88) ensures

$$|r(t)| = r(t) < \sqrt{\frac{q}{c_1}}, \tag{2.91}$$

for all $t \in [\tilde{t}^k, \tilde{t}^{k+1})$, where the equality holds because r is always positive, see its definition in (2.83). With this result in mind, the inequality (2.82) can be rewritten as

$$\dot{V}(\zeta) \leq -\left(q - \frac{|r|^2}{c_1}\right) \|\zeta\|^2,$$

which allows us to use (2.91) to conclude that the origin is globally exponentially stable for system (2.78) (cf. [91, Th. 4.10]). Specifically, there exist $a_4, a_8 > 0$ such that

$$\|\zeta(t)\| \leq a_8 \|\zeta(0)\| \exp(-a_4 t), \tag{2.92}$$

for any $\zeta(0) \in \mathbb{R}^d$. By noticing that $\|x_i(t) - x^\star\| \leq \|x(t) - \mathbf{1}_{N,n} x^\star\| = \|y(t)\| \leq \|\zeta(t)\|$, the proof follows by (2.92) by setting $a_3 = a_8 \|\zeta(0)\|$. ■

Asynchronous Triggered Gradient Tracking

We now investigate the case in which the agents choose their triggering time $t_i^{k_i}$ in a fully asynchronous way giving rise to an algorithm termed Asynchronous Triggered Gradient Tracking. This scheme is motivated by the fact that the synchronous communication executed according to (2.75) is rather conservative with a consequent non-efficient usage of the available resources. An asynchronous communication protocol allows agents to exchange information only when really needed. This requires a modification of the synchronous scheme. In particular, each agent has to check a local triggering condition and to maintain an additional auxiliary variable. The latter is important to take into account the so-called Zeno behavior. Specifically, an infinite number of triggerings over a finite interval of time must be avoided. Indeed, for agent i, a triggering law suffers from the Zeno effect if

$$\lim_{k_i \to \infty} t_i^{k_i} = \sum_{k_i=0}^{\infty} (t_i^{k_i+1} - t_i^{k_i}) = t_i^\infty,$$

for some (finite) $t_i^\infty > 0$ termed the Zeno time.

The local dynamics is again described by (2.74). But, in order to perform communication only when needed, each agent i chooses the next triggering time instant $t_i^{k_i+1}$

according to a locally verifiable condition. A possible choice for such a condition may be

$$t_i^{k_i+1} := \inf_{t>t_i^{k_i}} \{\|e_i(t)\| > \lambda \|h_i(t)\|\}, \tag{2.93}$$

with $e_i(t) := \text{col}(x_i(t) - \hat{x}_i^k, z_i(t) - \hat{z}_i^k, \nabla f_i(x_i(t)) - \nabla f_i^k)$, $h_i(t) := z_i(t) + \nabla f_i(x_i(t))$, and $\lambda > 0$ a constant to be properly specified later. The rationale for the triggering mechanism is to *(i)* keep the triggered scheme close to the original dynamics (2.47), and *(ii)* avoid the Zeno behavior. To this end, the right-hand side of the inequality within (2.93) must be asymptotically vanishing when the algorithm approaches a steady state. This, in turn, gives rise to a vanishing quantity on the left term of the inequality. Indeed, looking also to the discrete-time version (2.3), the (local) quantity $z_i(t) + \nabla f_i(x_i(t))$ can be seen as a proxy for $\sum_{i=1}^N \nabla f_i(x_i(t))$, i.e., a quantity that vanishes at a consensual optimal solution. However, $\|h_i(t)\|$ vanishes not only when the algorithm approaches the equilibrium, but also if $(x_i(t), z_i(t)) \in \mathcal{S}_i := \{(x_i, z_i) \in \mathbb{R}^{2n} \mid z_i = -\nabla f_i(x_i)\}$, possibly giving rise to the Zeno behavior. Thus, in order to exclude this situation, the triggering condition (2.93) is further modified as

$$t_i^{k_i+1} := \inf_{t>t_i^{k_i}} \Big\{ \|e_i(t)\| > \lambda \|h_i(t)\| + |\xi_i(t)| \Big\}, \tag{2.94}$$

where $\xi_i \in \mathbb{R}$ is a local, auxiliary variable maintained by each agent i evolving as

$$\dot{\xi}_i(t) = -\nu \xi_i(t), \tag{2.95}$$

where $\nu > 0$ is a parameter ruling the decay of $\xi_i(t)$.

As formally shown next, if the ξ_i are initialized to nonzero values, then algorithm (2.74) with triggering law (2.94) does not incur in the Zeno behavior.

As in Theorem 2.4, also the convergence properties of Asynchronous Triggered Gradient Tracking can be shown by properly reformulating its aggregate form (which is still given by (2.76)) as a perturbed instance of the Continuous Gradient Tracking dynamics (2.47) with a vanishing perturbation. For this asynchronous triggering law, (2.94), an upper bound on the perturbation magnitude is provided. It is proportional to *(i)* the term $\lambda \|z(t) + G(x(t))\|$, which, as already stated, represents a surrogate for the distance from the equilibrium point $\text{col}(\mathbf{1}_{N,n} x^\star, -G(\mathbf{1}_{N,n} x^\star))$, and *(ii)* to the exponentially decaying term $\|\xi\|$. Thus, considering a Lyapunov function derived from the one used in Theorem 2.3, it is possible to show that, by picking suitable λ and ν, the perturbation does not affect the sign of the Lyapunov derivative. The next theorem formalizes these concepts.

Theorem 2.5. *Consider the algorithm described by* (2.74) *with the asynchronous communication protocol given by* (2.94). *Let Assumptions 2.3, 2.4, 2.5 hold and pick any*

$\textsc{col}(x(0), z(0), \xi(0))$ such that $\mathbf{1}_{N,n}^\top z(0) = 0$ and with $\xi(0) = \textsc{col}(\xi_1(0), \dots, \xi_N(0)) \neq 0$. Then, there exist $\lambda^\star > 0$, $\nu^\star > 0$, $a_5 > 0$, and $a_6 > 0$ such that for any $\lambda \in (0, \lambda^\star)$ in (2.94) *and any $\nu > \nu^\star$, it holds*

$$\|x_i(t) - x^\star\| \leq a_5 \exp(-a_6 t), \quad \forall i \in \{1, \dots, N\}.$$

Moreover, system (2.74) *does not exhibit the Zeno behavior.* △

Proof.

The proof of Theorem 2.5 traces the same initial steps of the proof of Theorem 2.4. Specifically, we reformulate the Asynchronous Triggered Gradient Tracking as a perturbed, extended version of Continuous Gradient Tracking in which the perturbation is due to the event-triggered communication. By exploiting the steps leading to (2.78), the aggregate form of (2.74) and (2.95) reads

$$\dot{\zeta} = A\zeta + Bu + De \tag{2.96a}$$

$$\dot{\xi} = -\nu\xi, \tag{2.96b}$$

where the vectors $\zeta \in \mathbb{R}^d$, $u \in \mathbb{R}^{Nn}$ and the matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times Nn}$ are as in (2.57), $e \in \mathbb{R}^{3Nn}$ has the same meaning as in (2.114), $\xi := \textsc{col}(\xi_1, \dots, \xi_N) \in \mathbb{R}^N$, while the matrix D is given by

$$D := T_{\tilde{y}}^\top T_1 B_2 = \begin{bmatrix} -L & 0 & 0 \\ -R^\top L^2 & R^\top L & R^\top L \end{bmatrix}, \tag{2.97}$$

where the matrices T_1, $T_{\tilde{y}}$ and B_2 are as in (2.51), (2.54), and (2.76), respectively. We underline that the dynamics of ζ and ξ are decoupled while both quantities affect the triggering law (2.94).

Next, we show how to properly choose the value for ν in (2.95) and for λ in the triggering law (2.94) to guarantee that the perturbation term De and the auxiliary variable ξ do not alter the stability property associated the nominal system $\dot{\zeta} = A\zeta + Bu$ (cf. Theorem 2.3). To this end, an upper bound for $\|De\|$, proportional to $\|\zeta\|$ and $\|\xi\|$, is derived. We start by using the Cauchy-Schwarz inequality to write $\|De\| \leq \|D\| \, \|e\| \leq c_3 \sum_{i=1}^N \|e_i\|$, with $c_3 := \|D\|$. In light of the triggering law (2.94), the latter inequality can be upper bounded as

$$\begin{aligned} \|De\| &\leq \lambda c_3 \sum_{i=1}^N \|z_i + \nabla f_i(x_i)\| + c_3 \sum_{i=1}^N |\xi_i| \\ &\overset{(a)}{\leq} \lambda c_3 \sqrt{N} \, \|z + G(x)\| + c_3 \sqrt{N} \, \|\xi\| \\ &\overset{(b)}{=} \lambda c_4 \, \|\tilde{z} + G(\tilde{x} + \mathbf{1}_{N,n} x^\star) - G(\mathbf{1}_{N,n} x^\star)\| + c_4 \, \|\xi\| \end{aligned}$$

$$\overset{(c)}{\leq} \lambda c_4 \|\tilde{z}\| + \lambda c_4 \overline{L} \|\tilde{x}\| + c_4 \|\xi\|, \tag{2.98}$$

where in (a) we apply the basic algebraic relation $\sum_{i=1}^{N} \|\theta_i\| \leq \sqrt{N} \|\theta\|$ for a vector $\theta = \text{col}(\theta_1, \dots, \theta_N)$, in (b) we perform the change of coordinates given in (2.49) and introduce the constant $c_4 := c_3\sqrt{N}$, and in (c) we use the triangle inequality and the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.4). According to (2.51) and (2.53), it holds

$$\begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} = T_1 T_2^\top \begin{bmatrix} \zeta \\ \tilde{\eta}_{\text{avg}} \end{bmatrix} = T_1 T_2^\top \begin{bmatrix} \zeta \\ 0 \end{bmatrix}, \tag{2.99}$$

where we use the fact that the initialization $z(0)$ leads to $\tilde{\eta}_{\text{avg}}(t) \equiv 0$. We rearrange the inequality (2.98) to reconstruct the term $\|\text{col}(\tilde{x}, \tilde{z})\|$ as

$$\begin{aligned} \|De\| &\leq \lambda c_4 \max\{1, \overline{L}\}\sqrt{2} \|\text{col}(\tilde{x}, \tilde{z})\| + c_4 \|\xi\| \\ &\overset{(a)}{\leq} \lambda c_5 \|\zeta\| + c_4 \|\xi\|, \end{aligned} \tag{2.100}$$

where in (a) we combine (2.99) with the Cauchy-Schwarz inequality and set $c_5 := c_4 \max\{1, \overline{L}\}\sqrt{2} \left\|T_1 T_2^\top\right\|$. Given the linear bound in (2.100), we can pursue a Lyapunov approach to conclude the global exponential stability of the origin. Let us consider a quadratic, candidate Lyapunov function $\tilde{V}(\zeta, \xi) = \zeta^\top P \zeta + \frac{1}{2}\xi^\top \xi$. derived from the one considered in (2.59) with the blocks of P set as in (2.61). Using similar arguments leading to (2.82), the time-derivative of $\tilde{V}$ along trajectories of (2.96) can be upper bounded as

$$\dot{\tilde{V}}(\zeta, \xi) \leq -\tilde{q} \|\zeta\|^2 + 2\zeta^\top P De - \nu \|\xi\|^2. \tag{2.101}$$

By using the Cauchy-Schwarz inequality, we can plug (2.100) in (2.101) to obtain

$$\begin{aligned} \dot{\tilde{V}}(\zeta, \xi) &\leq -\tilde{q} \|\zeta\|^2 + 2 \|\zeta\| \|P\| \|De\| \\ &\leq -\tilde{q}_\lambda \|\zeta\|^2 + 2c_4 \|P\| \|\zeta\| \|\xi\| - \nu \|\xi\|^2, \end{aligned} \tag{2.102}$$

where we introduce $\tilde{q}_\lambda := (\tilde{q} - 2\lambda c_5 \|P\|)$. Then, for any $\lambda < \frac{\tilde{q}}{2c_5} \|P\| =: \lambda^\star$, it holds $\tilde{q}_\lambda > 0$. Setting $c_6 := c_4 \|P\|$, the inequality (2.102) can be arranged in a matrix form as

$$\dot{\tilde{V}}(\zeta, \xi) \leq - \begin{bmatrix} \|\zeta\| \\ \|\xi\| \end{bmatrix}^\top \underbrace{\begin{bmatrix} \tilde{q}_\lambda & -c_6 \\ -c_6 & \nu \end{bmatrix}}_{U} \begin{bmatrix} \|\zeta\| \\ \|\xi\| \end{bmatrix}. \tag{2.103}$$

Being $U \in \mathbb{R}^{2\times 2}$ symmetric, by the Sylvester criterion $U > 0$ if and only if $\tilde{q}_\lambda \nu > c_6^2$. Therefore, by taking any $\nu > \nu^\star := \frac{c_6^2}{\tilde{q}_\lambda}$, the matrix U is positive definite. Thus

the inequality (2.103) guarantees that the origin is globally exponentially stable for system (2.96) (cf. [91, Lemma 4.10]). Specifically, there exist $a_6, a_9 > 0$ such that

$$\|\text{col}(\zeta(t), \xi(t))\| \leq \underbrace{a_9 \|\text{col}(\zeta(0), \xi(0))\|}_{a_5} \exp(-a_6 t), \tag{2.104}$$

for any $\text{col}(\zeta(0), \xi(0)) \in \mathbb{R}^{N+d}$. By noticing that

$$\|x_i(t) - x^\star\| \leq \|x(t) - \mathbf{1}_{N,n} x^\star\| = \|y(t)\| \leq \|\text{col}(\zeta(t), \xi(t))\|,$$

the proof of the first part of the theorem follows by (2.104).

Next, we prove by contradiction that (2.74) does not exhibit the Zeno behavior. Suppose, without loss of generality, that an agent i exhibits the Zeno behavior, namely

$$\lim_{k_i \to \infty} t_i^{k_i} = t_i^\infty. \tag{2.105}$$

For any $k \geq 0$, we have

$$\begin{aligned}
\frac{d}{dt}\|e_i(t)\| &= \frac{e_i^\top \dot{e}_i}{\|e_i(t)\|} \overset{(a)}{\leq} \|\dot{e}_i(t)\| \\
&\overset{(b)}{=} \left\|\text{col}(\dot{x}_i(t), \dot{z}_i(t), \nabla^2 f_i(x_i(t))\dot{x}_i(t))\right\| \\
&\overset{(c)}{=} \left\|\text{col}(\dot{\tilde{x}}_i(t), \dot{\tilde{z}}_i(t), \nabla^2 f_i(\tilde{x}_i(t) + x^\star)\dot{\tilde{x}}_i(t))\right\|,
\end{aligned} \tag{2.106}$$

where in (a) we use the Cauchy-Schwarz inequality, in (b) we use the definition of $e_i(t)$, and in (c) we locally perform the change of variables given in (2.49). Combining the latter change of variables with (2.74), it holds

$$\dot{\tilde{x}}_i(t) = -\sum_{j \in \mathcal{N}_i} w_{ij}(\hat{\tilde{x}}_i^k - \hat{\tilde{x}}_j^k) - \tilde{z}_i(t) + u_i(\tilde{x}_i(t)) \tag{2.107a}$$

$$\dot{\tilde{z}}_i(t) = -\sum_{j \in \mathcal{N}_i} w_{ij}(\hat{\tilde{z}}_i^k - \hat{\tilde{z}}_j^k) - \sum_{j \in \mathcal{N}_i} w_{ij}(\nabla f_i^k - \nabla f_j^k), \tag{2.107b}$$

where we use $u_i(\tilde{x}_i(t)) := (\nabla f_i(\tilde{x}_i(t) + x^\star) - \nabla f_i(x^\star))$ and the local components of the shorthands given in (2.73). By (2.104), the variables $\tilde{x}_i(t)$ and $\tilde{z}_i(t)$ are bounded for all $i \in \{1, \dots, N\}$ and $k \geq 0$. Then, by defining $c_7 := \max_{i,t} \|\tilde{x}_i(t)\|$ and $c_8 := \max_{i,t} \|\tilde{z}_i(t)\|$, (2.107a) and the triangle inequality can be combined to get

$$\left\|\dot{\tilde{x}}_i(t)\right\| \leq \sum_{j \in \mathcal{N}_i} w_{ij} 2c_7 + c_8 + \|u_i(\tilde{x}_i(t))\| \overset{(a)}{\leq} (2c_9 + \overline{L})c_7 + c_8,$$

where in (a) we introduce $c_9 := \sum_{j \in \mathcal{N}_i} w_{ij}$ and we use the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.4). Using again the boundedness of the

quantities, and by adding and subtracting $\nabla f_i(x^\star)$ within the second sum of (2.107b), it holds

$$\left\|\dot{\tilde{z}}_i(t)\right\| \leq 2c_9(c_8 + \overline{L}c_7).$$

Moreover, the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.4) also ensures that $\left\|\nabla^2 f_i(v)\right\| \leq \overline{L}$, for all $v \in \mathbb{R}^d$ and all $i \in \{1, \dots, N\}$. By combining the latter with the two previous equations, the inequality (2.106) can be upper bounded as

$$\frac{d}{dt}\left\|e_i(t)\right\| \leq c_{10}, \tag{2.108}$$

with $c_{10} := (1 + \overline{L})(2c_9 + \overline{L})c_7 + c_8 + 2c_9(c_8 + \overline{L}c_7)$.

Since the protocol (2.94) imposes $e_i(t) = 0$ at the beginning of each time interval $[t_i^{k_i}, t_i^{k_i+1})$, then by also using (2.108), we can write

$$e_i(t) = e_i(t_i^{k_i}) + \int_{t_i^{k_i}}^{k} \frac{d\left\|e_i(\tau)\right\|}{d\tau} d\tau \leq c_{10}(t - t_i^{k_i}). \tag{2.109}$$

By (2.95), it holds $\xi_i(t) = \xi_i(0)\exp(-\nu t)$ for all $k \geq 0$. Thus, being $\lambda\left\|h_i(t)\right\| \geq 0$ for any $k \geq 0$, the bound in (2.109) imposes, as a necessary condition to satisfy the triggering in (2.94), that

$$c_{10}(t_i^{k_i+1} - t_i^{k_i}) \geq |\xi_i(0)|\exp(-\nu t_i^{k_i+1}) \tag{2.110}$$

From (2.105), for all $\epsilon > 0$ there exists $k_{i,\epsilon} \in \mathbb{N}$ such that

$$t_i^{k_i} \in [t_i^\infty - \epsilon, t_i^\infty], \quad \forall k_i \geq k_{i,\epsilon}. \tag{2.111}$$

Set

$$\epsilon := \frac{|\xi_i(0)|}{2c_{10}}\exp(-\nu t_i^\infty), \tag{2.112}$$

and suppose that the $k_{i,\epsilon}$-th triggering time of agent i, namely $t_i^{k_{i,\epsilon}}$, has occurred. Let $t_i^{k_{i,\epsilon}+1}$ be the next triggering time determined by (2.94). Then, using the necessary condition (2.110) we can write

$$t_i^{k_{i,\epsilon}+1} - t_i^{k_{i,\epsilon}} \geq \frac{|\xi_i(0)|}{c_{10}}\exp(-\nu t_i^{k_{i,\epsilon}+1}) \overset{(a)}{\geq} \frac{|\xi_i(0)|}{c_{10}}\exp(-\nu t_i^\infty) \overset{(b)}{=} 2\epsilon, \tag{2.113}$$

where in (a) we use $t_i^{\infty} \geq t_i^{k_{i,\epsilon}+1}$, while in (b) we use (2.112). However (2.113) implies

$$t_i^{k_{i,\epsilon}} \leq t_i^{k_{i,\epsilon}+1} - 2\epsilon \leq t_i^{\infty} - 2\epsilon,$$

which contradicts (2.111) and concludes the proof. ■

Robustness Against Inexact Computation

This section considers a more general scenario in which agents can access only inexact evaluations of their local state (x_i, z_i) and/or of the local gradients ∇f_i. Let $v_{i,\nabla}(t) \in \mathbb{R}^n$ represents the mismatch between the exact value of $\nabla f_i(x_i(t))$ and the one available to agent i for the local updates. The presence of this mismatch may be due to several reasons as, e.g., quantization errors of the computing unit, measurement errors in the sensor providing $\nabla f_i(x_i(t))$, or model uncertainties affecting the available gradient. Similarly, also mismatches affecting the states x_i and z_i can be considered. Thus, we consistently introduce $v_{i,x}(t) \in \mathbb{R}^n$ and $v_{i,z}(t) \in \mathbb{R}^n$ to model such uncertainties. This framework can be formalized by writing

$$\begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} -L & -I \\ 0 & -L \end{bmatrix} \begin{bmatrix} x(t) \\ z(t) \end{bmatrix} - \begin{bmatrix} I \\ L \end{bmatrix} G(x(t)) + \delta_1 B_2 e(t) + B_3 v(t), \tag{2.114}$$

where $e(t) := \text{col}(\hat{x}^k - x(t), \hat{z}^k - z(t), G^k - G(x(t)))$ collects (possible) mismatches due to discrete-time communication and $v(t) := \text{col}(v_{\nabla}(t), v_x(t), v_z(t))$ collects the mentioned local mismatches between the gradients, the solution estimates and the auxiliary variables, the matrix B_2 has the same meaning as in (2.76), and B_3 is defined as

$$B_3 := \begin{bmatrix} -L & -I & -I \\ 0 & -L & -L \end{bmatrix}. \tag{2.115}$$

Finally, δ_1 is equal to 0 for Continuous Gradient Tracking and equal to 1 for both the Synchronous and Asynchronous Triggered Gradient Tracking. Similarly, we denote as δ_2 is equal to 0 for both Continuous Gradient Tracking and Synchronous Triggered Gradient Tracking and equal to 1 for Asynchronous Triggered Gradient Tracking.

Next, the robustness of the algorithm in terms of input-to-state stability is studied. Specifically, we guarantee that within the framework modeled by (2.114), the proposed algorithms behave as input-to-state stable systems. Therefore, in presence of mismatches on variables and gradients, the distance between the solution of problem (2.1) and the computed estimates stay bounded according to the error magnitude.

Proposition 2.1. *Consider the algorithm described by* (2.114)*. Let Assumptions 2.3, 2.4, 2.5 hold and pick any* $\text{col}(x(0), z(0))$ *such that* $\mathbf{1}_{N,n}^{\top} z(0) = 0$*. Then, there exist a* $\mathcal{KL}$ *function*

$g_1(\cdot)$ and a $\mathcal{K}_\infty$ function $g_2(\cdot)$ such that for any $x(0) \in \mathbb{R}^{Nn}$ it holds $\|x(t) - \mathbf{1}_{N,n}x^\star\| \leq g_1(\|\chi(0)\|, t) + g_2(\|v(\cdot)\|_\infty)$, with $\chi(0) := \text{col}(x(0) - \mathbf{1}_{N,n}x^\star, z(0) + G(\mathbf{1}_{N,n}x^\star), \delta_2\xi(0))$ and for any $v(\cdot) \in \mathcal{L}_\infty^{3Nn}$.[1] △

Proof.

The proof of Proposition 2.1 traces the same initial steps of the proof of Theorem 2.4 and 2.5. Using the change of coordinates in (2.49), (2.51), (2.53), system (2.114) can be recast as

$$\dot{\zeta} = A\zeta + Bu + \delta_1 E e_{\zeta,\nabla} + T_{\tilde{y}}^\top T_1 B_3 v_{xz\nabla}, \tag{2.116}$$

with $\zeta \in \mathbb{R}^d$, $u \in \mathbb{R}^{Nn}$, where $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times Nn}$ are as in (2.57), E and $e_{\zeta,\nabla}$ are as in (2.79a) and (2.79b), B_3, $T_{\tilde{y}}$, and T_1 are as in (2.115), (2.51), and (2.54), while $v_{xz\nabla} := \text{col}(v_x, v_z, v_\nabla)$. We remark that $e_{\zeta,\nabla}$ changes according to the implemented communication protocol. Moreover, when Asynchronous Triggered Gradient Tracking is considered, also dynamics (2.95) has to be taken into account. However, when $v_\nabla \equiv v_{xz} \equiv 0$, then $v_{xz\nabla} \equiv 0$ and system (2.116) reduces to

$$\dot{\zeta} = A\zeta + Bu + \delta_1 E e_{\zeta,\nabla}. \tag{2.117}$$

Theorems 2.3, 2.4, and 2.5 ensure that the origin is globally exponentially stable for (2.117) for both $\delta_1, \delta_2 \in \{0, 1\}$ and for both communication protocols (2.75) and (2.94). In light of [91, Lemma 4.6], this condition is sufficient to assert that system (2.116) is input-to-state stable and the proof follows (cf. [175, Section 2.9]). ■

2.4.3 Numerical Simulations

We next present numerical simulations to confirm and support the theoretical findings. The simulations are done using Matlab with its numerical solver "ode45" to integrate the Continuous Gradient Tracking.

We consider a network of agents that want to cooperatively solve the data analytics problem presented in Section 1.2.2. We briefly recall the problem as follows. The agents want to train a linear classifier and each agent i is equipped with $m_i \in \mathbb{N}$ points $p_{i,1}, \ldots, p_{i,m_i} \in \mathbb{R}^n$ with binary labels $l_{i,q} \in \{-1, 1\}$ for all $q \in \{1, \ldots, m_i\}$. To this end, we consider a logistic regression problem given by

$$\min_{w,b} \sum_{i=1}^{N} \sum_{q=1}^{m_i} \log\left(1 + \exp(-l_{i,q}(w^\top p_{i,q} + b))\right) + \frac{C(\|w\|^2 + b^2)}{2},$$

where the optimization variables $w \in R^{n-1}$ and $b \in \mathbb{R}$ define the separating hyperplane,

[1] See [91, Chapter 4] for the function classes' definitions.

while $C > 0$ is the so-called regularization parameter. Notice that the presence of the regularization makes the cost function strongly convex. In our simulations, we pick $n = 3$, $m_i = 10$ for all $i \in \{1, \dots, N\}$, and $C = 0.1$.

Continuous Gradient Tracking

In this subsection, the effectiveness of Continuous Gradient Tracking is shown on a network of $N = 50$ agents communicating according to an undirected and connected Erdős-Rényi graph with parameter 0.4. In Figure 2.7 the convergence performances of Continuous Gradient Tracking algorithm are shown. Specifically, the distance of the local estimates $x(t) := \text{col}(x_1(t), \dots, x_N(t))$ from the optimum $\|x(t) - \mathbf{1}_{N,n} x^\star\|$, converges to zero exponentially fast as expected from Theorem 2.3.

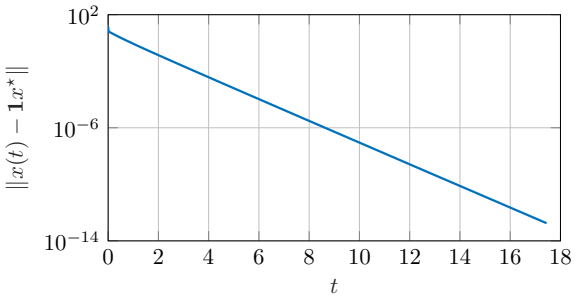


Figure 2.7: Evolution of the distance from the optimum of local estimates generated by Continuous Gradient Tracking.

Synchronous and Asynchronous Triggered Gradient Tracking

In this subsection, the effectiveness of the triggered algorithms is shown for a network of $N = 10$ agents communicating according to an undirected and connected Erdős-Rényi graph with parameter 0.4. We tested Synchronous Triggered Gradient Tracking and Asynchronous Triggered Gradient Tracking for different values of their key parameters Δ and λ, respectively. Moreover, we experimentally tuned the step-size for the discrete Gradient Tracking as $\gamma = 0.1$ in order to optimize its convergence rate. Finally, we set $\nu = 5$ for the dynamics of ξ_i in (2.95). For the simulation of Asynchronous Triggered Gradient Tracking, the triggering condition (cf. (2.94)) is checked every 0.001 seconds. Figure 2.8 compares the evolution of the optimality error obtained with different Δ and λ, for Synchronous Triggered Gradient Tracking, Asynchronous Triggered Gradient Tracking, and the discrete Gradient Tracking algorithm. Specifically, the comparison is done in terms of communication rounds. The plot considers the performances of the most efficient agent, say $i_\star$, that performs the smallest number of neighboring communications in Asynchronous Triggered Gradient Tracking. As for the discrete

Gradient Tracking algorithm, we denote, with a slight abuse of notation, $x_{i_\star}(t_{i_\star}^{k_{i_\star}}) = \mathrm{x}_{i_\star}^k$, with the sequence $\{\mathrm{x}_{i_\star}^k\}_{k\geq 0}$ generated by (2.44). As Figure 2.8 clearly highlights, the communication rounds decrease as λ increases. The same applies to Δ. In particular, we underline that Asynchronous Triggered Gradient Tracking results more efficient in finding the optimal solution with respect to both Synchronous Triggered Gradient Tracking and discrete gradient tracking.

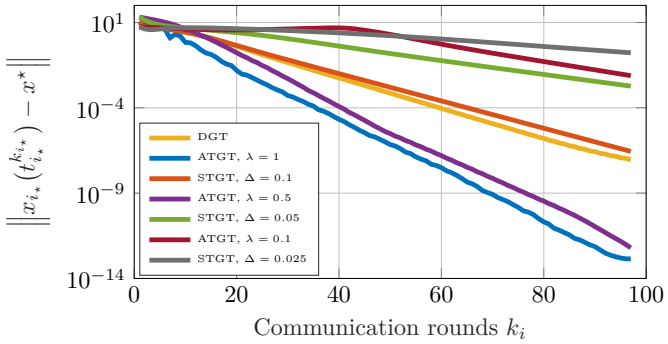


Figure 2.8: Comparison among Asynchronous Triggered Gradient Tracking (ATGT), Synchronous Triggered Gradient Tracking (STGT) and the discrete gradient tracking (DGT) in terms of evolution of the optimality error.

Finally, in Figure 2.9 each cross represents when the triggering condition occurred for each agent while running the Asynchronous Triggered Gradient Tracking with $\lambda = 0.1$. The plots demonstrate how event-triggered communication effectively reduces inter-agents communication.

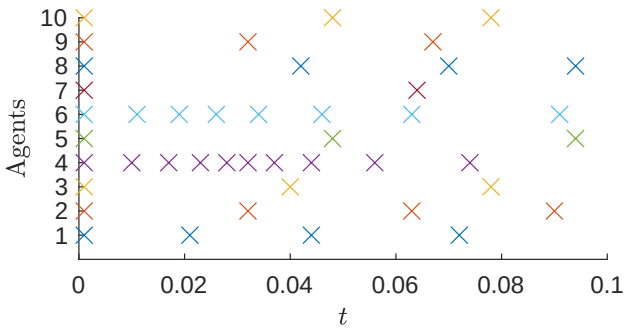


Figure 2.9: Occurrence of the triggering conditions in the Asynchronous Triggered Gradient Tracking.

2.5 Derivative-Free Distributed Consensus Optimization

In this section, we address problem (2.1) in a derivative-free manner, i.e., by assuming that the agents cannot access the gradients (or other derivatives) of the objective functions. In this setup, we propose Extremum Tracking Descent, i.e., a distributed

method whose algorithmic structure is inspired by the Gradient Tracking algorithm and that uses an equilibrium seeking technique to replace the unavailable gradients. In this section, we enforce the following assumptions.

Assumption 2.6. *Graph $\mathcal{G}$ is connected and the adjacency matrix $\mathcal{W}_{\mathcal{G}} \in^{N\times N}$ is symmetric.*$\triangle$

Assumption 2.7. *For all $i \in \{1, \dots, N\}$, the function f_i is $\underline{L}$-strongly convex for some $\underline{L} > 0$.* $\triangle$

Assumption 2.8. *Each function f_i is (at least) $\mathcal{C}^3$ and has $\overline{L}_i$-Lipschitz continuous gradients, namely there exists a constant $\overline{L}_i > 0$ such that for all $i \in \{1, \dots, N\}$ it holds*

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\| \leq \overline{L}_i \|x_1 - x_2\|,$$

for any x_1, $x_2 \in \mathbb{R}^n$. We denote $\overline{L} = \max\{\overline{L}_1, \dots, \overline{L}_N\}$. $\triangle$

Before the description of our derivative-free algortihm, we introduce the forward Euler discretization of Continuous Gradient Tracking (cf. (2.48)), namely the discrete-time scheme described by

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k - \gamma \sum_{j\in\mathcal{N}_i} \ell_{ij}\mathrm{x}_j^k - \gamma\left(\mathrm{z}_i^k + \nabla f_i(\mathrm{x}_i^k)\right) \tag{2.118a}$$

$$\mathrm{z}_i^{k+1} = \mathrm{z}_i^k - \gamma \sum_{j\in\mathcal{N}_i} \ell_{ij}\left(\mathrm{z}_j^k + \nabla f_j(\mathrm{x}_j^k)\right), \tag{2.118b}$$

where $\gamma > 0$ represents the time step and ℓ_{ij} the (i,j)-entry of the Laplacian matrix $\mathcal{L}$ associated to the graph $\mathcal{G}$. In this algorithm, agents exchange with their neighbors the information $\mathrm{col}(\mathrm{x}_i^k, \mathrm{z}_i^k + \nabla f_i(\mathrm{x}_i^k))$ involving $2n$ components. The main idea consists of estimating $\nabla f_i(\mathrm{x}_i^k)$, supposed non-measurable, via an extremum-seeking algorithm. It is worth noting how system (2.118) is in the so-called averaging standard form for discrete-time systems [165], which will be useful in the analysis of the extremum seeking version of this protocol.

2.5.1 Extremum Tracking Descent: Algorithm Description and Analysis

The proposed algorithm is inspired by (2.118), which is redesigned via extremum seeking by replacing local gradients with a proper estimation based on the local cost function values and suitable dithering signals defined as

$$\mathrm{d}_i^k = \mathrm{col}\left(\sin\left(\frac{2\pi k}{\tau_{i_1}} + \phi_{i_1}\right), \dots, \sin\left(\frac{2\pi k}{\tau_{i_n}} + \phi_{i_n}\right)\right), \tag{2.119}$$

where $\tau_{i_p} \in \mathbb{N}$ and $\phi_{i_p} \in \mathbb{R}$ such that, given $p, q, r \in \{1, \dots, n\}$, $p \neq q$, $q \neq r$, $q \neq r$, it holds

$$\sum_{k=0}^{\tau-1} \sin\left(\frac{2\pi k}{\tau_{i_p}} + \phi_{i_p}\right) = 0 \tag{2.120a}$$

$$\sum_{k=0}^{\tau-1} \sin\left(\frac{2\pi k}{\tau_{i_p}} + \phi_{i_p}\right) \sin\left(\frac{2\pi t}{\tau_{i_q}} + \phi_{i_q}\right) = \frac{\tau}{2} \tag{2.120b}$$

$$\sum_{k=0}^{\tau-1} \sin\left(\frac{2\pi k}{\tau_{i_p}} + \phi_{i_p}\right) \sin\left(\frac{2\pi t}{\tau_{i_p}} + \phi_{i_p}\right) \sin\left(\frac{2\pi k}{\tau_{i_r}} + \phi_{i_r}\right) = 0, \tag{2.120c}$$

for $i \in \{1, \dots, N\}$. Here, $\tau \in \mathbb{N}$ is the least common multiple of all periods τ_{ip}.Extremum Tracking Descent is described in Algorithm 2 from the perspective of agent i. In the table, the parameter $\delta_i > 0$ represents the amplitude of the dither signal d_i^t. Notice that, as in (2.118), agents communicate to neighbors a total of $2n$ components.

Algorithm 2 Extremum Tracking Descent (agent i)

initialization: $x_i^0 \in \mathbb{R}^n$ and $z_i^0 = 0$

for $t = 0, 1, \dots$ **do**

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k - \gamma \left(\sum_{j \in \mathcal{N}_i} \ell_{ij} \mathrm{x}_j^k + \mathrm{z}_i^k + \frac{2 f_i(\mathrm{x}_i^k + \delta_i \mathrm{d}_i^k)\mathrm{d}_i^k}{\delta_i} \right) \tag{2.121a}$$

$$\mathrm{z}_i^{k+1} = \mathrm{z}_i^k - \gamma \sum_{j \in \mathcal{N}_i} \ell_{ij} \left(\mathrm{z}_j^k + \frac{2 f_j(\mathrm{x}_j^k + \delta_j \mathrm{d}_j^k)\mathrm{d}_j^k}{\delta_j} \right) \tag{2.121b}$$

end for

The convergence properties of Extremum Tracking Descent are formalized in the next theorem.

Theorem 2.6. *Consider* (2.121) *and let Assumptions 2.6, 2.7, and 2.8 hold. Then, for any $r, \bar{\rho} > 0$, there exist $\gamma^\star, \delta^\star, k_1 > 0$, $\epsilon \leq \bar{\rho}$, and $k_2 \geq (\bar{\rho} - \epsilon)$ such that, for any $\gamma \in (0, \gamma^\star)$, $\textsc{col}(\mathrm{x}_i, \mathrm{z}_i) \in \{\textsc{col}(\mathrm{x}_i^0, \mathrm{z}_i^0) \in \mathbb{R}^{2n} \mid \|\mathrm{x}_i - x^\star\| \leq r, z_i = 0\}$, $\delta_i \in (0, \delta^\star)$, $i \in \{1, \dots, N\}$, the trajectories of* (2.121) *are bounded and for each $i \in \{1, \dots, N\}$*

$$\left\| \mathrm{x}_i^k - x^\star \right\| \leq \bar{\rho}, \tag{2.122}$$

for all $k \geq k^\star$, where $k^\star := -\frac{1}{\gamma k_1} \ln((\bar{\rho} - \epsilon)/k_2)$, i.e., the convergence to the set $\{x_i \in \mathbb{R}^n \mid \left\| \mathrm{x}_i^k - x^\star \right\| \leq \bar{\rho}\}$ is linear. △

The proof of Theorem 2.6 is provided in Section 2.5.1. It is worth noting that Theorem 2.6 provides a semi-global, practical exponential-stability result on the discrete-time dynamics described by Extremum Tracking Descent. Indeed, the parameters $\gamma^\star$

and $\delta^\star$ depend on both r and $\bar{\rho}$.

We first rewrite the local updates in (2.121) in an aggregate form as

$$\mathrm{x}^{k+1} = \mathrm{x}^k - \gamma \left(L\mathrm{x}^k + \mathrm{z}^k + f_{\mathrm{d}}(\mathrm{x}^k + \delta \mathrm{d}^k) \right) \tag{2.123a}$$

$$\mathrm{z}^{k+1} = \mathrm{z}^k - \gamma \left(L\mathrm{z}^k + L f_{\mathrm{d}}(\mathrm{x}^k + \delta \mathrm{d}^k) \right), \tag{2.123b}$$

where we introduced $L := \mathcal{L} \otimes I_n$, $\mathrm{x}^k := \mathrm{col}(\mathrm{x}_1^k, \dots, x_N^k)$, $\mathrm{z}^k := \mathrm{col}(\mathrm{z}_1^k, \dots, \mathrm{z}_N^k)$, $f_{\mathrm{d}}(\mathrm{x}^k + \delta \mathrm{d}^k) := \mathrm{col}(2f_1(\mathrm{x}_1^k + \delta_1 \mathrm{d}_1^k)\mathrm{d}_1^k/\delta_1, \dots, 2f_N(\mathrm{x}_N^k + \delta_N \mathrm{d}_N^k)\mathrm{d}_N^k/\delta_N)$, $\mathrm{d}^k := \mathrm{col}(\mathrm{d}_1^k, \dots, \mathrm{d}_N^k)$, and $\delta := \mathtt{diag}(I_n, \dots, I_n) \otimes I_n$. We point out that, see also Fig. 2.10, system (2.123) can be conceived as an extremum seeking scheme with output map $f(\mathrm{x} + \delta \mathrm{d}^k)$.

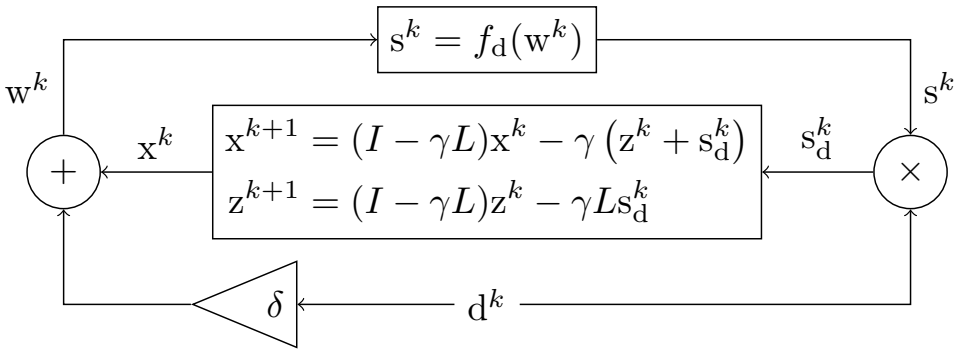


Figure 2.10: Block scheme of the proposed Extremum Tracking Descent algorithm in the (x, z) coordinates.

We now give an overview of the main steps of the stability analysis carried out to prove Theorem 2.6:

(i) We perform a first change of variables to describe the dynamics (2.123) in terms of the mean value (over the agents) $\tilde{\bar{z}}$ of z and the orthogonal part $\tilde{\mathrm{z}}_\perp$ associated to the consensus error. Then, by relying on averaging theory [165], we introduce a so-called average system obtained by averaging the dithering signals over a common period. This system is shown to be driven by the local function gradients with additive estimation errors.

(ii) When neglecting these errors, the average system corresponds to an equivalent form of (2.118). Based on this observation, we rely on existing stability properties of the continuous gradient tracking to demonstrate that the trajectories of the average system exponentially converge to an arbitrarily small neighborhood of $\mathrm{col}(\mathbf{1}x^\star, \tilde{\mathrm{z}}_\perp^{\mathrm{eq}})$ for some $\tilde{\mathrm{z}}_\perp^{\mathrm{eq}}$ arising from the analysis.

(iii) Finally, we prove Theorem 2.6 by exploiting the steps above and by using averaging theory to show the closeness between the trajectories of (2.123) and those of its average system.

We start by introducing a change of coordinates to highlight the error dynamics. To this end, let $G : \mathbb{R}^{Nn} \to \mathbb{R}^{Nn}$ be

$$G(\mathrm{x}^k) := \mathrm{col}(\nabla f_1(\mathrm{x}_1^k), \dots, \nabla f_N(\mathrm{x}_N^k)).$$

Then, let the error coordinates $\tilde{\mathrm{x}}^k, \tilde{\mathrm{z}}^k \in \mathbb{R}^{Nn}$ be

$$\tilde{\mathrm{x}}^k := \mathrm{x}^k - \mathbf{1}_{N,n} x^\star, \quad \tilde{\mathrm{z}}^k = \mathrm{z}^k + G(\mathbf{1}_{N,n} x^\star), \tag{2.124}$$

and let us introduce $\phi_{\mathrm{xz}} : \mathbb{N} \times \mathbb{R}^{Nn} \times \mathbb{R}^{Nn} \to \mathbb{R}^{2Nn}$ as

$$\phi_{\mathrm{xz}}(t, \tilde{\mathrm{x}}^k, \tilde{\mathrm{z}}^k) := \begin{bmatrix} -L\tilde{\mathrm{x}}^k - \tilde{\mathrm{z}}^k - f_{\mathrm{d}}(\tilde{\mathrm{x}}^k + \mathbf{1}_{N,n} x^\star + \delta \mathrm{d}^k) + G(\mathbf{1}_{N,n} x^\star) \\ -L\tilde{\mathrm{z}}^k - L(f_{\mathrm{d}}(\tilde{\mathrm{x}}^k + \mathbf{1}_{N,n} x^\star + \delta \mathrm{d}^k) + G(\mathbf{1}_{N,n} x^\star)) \end{bmatrix},$$

Then, by using the new coordinates, we rewrite (2.123) as

$$\begin{bmatrix} \tilde{\mathrm{x}}^{k+1} \\ \tilde{\mathrm{z}}^{k+1} \end{bmatrix} = \begin{bmatrix} \tilde{\mathrm{x}}^k \\ \tilde{\mathrm{z}}^k \end{bmatrix} + \gamma \phi_{\mathrm{xz}}(t, \tilde{\mathrm{x}}^k, \tilde{\mathrm{z}}^k), \tag{2.125}$$

where we have used the property $L\mathbf{1}_{N,n} = 0$. As in the previous sections, we take advantage of the initialization $\mathrm{z}_i^0 = 0$ for all $i \in \{1, \dots, N\}$ by defining

$$\begin{bmatrix} \tilde{\bar{\mathrm{z}}}^k \\ \tilde{\mathrm{z}}_\perp^k \end{bmatrix} := \begin{bmatrix} \frac{\mathbf{1}_{N,n}^\top}{N} \\ \mathbb{R}^\top \end{bmatrix} \tilde{\mathrm{z}}^k, \quad \xi^k := \begin{bmatrix} \tilde{\mathrm{x}}^k \\ \tilde{\mathrm{z}}_\perp^k \end{bmatrix}, \tag{2.126}$$

where we introduced the matrix $R \in \mathbb{R}^{Nn \times (N-1)n}$ such that $R^\top \mathbf{1}_{N,n} = 0$, $R^\top R = I$. Then, since $\mathbf{1}_{N,n}^\top L = 0$ in light of Assumption 2.6 and $\mathbf{1}_{N,n}^\top G(\mathbf{1}_{N,n} x^\star) = \sum_{i=1}^N \nabla f_i(x^\star) = 0$ (since $x^\star$ is the minimizer of problem (2.1)), system (2.125) reads as

$$\xi^{k+1} = \xi^k + \gamma \phi_\xi(k, \mathrm{col}(\tilde{\mathrm{x}}^k, \mathbf{1}_{N,n} \tilde{\bar{\mathrm{z}}}^k + R\tilde{\mathrm{z}}_\perp^k)) \tag{2.127a}$$

$$\tilde{\bar{\mathrm{z}}}^{k+1} = \tilde{\bar{\mathrm{z}}}^k, \tag{2.127b}$$

where we introduced $\phi_\xi : \mathbb{N} \times \mathbb{R}^{2Nn} \to \mathbb{R}^{(2N-1)n}$ defined as

$$\phi_\xi(k, \mathrm{col}(\tilde{\mathrm{x}}, \tilde{\mathrm{z}})) := \begin{bmatrix} I & 0 \\ 0 & R \end{bmatrix} \phi_{\mathrm{xz}}(k, \tilde{\mathrm{x}}, \tilde{\mathrm{z}}).$$

The equation (2.127b) allows us to claim that $\tilde{\bar{\mathrm{z}}}^k = \tilde{\bar{\mathrm{z}}}^0$ for all $k \geq 0$. Moreover, we recall that (i) $\mathrm{z}_i^0 = 0$ for all $i \in \{1, \dots, N\}$, and (ii) $\mathbf{1}_{N,n}^\top G(\mathbf{1}_{N,n} x^\star) = 0$. Hence, it holds $\tilde{\bar{\mathrm{z}}}^0 = 0$ which allows us to ignore (2.127b) and rewrite (2.127) according to the equivalent,

reduced system

$$\xi^{k+1} = \xi^k + \gamma\phi(t, \xi^k), \tag{2.128}$$

where $\phi(k, \xi^k) := \phi_\xi(k, \text{col}(\tilde{\text{x}}^k, R\tilde{\text{z}}^k_\perp))$,with $\xi^k = \text{col}(\tilde{\text{x}}^k, \tilde{\text{z}}^k)$.

We define the *average system* associated to (2.128) as

$$\xi_a^{k+1} = \xi_a^k + \gamma\phi_a(\xi_a^k), \qquad \xi_a^0 = \xi_0 \tag{2.129}$$

with $\phi_a(\xi_a^k) = \dfrac{1}{\tau}\sum_{q=k+1}^{k+\tau}\phi(q, \xi_a^k)$. We need the following result to detail $\phi_a(\xi_a^k)$.

Lemma 2.3 (Gradient estimation)**.** *For all $i \in \{1, \dots, N\}$, there exists $\ell_i : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$\frac{2}{\delta_i\tau}\sum_{q=k+1}^{k+\tau} f_i(x_i + \delta_i d_i^q)d_i^q = \nabla f_i(x_i) + \delta_i^2\ell_i(x_i),$$

for all $x_i \in \mathbb{R}^n$ and all $k \geq 0$. Moreover, given any compact set $\mathcal{S}_i \subset \mathbb{R}^n$, if $\delta_i \in (0, 1]$, there exists $L_{i,\mathcal{S}_i} > 0$ such that

$$\|\ell_i(x_i)\| \leq L_{i,\mathcal{S}_i}, \tag{2.130}$$

for any $x_i \in \mathcal{S}_i$ and all $i \in \{1, \dots, N\}$. △

Proof. Given $\alpha = \text{col}(\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$, $y = \text{col}(y_1, \dots, y_n) \in \mathbb{R}^n$, and a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, we define

$$\alpha! := \alpha_1! \dots \alpha_n!, \qquad y^\alpha := y_1^{\alpha_1} \dots y_n^{\alpha_n},$$
$$\partial^\alpha f(y) := \frac{\partial^{\alpha_1}}{\partial y_1^{\alpha_1}} \dots \frac{\partial^{\alpha_n}}{\partial y_n^{\alpha_n}} f(y), \qquad |\alpha| := \alpha_1 + \dots + \alpha_n.$$

Being each function f_i smooth (cf. Assumption 2.8), we can apply Taylor's expansion (cf. [68, Theorem 2]) and write

$$f_i(x_i + \delta_i \text{d}_i^k) = f_i(x_i) + \delta_i {\text{d}_i^k}^\top \nabla f_i(x_i) + \frac{\delta_i^2}{2}{\text{d}_i^k}^\top \nabla^2 f_i(x_i)\text{d}_i^k + \delta^3 R_{i,2}(x_i, \delta_i \text{d}_i^k), \tag{2.131}$$

where the remainder $R_{i,2}(x_i, \delta_i \text{d}_i^k)$ is given by

$$R_{i,2}(x_i, \delta_i \text{d}_i^k) = \sum_{|\alpha|=3} \frac{\partial^\alpha f_i(x_i + c\delta_i \text{d}_i^k)}{\alpha!}(\delta_i \text{d}_i^k)^\alpha, \tag{2.132}$$

for some $c \in (0,1)$. Then, we can use (2.131) to write

$$\frac{2}{\delta_i \tau} \sum_{q=k+1}^{k+\tau} \mathrm{d}_i^q f_i(x_i + \delta_i \mathrm{d}_i^q) = \frac{2 f_i(x_i)}{\delta_i \tau} \sum_{q=k+1}^{k+\tau} \mathrm{d}_i^q + \left[\frac{2}{\tau} \sum_{q=k+1}^{k+\tau} \left(\mathrm{d}_i^q \mathrm{d}_i^{q\top} \right) \right] \nabla f_i(x_i) + \frac{\delta_i}{\tau} \sum_{q=k+1}^{k+\tau} \left(\mathrm{d}_i^q \mathrm{d}_i^{q\top} \right) \nabla^2 f_i(x_i) \mathrm{d}_i^q + \frac{2}{\delta_i \tau} \sum_{q=k+1}^{k+\tau} \mathrm{d}_i^q R_{i,2}(x_i, \delta_i \mathrm{d}_i^q). \tag{2.133}$$

Since the frequencies of d_i^q satisfy (2.120), we get

$$\begin{aligned} &\sum_{q=k+1}^{k+\tau} \mathrm{d}_i^q = 0 \\ &\frac{2}{\tau} \sum_{q=k+1}^{k+\tau} \left(\mathrm{d}_i^q \mathrm{d}_i^{q\top} \right) = I_n \\ &\sum_{q=k+1}^{k+\tau} \left(\mathrm{d}_i^q \mathrm{d}_i^{q\top} \right) \nabla^2 f_i(x_i) \mathrm{d}_i^q = 0, \end{aligned}$$

which combined with (2.133), allows us to write

$$f_i(x_i + \delta_i \mathrm{d}_i^k) = \nabla f_i(x_i) + \frac{2}{\delta_i \tau} \sum_{q=k+1}^{k+\tau} \mathrm{d}_i^q R_{i,2}(x_i, \delta_i \mathrm{d}_i^q).$$

The proof follows by setting $\ell_i(x_i) = \frac{2}{\tau} \sum_{q=k+1}^{k+\tau} \mathrm{d}_i^q R_{i,2}(x_i, \delta_i \mathrm{d}_i^q)/\delta_i^3$. Finally, given a compact set $\mathcal{S}_i \subset \mathbb{R}^n$, let us bound $\|\ell_i(x_i)\|$ for any $x_i \in \mathcal{S}_i$. Note that $\left\|\delta_i \mathrm{d}_i^k\right\| \leq \delta_i \sqrt{n}$ for all $k \geq 0$ and let $\mathcal{S}_i' \subset \mathbb{R}^n$ be a compact set such that (i) $\mathcal{S}_i \subseteq \mathcal{S}_i' \subset \mathbb{R}^n$, and (ii) $x_i + \delta \mathrm{d}_i^k \in \mathcal{S}'$ for any $x_i \in \mathcal{S}_i$, $\delta_i \in (0,1]$, and all $k \geq 0$. Thus, we can write

$$\begin{aligned} \sup_{x_i \in \mathcal{S}_i} \left\| \frac{R_{i,2}(x_i, \delta_i \mathrm{d}_i^q)}{\delta_i^3} \right\| &\overset{(a)}{\leq} \sup_{x_i \in \mathcal{S}_i'} \left\| \frac{1}{\delta_i^3} \sum_{|\alpha|=3} \frac{\partial^\alpha f_i(x_i)}{\alpha!} (\delta_i \mathrm{d}_i^q)^\alpha \right\| \\ &\overset{(b)}{=} \sup_{x_i \in \mathcal{S}_i'} \left\| \sum_{|\alpha|=3} \frac{\partial^\alpha f_i(x_i)}{\alpha!} (\mathrm{d}_i^q)^\alpha \right\| =: L_{i,\mathcal{S}_i}', \end{aligned} \tag{2.134}$$

where in (a) we use the expression (2.132) of $R_{i,2}(x_i, \delta_i \mathrm{d}_i^q)$, the definition of $\mathcal{S}_i'$, and the fact that $\delta_i \in (0,1]$, while in (b) we drop out the term δ_i^3 from $(\delta_i \mathrm{d}_i^q)^\alpha$. We underline that, since the set $\mathcal{S}_i'$ is compact and f_i is smooth, $L_{i,\mathcal{S}_i}$ exists and is finite. The bound of $\ell_i(x_i)$ follows by defining $L_{i,\mathcal{S}_i} := 2 L_{i,\mathcal{S}_i}' \sqrt{n}$ and combining the result (2.134) with the bound about the norm of the dither signal, i.e., $\left\|\mathrm{d}_i^k\right\| \leq \sqrt{n}$ for all $k \geq 0$. ■

Now, let

$$\ell(\mathrm{x}^k) := \mathrm{col}\left(\ell_1(\mathrm{x}_1^k), \dots, \ell_N(\mathrm{x}_N^k)\right).$$

Then, averaging (2.128) over τ samples and using Lemma 2.3 leads to $\phi_a(\xi_a^k) = \phi_{GT}(\xi_a^k) + BM_\delta u(\xi_a^k)$, where $M_\delta := \mathtt{diag}(\delta_1^2, \dots, \delta_N^2) \otimes I_n$, $\xi_a^k := \mathrm{col}(\tilde{\mathrm{x}}_a^k, \tilde{z}_{\perp,a}^k)$, and

$$\begin{aligned}
\phi_{GT}(\xi_a^k) &:= \begin{bmatrix} -L\tilde{\mathrm{x}}_a^k - R\tilde{z}_{\perp,a}^k - G(\tilde{\mathrm{x}}^k + \mathbf{1}_{N,n}x^\star) + G(\mathbf{1}_{N,n}x^\star) \\ -R^\top L\tilde{z}_a^k - R^\top L(G(\tilde{\mathrm{x}}^k + \mathbf{1}_{N,n}x^\star) - G(\mathbf{1}_{N,n}x^\star)) \end{bmatrix} \\
u(\xi_a^k) &:= \ell(\tilde{\mathrm{x}}_a^k + \mathbf{1}_{N,n}x^\star) \\
B &:= \begin{bmatrix} -I \\ -R^\top L \end{bmatrix}.
\end{aligned}$$

Remark 2.4. It is worth highlighting the main distinctive features of our method. First, the finite differences methods are characterized by estimation errors involving second-order terms of the local cost function expansion, while our estimation policy allows for estimation errors involving third-order terms and, thus, an higher precision. Second, the estimation of the gradients is performed according to a single-point estimator and, thus, the objective functions queries and, possibly, communications are reduced. Also, in some application scenarios, it may be not possible to have multiple samples of the cost function, but the user/agent/robot should decide just one. Third and final, the estimation policy is purely deterministic and, thus, the convergence guarantees are deterministic too. △

Average System Analysis

In this subsection, we analyze the average system (2.129). We study (2.129) as the system

$$\xi_a^{k+1} = \xi_a^k + \gamma\phi_{\mathrm{CGT}}(\xi_a^k), \tag{2.135}$$

perturbed by $\gamma BM_\delta u(\xi_a^k)$. The next lemma proves the global exponential stability of the origin for (2.135).

Lemma 2.4. *There exist $P_{EST} = P_{EST}^\top \in \mathbb{R}^{(2N-1)n\times(2N-1)n}$, $a_1, a_2, c_1 > 0$, and $\gamma_0 > 0$ such that, for any $\gamma \in (0, \gamma_0)$, along the trajectories of* (2.135) *it holds*

$$a_1 I \le P_{EST} \le a_2 I \tag{2.136a}$$

$$\xi_a^{k+1\top} P_{EST}\xi_a^{k+1} - \xi_a^{k\top} P_{EST}\xi_a^k \le -\gamma c_1 \left\|\xi_a\right\|^2, \tag{2.136b}$$

for any $\xi_a^k \in \mathbb{R}^{2(N-1)n}$. △

Proof. Theorem 2.3 (cf. Section 2.4) proves that, under the Assumptions 2.6, 2.7, and 2.8, the point $\xi^\star = (\mathbf{1}x^\star, \tilde{z}_\perp^{\text{eq}})$, with $\tilde{z}_\perp^{\text{eq}} := -R^\top G(\mathbf{1}x^\star)$, is a globally exponentially stable equilibrium for the continuous-time system

$$\dot{\xi}(t) = \phi_{\text{CGT}}(\xi(t)),$$

which is equivalent to (2.57), i.e., the system written according to the coordinate $\zeta := \text{col}(\tilde{y}, \tilde{\eta})$ (cf. (2.51)). In detail, Theorem 2.3 (cf. Section 2.4) introduces the matrix $P = P^\top \in \mathbb{R}^{(2N-1)n \times (2N-1)n}$ given by

$$P := \begin{bmatrix} mI & -R \\ -R^\top & mR^\top (L^2)^\dagger R \end{bmatrix},$$

and proves that there exist $\bar{m}, m_1, m_2, m_3 > 0$ such that, for any $m > \bar{m}$, it holds

$$m_1 I \leq P \leq m_2 I \tag{2.137a}$$

$$\zeta^\top P \bar{T} \phi_{\text{CGT}}(\bar{T}^{-1}\zeta) \leq -m_3 \left\|\zeta\right\|^2, \tag{2.137b}$$

for any $\zeta \in \mathbb{R}^{(2N-1)n}$, where we used a tranformation matrix $\bar{T}\mathbb{R}^{(2N-1)n \times (2N-1)n}$ to describe the transformation from ξ to ζ. Based on this observation, we guarantee that there exists $P_{\text{EST}} = P_{\text{EST}}^\top \in \mathbb{R}^{(2N-1)n \times (2N-1)n}$ such that

$$a_1 I \leq P_{\text{EST}} \leq a_2 I \tag{2.138a}$$

$$\xi_a^\top P_{\text{EST}} \phi_{\text{CGT}}(\xi_a) \leq -a_3 \left\|\xi_a\right\|^2, \tag{2.138b}$$

for any $\xi_a \in \mathbb{R}^{(2N-1)n}$. Then, we use P_{EST} to introduce the candidate Lyapunov function $V : \mathbb{R}^{(2N-1)n} \to \mathbb{R}$ considered in Theorem 2.3 defined as With this result at hand, we can bound $\Delta V(\xi_a^k) := V(\xi_a^{k+1}) - V(\xi_a^k)$ along the trajectories of (2.135) as

$$\Delta V(\xi_a^k) \leq -\gamma 2a_3 \left\|\xi_a^k\right\|^2 + \gamma^2 \phi_{\text{CGT}}(\xi_a^k)^\top P_{\text{EST}} \phi_{\text{CGT}}(\xi_a^k). \tag{2.139}$$

Moreover, by using the Lipschitz continuity of the gradients of the objective functions (cf. Assumption 2.8) and the definition of ϕ_{CGT}, there exists $a_4 > 0$ such that

$$\left\|\phi_{\text{CGT}}(\xi_a^k)\right\| \leq a_4 \left\|\xi_a^k\right\|. \tag{2.140}$$

Finally, for any $c \in (0, 2a_3)$, let $\gamma_0 := (2a_3 - c)/(a_2 a_4^2)$ and the proof follows by using (2.139) and (2.140). ■

Remark 2.5. Notice that Lemma 2.4 proves the algorithm (2.118) linearly converges to the minimizer of (2.1), since (2.135) is an equivalent formulation of (2.118). △

Now, we analyze the impact of $u(\cdot)$ on (2.129).

Lemma 2.5. *Assume that there exist $a_1, a_2, c_1 > 0$ and $P_{EST} = P_{EST}^\top \in \mathbb{R}^{(2N-1)n\times(2N-1)n} \to \mathbb{R}$ such that conditions (2.136) hold. Then, for any $r_\xi > 0, c_1' \in (0, c_1)$, and $\rho \in (0, r_\xi)$, there exist $c_3 \geq 0$ and $\delta^\star \in (0, 1]$ such that, for any $\gamma \in (0, 1]$, $\delta_i \in (0, \delta^\star)$, $i \in \{1, \dots, N\}$, and $\left\|\xi_a^0\right\| \leq r_\xi$, it holds*
(i) $\xi_a^k \in \mathcal{B}_{\sqrt{a_2/a_1} r_\xi}$ for all $k \geq 0$, and
(ii)

$$\|\xi_a^k\| \leq \sqrt{a_2/a_1} \exp\left(-k\gamma c_3\right) \|\xi_a^0\|, \tag{2.141}$$

for any $\left\|\xi_a^k\right\| \geq \rho$. △

Proof. The proof relies on (i) the matrix P satisfying (2.136), and (ii) the fact that the norm of the perturbation term $\gamma B M_\delta u(\xi_a^k)$ can be arbitrarily reduced through the parameters δ_i as long as ξ_a^k lies into a compact set. First of all, without loss of generality, we assume $\rho \leq r_\xi$. Indeed, we will use the parameter r_ξ to define a (compact) ball and arbitrarily bound the norm of the perturbation term $\gamma B M_\delta u(\xi_a^k)$ through the parameters δ_i as long as ξ_a^k lies into this ball. Hence, we can always use the more conservative condition. In detail, we define $V(\xi_a) := \xi_A^\top P_{\text{EST}} \xi_a$ and $\Omega_{r_\xi} := \{\xi_a^{(2N-1)n} \mid V(\xi_a) \leq a_2 r_\xi^2\} \subset \mathbb{R}^{(2N-1)n}$. Then, from (2.136a), we derive $\mathcal{B}_{r_\xi} \subseteq \Omega_{r_\xi} \subseteq \mathcal{B}_{r_\xi'}$, where $r_\xi' := \sqrt{a_2/a_1} r_\xi$. Thus, it holds $\xi_a^k \in \Omega_{r_\xi}$. Now, under the assumption $\xi_a^k \in \mathcal{B}_{r_\xi}$ (later verified by a proper selection of the algorithm parameters), we use (2.139), the Cauchy-Schwarz inequality, the result (2.140), and the parameter $\bar{\delta} := \max\{\delta_1, \dots, \delta_N\}$, to bound $\Delta V(\xi_a^k) := V(\xi_a^{k+1}) - V(\xi_a^k)$ along the trajectories of (2.129) as

$$\begin{aligned}\Delta V(\xi_a^k) \leq -\gamma c_1 \|\xi_a^k\|^2 + \bar{\delta}^2 \gamma 2 \left\|P_{\text{EST}} B\right\| \|\xi_a^k\| \left\|u(\xi_a^k)\right\| + \bar{\delta}^2 \gamma^2 2 a_4 \left\|P_{\text{EST}} B\right\| \left\|\xi_a^k\right\| \left\|u(\xi_a^k)\right\| \\ + \bar{\delta}^4 \gamma^2 \left\|B^\top P_{\text{EST}} B\right\| \left\|u(\xi_a^k)\right\|^2.\end{aligned} \tag{2.142}$$

Now, let us define the compact set $\mathcal{S}_i := \{x_i \in \mathbb{R}^n \mid \|x_i - x^\star\| \leq r_{\xi_a}'\} \subset \mathbb{R}^n$ and note that $\xi_a := (\tilde{x}_a, \tilde{z}_{\perp,a}) \in \Omega_{r_\xi} \implies \tilde{x} \in \mathcal{S} \subset \mathbb{R}^{Nn}$, where $\mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$. Then, we apply result (2.130) to claim that, for all $i \in \{1, \dots, N\}$, it holds $\ell_i(x_i) \leq L_{i,\mathcal{S}_i}$ for any $x_i \in \mathcal{S}_i$. Thus, by defining $L_{\mathcal{S}} := \max_i\{L_{1,\mathcal{S}_1}, \dots, L_{N,\mathcal{S}_N}\}$ and using the definition $u(\xi_a^k) = \ell(\tilde{\mathrm{x}}^k + \mathbf{1}_{N,n} x^\star)$, we get

$$\left\|u(\xi_a^k)\right\| \leq \sqrt{N} L_{\mathcal{S}}. \tag{2.143}$$

Hence, if $\gamma \in (0, 1]$ and $\delta \in (0, 1]$, we can bound (2.142) as

$$\Delta V(\xi_a^k) \leq -\gamma c_1 \|\xi_a^k\|^2 + \gamma \bar{\delta}^2 \left(b_1 \left\|\xi_a^k\right\| + b_2\right), \tag{2.144}$$

where we introduced

$$b_1 := 2\left\|P_{\text{EST}}B\right\|\sqrt{N}L_{\Omega_{r_\xi}} + 2a_4\left\|P_{\text{EST}}B\right\|\sqrt{N}L_{\Omega_{r_\xi}}$$
$$b_2 := \left\|B^\top P_{\text{EST}}B\right\| NL^2_{\Omega_{r_\xi}}.$$

Therefore, for any $\rho \in (0, r_\xi)$ and $c_1' \in (0, c_1)$, we define

$$\delta^\star := \min\left\{\sqrt{(c_1 - c_1')\rho^2/(b_1 r_\xi' + b_2)}, 1\right\}. \tag{2.145}$$

Consequently, by combining (2.144) and (2.145), we claim that, if $\delta_i \in (0, \delta^\star)$ for all $i \in \{1, \dots, N\}$, then, for any $\xi_a^k \in \Omega_{r_\xi'}$ such that $\left\|\xi_a^k\right\| \geq \rho$, it holds

$$\Delta V(\xi_a^k) < -\gamma c_1' \left\|\xi_a^k\right\|^2. \tag{2.146}$$

Thus, the inequality (2.146) ensures that the set Ω_{r_ξ} is invariant for system (2.129). Hence, if we pick $\xi_a^0 \in \mathcal{B}_{r_\xi}$, we prove that $\xi_a^k \in \Omega_{r_\xi}$ for all $k \geq 0$. Consequently, the bound (2.143) holds for all $k \geq 0$ and, in turn, also the inequality (2.146) is verified for all $k \geq 0$, namely we proved that the trajectories of system (2.129) enter the ball $\mathcal{B}_\rho$ exponentially fast. The result (2.141) follows from the inequality (2.146) and (2.136a) by setting $c_3 := c_1'/(2a_2)$. ■

Proof of Theorem 2.6

Since Assumptions 2.6, 2.7, and 2.8 hold, we apply Lemma 2.4 to claim that there exists $P : \mathbb{R}^{(2N-1)n \times (2N-1)n}$, $a_1, a_2, c_1 > 0$ such that, if $\gamma \in (0, \gamma_0)$, the conditions (2.136) are satisfied. Now, we evaluate the distance with respect to the origin of the initial conditions of system (2.128) and (2.129), i.e., $\left\|\xi^0\right\| = \left\|\xi_a^0\right\|$. By using the definition of ξ, the changes of variables (2.124) and (2.126), and the triangle inequality, we get

$$\begin{aligned}\left\|\xi^0\right\| &\leq \left\|\mathrm{x}^0 - \mathbf{1}_{N,n}x^\star\right\| + \left\|R^\top(\mathrm{z}^0 + G(\mathbf{1}_{N,n}x^\star))\right\| + \left\|\frac{\mathbf{1}_{N,n}^\top}{N}(\mathrm{z}^0 + G(\mathbf{1}_{N,n}x^\star))\right\| \\ &\overset{(a)}{\leq} r\sqrt{N} + \|R\|\left\|G(\mathbf{1}_{N,n}x^\star)\right\|,\end{aligned}$$

where in (a) we combine the initialization $\left\|\mathrm{x}_i^0 - x^\star\right\| \leq r$ and $\mathrm{z}_i^0 = 0$ for all $i \in \{1, \dots, N\}$ with the fact that $\mathbf{1}_{N,n}^\top G(\mathbf{1}_{N,n}x^\star) = \sum_{i=1}^N f_i(x^\star) = 0$. Hence, by defining $r_\xi := r\sqrt{N} + \|R\|\left\|G(\mathbf{1}_{N,n}x^\star)\right\|$, we claim that $\left\|\xi^0\right\| = \left\|\xi_a^0\right\| \leq r_\xi$. Once the initial distance from the origin has been evaluated, we choose any $\bar\rho > 0$, set $c_2 := \sqrt{a_2/a_1}$, and choose any $\epsilon \in (0, \bar\rho(1+c_2))$. Then, we pick $\rho \in (0, (\bar\rho - (1+c_2)\epsilon)/c_2)$, $c_1' \in (0, c_1)$, and use the matrix P_{EST} satisfying (2.136) to apply Lemma 2.5. Specifically, we claim that there exist $c_3 > 0$, and $\delta^\star \in (0, 1]$ such that, for any $\delta_i \in (0, \delta^\star)$ for all $i \in \{1, \dots, N\}$ and $\gamma \in (0, 1]$, it holds (i)

$\xi_a^k \in \mathcal{B}_{c_2 r_\xi}$ for all $k \geq 0$, and (ii) the inequality $\left\|\xi_a^k\right\| \leq c_2 \exp(-k\gamma c_3) \left\|\xi_a^0\right\|$, for any ξ_a^k such that $\left\|\xi_a^k\right\| \geq \rho$. Now, in order to bound $\|\xi^k - \xi_a^k\|$, let $\upsilon(k, \xi_a) := \sum_{q=0}^{k-1} (\phi(q, \xi_a) - \phi_a(\xi_a))$ and write

$$\upsilon(k+1, \xi_a^{k+1}) - \upsilon(k, \xi_a^k) = \phi(k, \xi_a^{k+1}) - \phi_a(\xi_a^{k+1}) + \upsilon(k, \xi_a^{k+1}) - \upsilon(k, \xi_a^k). \tag{2.147}$$

Then, let $r'_\xi := c_2 r_\xi$ and define $\Delta := \delta\sqrt{Nn}$. Under the assumption of $\xi^k \in \mathcal{B}_{r'_\xi + \epsilon}$ for all $k \geq 0$ (later verified by a proper selection of γ), we claim that the arguments of the functions f_i and their derivatives (embedded into the definitions of $\phi(k, \cdot)$ and $\phi_a(\cdot)$ and their derivatives) lie into the compact set $\mathcal{B}_{r'_\xi + \epsilon + \Delta}$. Thus, since the functions f_i and its derivatives are continuous (cf. Assumption 2.8) and the functions $\phi(\cdot, \cdot)$ and $\nu(\cdot, \cdot)$ are periodic in the first argument, we define

$$L_\phi := \sup_{\substack{\xi \in \mathcal{B}_{r'_\xi + \epsilon} \\ k \in [0, \tau]}} \left\{ \|\phi(k, \xi)\|, \|\phi_a(\xi)\|, \left\| \frac{\partial \phi(k, \xi)}{\partial \xi} \right\|, \left\| \frac{\partial \phi_a(\xi)}{\partial \xi} \right\|, \left\| \frac{\partial \nu(k, \xi)}{\partial \xi} \right\| \right\}.$$

Consequently, it holds

$$\|\upsilon(k, \xi)\| \leq 2L_\phi \tau \tag{2.148a}$$
$$\|\phi(k, \xi) - \phi(k, \xi')\| \leq L_\phi \|\xi - \xi'\| \tag{2.148b}$$
$$\|\phi_a(\xi) - \phi_a(\xi')\| \leq L_\phi \|\xi - \xi'\| \tag{2.148c}$$
$$\|\upsilon(k, \xi) - \upsilon(k, \xi')\| \leq 2L_\phi \tau \|\xi - \xi'\| \tag{2.148d}$$
$$\|\phi_a(\xi)\| \leq L_\phi, \tag{2.148e}$$

for any $\xi, \xi' \in \mathcal{B}_{r'_\xi + \epsilon}$ and $k \geq 0$. Let $\eta^k := \xi_a^k + \gamma \upsilon(k, \xi_a^k)$ and write

$$\xi^k - \eta^k = \sum_{q=0}^{k-1} (\xi^{q+1} - \xi^q) - (\eta^{q+1} - \eta^q),$$

add and subtract $\gamma \sum_{q=0}^{k-1} (\phi(q, \xi^q) + \phi(q, \xi_a^q))$, and use (2.147) to get

$$\begin{aligned} \xi^k - \eta^k = {} & \gamma \sum_{q=0}^{k-1} (\phi(q, \xi^q) - \phi(q, \eta^q)) + \gamma \sum_{q=0}^{k-1} (\phi(q, \eta^q) - \phi(q, \xi_a^q)) \\ & - \gamma \sum_{q=0}^{k-1} (\phi(q, \xi_a^{q+1}) - \phi(q, \xi_a^q)) + \gamma \sum_{q=0}^{k-1} (\phi_a(\xi_a^{q+1}) - \phi_a(\xi_a^q)) \\ & - \gamma \sum_{q=0}^{k-1} (\upsilon(q, \xi_a^{q+1}) - \upsilon(q, \xi_a^q)). \end{aligned}$$

Use (2.128), (2.129), and (2.148) to bound

$$\|\xi^k - \eta^k\| \leq \gamma L_\phi \sum_{q=0}^{k-1} \|\xi^q - \eta^q\| + \gamma^2 L_\phi^2 2\,(1+2\tau)\,k.$$

Apply the discrete Gronwall inequality (see [84, 153]) and

$$\sum_{q=0}^{k-1} \gamma L_\phi q \exp\left(-\gamma L_\phi q\right) \leq \sum_{q=0}^{\infty} \gamma L_\phi q \exp\left(-\gamma L_\phi q\right) = 1$$

to get

$$\|\xi^k - \eta^k\| \leq \gamma^2 L_\phi^2 2\,(1+2\tau)\,k + \gamma L_\phi 2\,(1+2\tau)\exp\left(\gamma L_\phi k\right),$$

from which

$$\|\xi^k - \eta^k\| \leq \gamma^2 L_\phi^2 2\,(1+2\tau)\,k + \gamma L_\phi 2\,(1+2\tau)\exp\left(\gamma L_\phi k\right) + \gamma 2 L_\phi \tau.$$

Then, set $\theta^\star \in \mathbb{N}$ such that

$$\theta^\star \geq -\frac{1}{c_3}\ln\left(\frac{(\bar\rho - \epsilon)/c_2}{c_2 r_\xi}\right). \tag{2.149}$$

Let $\gamma_2 := \frac{\epsilon/3}{L_\phi^2 2(1+2\tau)\theta^\star}$, $\gamma_3 := \frac{\epsilon/3}{2L_\phi(1+2\tau)\exp(L_\phi\theta^\star)}$,$\gamma_4 := \frac{\epsilon/3}{2L_\phi\tau}$, $\gamma_1 := \min\{\gamma_2, \gamma_3, \gamma_4, 1\}$, and $k^\star := \theta^\star/\gamma$. Then, for any $\gamma \in (0, \gamma_1)$, it holds

$$\left\|\xi^k - \xi_a^k\right\| \leq \epsilon, \tag{2.150}$$

for all $k \in [0, k^\star]$. As a consequence, since $\xi_a^k \in \mathcal{B}_{r'_\xi}$ for all $k \geq 0$, it holds $\xi^k \in \mathcal{B}_{r'_\xi + \epsilon}$ for all $k \in [0, k^\star]$, i.e., we have verified that the bounds (2.148) can be used into the interval $[0, k^\star]$. Moreover, the exponential law (2.141) and the expression of $\theta^\star$ (cf. (2.149)) ensure that it holds

$$\|\xi_a^k\| \leq (\bar\rho - \epsilon)/c_2, \tag{2.151}$$

for all $k \geq k^\star$. Now, by using the triangle inequality, we write

$$\|\xi^{k^\star}\| \leq \|\xi^{k^\star} - \xi_a^{k^\star}\| + \|\xi_a^{k^\star}\| \overset{(a)}{\leq} \bar\rho/c_2, \tag{2.152}$$

where in (a) we combined (2.150) and (2.151). The inequality (2.152) guarantees that $\xi^{t^\star} \in \mathcal{B}_{\bar\rho/c_2}$, hence we proved that the trajectories of (2.129) enters into $\mathcal{B}_{\bar\rho/c_2}$ with linear rate. Next, in order to show that $\xi^k \in \mathcal{B}_{\bar\rho}$ for any $k \geq k^\star$, we divide the set of natural numbers in intervals as $\mathbb{N} = [0,\ k^\star] \cup [k^\star, 2k^\star] \cup \ldots$. Define $\psi_a(q + k^\star, \xi^{k^\star})$ as the solution

to (2.129) for $\xi_a^0 = \xi^{k^\star}$ and $q \in [0, k^\star]$. Thus, at the beginning of the time interval $[k^\star, 2k^\star]$, the initial condition of the trajectory of (2.129) (i) coincides with the one of $\psi_a(q+k^\star, \xi^{k^\star})$, and (ii) lies into $\mathcal{B}_{\bar{\rho}} \subseteq \mathcal{B}_{r_\xi}$. Thus, we apply the same arguments above to guarantee that, for any $\gamma \in (0, \gamma^\star)$, it holds (i) $\left\|\xi^{q+k^\star} - \psi_a(q+k^\star, \xi^{k^\star})\right\| \leq \epsilon$, for all $q \in [0, k^\star]$, and (ii) $\psi_a(2k^\star, \xi^{k^\star}) \in \mathcal{B}_{(\bar{\rho}-\epsilon)/c_2}$. Moreover, using the arguments of Lemma 2.5, we guarantee that system (2.129) cannot escape from the set $\mathcal{B}_{\bar{\rho}-\epsilon}$, namely $\xi_a^k \in \mathcal{B}_{\bar{\rho}-\epsilon}$ for all $k \geq k^\star$. Thus, we get $\xi^k \in \mathcal{B}_{\bar{\rho}}$ for all $k \in [k^\star, 2k^\star]$. The proof follows by recursively applying the same arguments above for each time interval $[jk^\star, (j+1)k^\star]$ with $j = 2, 3, \ldots$, and by using the trivial inequality $\left\|\mathrm{x}_i^k - x^\star\right\| \leq \left\|\xi^k\right\|$ for all $i \in \{1, \ldots, N\}$ and $k \geq 0$.

2.5.2 Numerical Simulations

To corroborate the theoretical analysis, in this section we provide numerical computations for the proposed distributed algorithm on a personalized optimization framework.

In several engineering applications, a problem of interest consists of optimizing a performance metric while keeping into account user discomfort terms [146, 203]. In these scenarios, the user discomfort term is usually not known in advance but can be only accessed by measurements. Specifically, we associate to each agent $i \in \{1, \ldots, N\}$ a cost function in the form $f_i(w) = w^\top Q_i w + r_i^\top w + \log(\sum_{\ell=1}^n a_{i\ell} e^{b_{i\ell} w_\ell})$ with $Q_i = Q_i^\top \in^{n \times n}$, $r_i \in^n$ and $a_{i\ell}, b_{i\ell}$, for all $\ell \in \{1, \ldots, n\}$. For all $i \in \{1, \ldots, N\}$ and $\ell \in \{1, \ldots, n\}$, we uniformly randomly choose the eigenvalues of Q_i from the interval $[10^{-3}, 5 \cdot 10^{-3}]$, the components of r_i within the interval $[-10^{-2}, 3 \cdot 10^{-2}]$, and the parameters $a_{i\ell}, b_{i\ell}$ within the interval $[0, 10^{-3}]$. Agents communicate according to Erdős-Rényi random graphs with edge probabilities equal to 0.2. We choose the parameters τ_{i_p} and ϕ_{i_p} as follows. Define $\mathbb{O}_3$ as the set of odd umbers greater than 3. Then, for all $i \in \{1, \ldots, N\}$, we take $\delta_i = 0.2$, $\phi_{i_p} = \dfrac{\pi}{4}(1 + (-1)^p)$ for all $p = 1, \ldots, n$, while τ_{i_p} have been chosen as the first $\lfloor (n+1)/2 \rfloor$ elements of $\mathbb{O}_3$. Simulations are performed using DISROPT [63], a Python package based on MPI which provides libraries to encode and simulate distributed optimization algorithms.

In the first set of runs, we consider for the number of agents the values $N = 5, 10, 20, 30$. For each value of N, we generate 50 random instances. We generate communication graphs with a diameter d such that the ratio N/d is constant while varying N. Results are depicted in Fig. 2.11, where for each problem instance, we evaluated the relative errors $|\sum_i f_i(\bar{\mathrm{x}}^k) - f(x^\star)|/|f(x^\star)|$ and $\|\bar{\mathrm{x}}^k - x^\star\|/\|x^\star\|$, where $\bar{\mathrm{x}}^k := \frac{1}{N}\sum_{i=1}^N \mathrm{x}_i^k$.

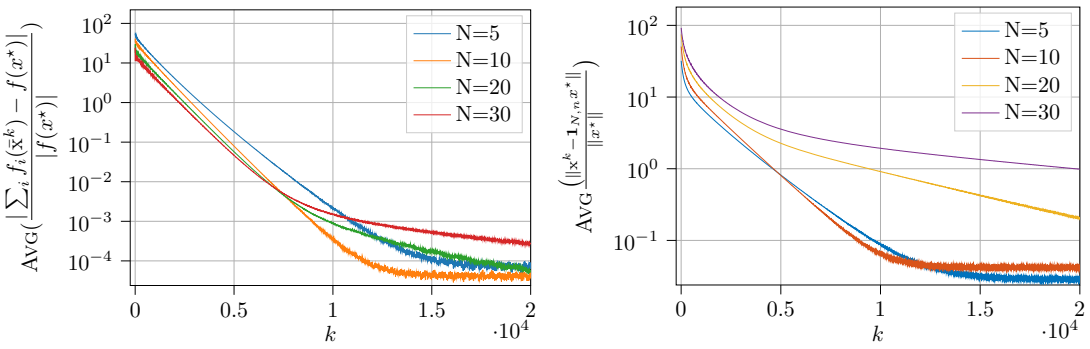


Figure 2.11: Cost error (left) and optimization variable error (right) in Monte Carlo simulations for varying number of agents.

Then, we perform numerical simulations over a larger network made of $N = 250$ agents. We consider different optimization variable sizes, namely $n = 10, 20$. For each value of N, we generate 20 random instances. Part of these performances has been run on the Marconi100 HPC Cluster of the Italian Cineca. We used 10 nodes of the cluster and, for each node, we used 25 cores and 4 GPUs. The code has been adapted in order to perform part of the computation directly on GPUs. The results are shown in Fig. 2.12.

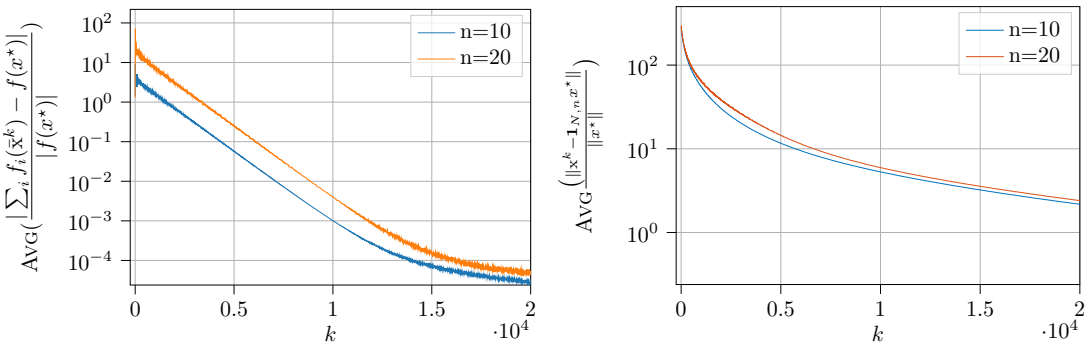


Figure 2.12: Cost error (left) and optimization variable error (right) in Monte Carlo simulations for different problem size and 250 agents.

To conclude, we perform a numerical comparison against with the zero-order distributed scheme *Algorithm 1* in [179]. At each communication round, this algorithm estimates the gradient with two queries of the objective function, which is similar to the one-query estimation in our proposed algorithm. We run *Algorithm 1* in [179] on the same set of simulations with $n = 10$ and $N = 10$. We used the same communication graphs, cost functions and initial conditions. Results are in Fig. 2.13. As it can be seen, our scheme performs better.

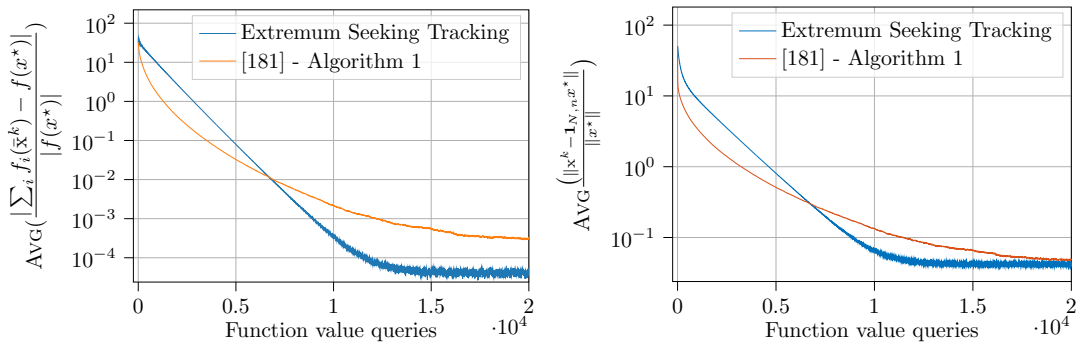


Figure 2.13: Cost error (left) and optimization variable error (right) comparison between the proposed algorithm and a zero order scheme.

In detail, increasing the number of agents increases the Lipschitz constant of the system to be averaged. Moreover, a larger domain of initial conditions also implies a potentially larger L_ϕ constant (cf. Section 2.5.1). This implies smaller $\gamma^\star$, which, fixed the other parameters, makes the convergence slower. The decision variable dimension instead impacts the selection of the dither signal. A larger number of states implies a larger number of frequencies. This, in turn, means a longer time to estimate the gradient (cf. Lemma 2.3). Notice that, however, the accuracy of the final estimate is guaranteed by design. Indeed, since δ and γ are designed on $\bar{\rho}$, the trajectories of (5) converge to a ball of radius $\bar{\rho}$ independently of the dimensions of the optimization problem.

2.6 Distributed Online Consensus Optimization

In this section, we address online instances of (2.1), namely optimization problems in the form

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{N} f_i^k(x), \quad k \geq 0, \tag{2.153}$$

where each $f_i^k : \mathbb{R}^n \to \mathbb{R}$ is a local function revealed only to agent i at iteration k. In the following, we let $f^k(x) := \sum_{i=1}^N f_i^k(x)$.

We address the distributed solution of the online optimization problem (2.153) in terms of *dynamic regret* (see, e.g., [105]). In particular, let x_i^k be the solution estimate of the problem at time t maintained by agent i, and let $x_\star^k$ be a minimizer of $\sum_{i=1}^N f_i^k$. Then, the agents want to minimize the dynamic regret defined as

$$R_T := \sum_{k=1}^{T} f^k(\bar{\mathrm{x}}^k) - \sum_{k=1}^{T} f^k(x_\star^k), \tag{2.154}$$

for a finite value $T > 1$ with $\bar{\mathrm{x}}^k := \frac{1}{N}\sum_{i=1}^N \mathrm{x}_i^k$.

Another possible performance metric is the so-called static regret (see, e.g., [105]).

The dynamic regret (2.154) is known to be more challenging than the static one [105] and, for this reason, consistently with the majority of the recent papers in literature, this work focuses on the dynamic regret (2.154). As it is customary in the distributed setting, we also complement these measures with the consensus metric $\sum_{i=1}^{N} \left\|\mathrm{x}_i^T - \bar{\mathrm{x}}^T\right\|^2$, quantifying how far from consensus the local decisions are.

Along this section, we enforce the following assumptions.

Assumption 2.9 (Lipschitz continuous gradients)**.** *The functions f_i^k have $\overline{L}$-Lipschitz continuous gradients for all $i \in \{1, \dots, N\}$ and $k \geq 0$.* △

Assumption 2.10 (Strong convexity)**.** *The functions f_i^k are μ-strongly convex for all $i \in \{1, \dots, N\}$ and $k \geq 0$.* △

Finally, the following characterizes the communication structure.

Assumption 2.11 (Network Structure)**.** *The weighted graph $\mathcal{G}$ is connected with doubly stochastic matrix $\mathcal{W}_{\mathcal{G}}$ stochastic.* △

We point out that, in light of Assumption 2.10, the minimizer $x_\star^k$ is unique for all $k \geq 0$ (cf. Proposition A.2 in Appendix A).

In order to address in a distributed fashion problem (2.153), we propose GTAdam, i.e., a novel method taking inspiration both from the Gradient Tracking (see Section 2.2) distributed algorithm and Adam.

Adam centralized algorithm

Adam [94] is an optimization algorithm that solves problems in the form (2.153) in a *centralized* computation framework. It is an iterative gradient-like procedure in which, at each iteration k, a solution estimate x^k is updated by means of a descent direction which is enhanced by a proper use of the gradient history, i.e., through estimates of their first- and second-order momenta. Specifically, the (time-varying) gradient $\mathrm{g}^k = \nabla f^k(\mathrm{x}^k)$ of the function drives two exponential moving average estimators. The two estimates, denoted by m^k and v^k, represent, respectively, mean and variance (1^{st} and 2^{nd} momentum) of the gradient sequence and are nonlinearly combined to build the descent direction. A pseudo-code of Adam algorithm is reported in Algorithm 3 in which $\gamma > 0$ is the step-size, the constant $0 < \epsilon \ll 1$ is introduced to guarantee numerical robustness of the scheme, while the hyper-parameters $\beta_1, \beta_2 \in (0, 1)$ control the exponential-decay rate of the moving average dynamics.

We point out that in the algorithm above the ratio $\frac{\mathrm{m}^{k+1}}{\sqrt{\mathrm{v}^{k+1}+\epsilon}}$ is meant element-wise. Typical choices for the algorithmic parameters are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

Algorithm 3 Adam

initialization: $\mathrm{x}^0 \in \mathbb{R}^n$, $\mathrm{m}^0 = \mathrm{v}^0 = 0$, $\mathrm{g}^0 = \nabla f^0(\mathrm{x}^0)$
for $k = 1, 2 \dots$ **do**
$\mathrm{m}^{k+1} = \beta_1 \mathrm{m}^k + (1-\beta_1)\mathrm{g}^k$
$\mathrm{v}^{k+1} = \beta_2 \mathrm{v}^k + (1-\beta_2)\mathrm{g}^k \odot \mathrm{g}^k$
$\mathrm{x}^{k+1} = \mathrm{x}^k - \gamma \frac{\sqrt{1-\beta_2}}{1-\beta_1} \frac{\mathrm{m}^{k+1}}{\sqrt{\mathrm{v}^{k+1}+\epsilon}}$
$\mathrm{g}^{k+1} = \nabla f^{k+1}(\mathrm{x}^{k+1})$
end for

2.6.1 GTAdam: Algorithm Description and Analysis

In this section, we present GTAdam. Along the evolution of the algorithm, each agent i maintains four local states:

(i) a local estimate x_i^k of the current optimal solution $x_\star^k$;

(ii) an auxiliary variable s_i^k whose role is to track the gradient of the whole cost function;

(iii) an estimate m_i^k of the 1^{st} momentum of s_i^k;

(iv) an estimate v_i^k of the 2^{nd} momentum of s_i^k.

The momentum estimates of s_i^k are initialized as $\mathrm{m}_i^0 = \mathrm{v}_i^0 = 0$, while the tracker of the gradient is initialized as $\mathrm{s}_i^0 = \nabla f_i^0(\mathrm{x}_i^0)$.

The algorithm works as follows. At each iteration k, each agent i performs the following operations

(i) it updates the moving averages m_i^k and v_i^k;

(ii) it computes a weighted average of the solution estimates of its neighbors and, starting from this point, it uses the update direction $\frac{\mathrm{m}_i^{k+1}}{\sqrt{\mathrm{v}_i^{k+1}+\epsilon}}$ to compute the new solution estimate x_i^{k+1};

(iii) it updates the local gradient tracker s_i^k via a "dynamic consensus" mechanism.

A pseudo-code of GTAdam is reported in Algorithm 4.

Some remarks are in order. The algorithm proposed in this paper is different from [134]. In fact, although they both use a similar strategy involving first- and second-order momenta, in that work only local gradients are considered, without resorting to any tracking mechanism. Note that a saturation term $G \gg 0$ is introduced in the update of v_i^k, where the $\min$ operator is to be intended element-wise. The value of G guarantees a bound for the scaling factor that multiplies the descent direction.

Algorithm 4 GTAdam (for agent i)

initialization: $\mathrm{x}_i^0 \in \mathbb{R}^n$, $\mathrm{s}_i^0 = \mathrm{g}_i^0 = \nabla f_i^0(\mathrm{x}_i^0)$, $\mathrm{m}_i^0 = \mathrm{v}_i^0 = 0$
for $k = 1, \dots, T$ **do**
$\mathrm{m}_i^{k+1} = \beta_1 \mathrm{m}_i^k + (1-\beta_1)\mathrm{s}_i^k$
$\mathrm{v}_i^{k+1} = \min\{\beta_2 \mathrm{v}_i^k + (1-\beta_2)\mathrm{s}_i^k \odot \mathrm{s}_i^k, G\}$
$\mathrm{x}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{x}_j^k - \gamma \frac{\mathrm{m}_i^{k+1}}{\sqrt{\mathrm{v}_i^{k+1}+\epsilon}}$
$\mathrm{g}_i^{k+1} = \nabla f_i^{k+1}(\mathrm{x}_i^{k+1})$
$\mathrm{s}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{s}_j^k + \mathrm{g}_i^{k+1} - \mathrm{g}_i^k$
end for

Such a bound will turn out to be important for analysis purposes. We suggest to take it proportional to the initial estimates v_i^0.

In order to analyze GTAdam, we rewrite it into an aggregate form. We define $\mathrm{x}^k := \mathrm{col}(\mathrm{x}_1^k, \dots, \mathrm{x}_N^k)$ and their average as $\bar{\mathrm{x}}^k := \frac{1}{N}\sum_{i=1}^N \mathrm{x}_i^k$. Similar definitions apply to the quantities $\mathrm{m}^k, \mathrm{v}^k, \mathrm{d}^k, \mathrm{g}^k, \mathrm{s}^k$ and their averages $\bar{\mathrm{m}}^k, \bar{\mathrm{v}}^k, \bar{\mathrm{d}}^k, \bar{\mathrm{s}}^k$. With these definitions at hand, GTAdam can be rephrased from a global perspective as

$$\mathrm{m}^{k+1} = \beta_1 \mathrm{m}^k + (1-\beta_1)\mathrm{s}^k \tag{2.155a}$$
$$\mathrm{v}^{k+1} = \min\{\beta_2 \mathrm{v}^k + (1-\beta_2)\mathrm{s}^k \odot \mathrm{s}^k, \mathbf{1}_{N,n} G\} \tag{2.155b}$$
$$\mathrm{d}^{k+1} = (\mathrm{V}^{k+1} + \epsilon I)^{-1/2}\mathrm{m}^{k+1} \tag{2.155c}$$
$$\mathrm{x}^{k+1} = \mathcal{W}\mathrm{x}^k - \gamma \mathrm{d}^{k+1} \tag{2.155d}$$
$$\mathrm{s}^{k+1} = \mathcal{W}\mathrm{s}^k + \mathrm{g}^{k+1} - \mathrm{g}^k, \tag{2.155e}$$

where we set $W := \mathcal{W} \otimes I_n$, $\mathrm{V}^k := \mathrm{diag}(\mathrm{v}^k)$, and $\bar{\mathrm{V}}^k := \mathrm{diag}(\bar{\mathrm{v}}^k)$. Moreover, the averaged quantities of (2.155) satisfy

$$\bar{\mathrm{m}}^{k+1} = \beta_1 \bar{\mathrm{m}}^k + (1-\beta_1)\bar{\mathrm{s}}^k \tag{2.156a}$$
$$\bar{\mathrm{v}}^{k+1} = \min\{\beta_2 \bar{\mathrm{v}}^k + (1-\beta_2)\bar{\mathrm{s}}^k \odot \bar{\mathrm{s}}^k, G\} \tag{2.156b}$$
$$\bar{\mathrm{d}}^{k+1} = \frac{1}{N}\mathbf{1}_{N,n}^\top \mathrm{d}^{k+1} \tag{2.156c}$$
$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k - \gamma \bar{\mathrm{d}}^{k+1} \tag{2.156d}$$
$$\bar{\mathrm{s}}^{k+1} = \bar{\mathrm{s}}^k + \frac{1}{N}\sum_{i=1}^N (\mathrm{g}_i^{k+1} - \mathrm{g}_i^k). \tag{2.156e}$$

Our analysis is based on studying the aggregate dynamical evolution of the following: average first momentum $\|\bar{\mathrm{m}}^k\|$, average tracking momentum difference $\|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\|$, first momentum error $\|\mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k\|$, gradient tracking error $\|\mathrm{s}^k - \mathbf{1}_{N,n}\bar{\mathrm{s}}^k\|$, consensus error $\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\|$ and solution error $\|\bar{\mathrm{x}}^k - x_\star^k\|$. Let y^k be the vector stacking the above

quantities at iterations k

$$\mathrm{y}^k := \begin{bmatrix} \|\bar{\mathrm{m}}^k\| \\ \|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| \\ \|\mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k\| \\ \|\mathrm{s}^k - \mathbf{1}_{N,n}\bar{\mathrm{s}}^k\| \\ \|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\| \\ \|\bar{\mathrm{x}}^k - x_\star^k\| \end{bmatrix}. \tag{2.157}$$

Notice that, due to the distributed context and no assumptions on the boundedness of the gradients, we need to take into account all these quantities to study the convergence. Let us introduce two useful variables that will be used to provide the main result of the paper, namely

$$\begin{aligned} \eta^k &:= \sup_i \sup_{x\in\mathbb{R}^n} \|\nabla f_i^{k+1}(x) - \nabla f_i^k(x)\|, \\ \zeta^k &:= \|x_\star^{k+1} - x_\star^k\|. \end{aligned} \tag{2.158}$$

We now give a sequence of intermediate results, providing proper bounds on the components of y^k (defined in (2.157)), that are then used as building blocks for proving the main result regarding GTAdam, i.e., an upper bound for the dynamic regret.

Preparatory Lemmas

Lemma 2.6 (Average first momentum magnitude)**.** *Let Assumption 2.9 holds. Then, for all $k \geq 1$, it holds*

$$\left\|\bar{\mathrm{m}}^{k+1}\right\| \leq \beta_1 \left\|\bar{\mathrm{m}}^k\right\| + \frac{(1-\beta_1)\overline{L}}{\sqrt{N}} \left\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\right\| + (1-\beta_1)\overline{L} \left\|\bar{\mathrm{x}}^k - x_\star^k\right\|.$$

Proof. By using the update (2.156a), we can write

$$\left\|\bar{\mathrm{m}}^{k+1}\right\| = \left\|\beta_1\bar{\mathrm{m}}^k + (1-\beta_1)\bar{\mathrm{s}}^k\right\| \leq \beta_1 \left\|\bar{\mathrm{m}}^k\right\| + (1-\beta_1)\left\|\bar{\mathrm{s}}^k\right\|, \tag{2.159}$$

in which we use the triangle inequality. Regarding the term $\left\|\bar{\mathrm{s}}^k\right\|$, we use the relation $\bar{\mathrm{s}}^k = \frac{1}{N}\sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k)$, and we add $\frac{1}{N}\sum_{i=1}^N \nabla f_i^k(x_\star^k) = 0$, thus obtaining

$$\begin{aligned} \|\bar{\mathrm{s}}^k\| &= \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) - \frac{1}{N}\sum_{i=1}^N \nabla f_i^k(x_\star^k)\right\| \overset{(a)}{\leq} \frac{\overline{L}}{N}\sum_{i=1}^N \left\|\mathrm{x}_i^k - x_\star^k\right\| \\ &\overset{(b)}{\leq} \frac{\overline{L}}{\sqrt{N}} \left\|\mathrm{x}^k - \mathbf{1}_{N,n}x_\star^k\right\| \overset{(c)}{\leq} \frac{\overline{L}}{\sqrt{N}} \left\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\right\| + \overline{L}\left\|\bar{\mathrm{x}}^k - x_\star^k\right\|, \end{aligned} \tag{2.160}$$

where in (a) we exploit the Lipschitz continuity of the gradients of the cost functions

(cf. Assumptions 2.9), in (b) we use the basic algebraic property $\sum_{i=1}^{N} \|\theta_i\| \leq \sqrt{N}\|\theta\|$ for a generic vector $\theta := \mathrm{col}(\theta_1, \dots, \theta_N)$, and in (c) we add and subtract the term $\mathbf{1}_{N,n}\bar{\mathrm{x}}^k$ and apply the triangle inequality. The proof follows by combining the bounds (2.159) and (2.160). ■

Lemma 2.7 (First momentum error)**.** *For all $k \geq 1$, it holds*

$$\left\|\mathrm{m}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}\right\| \leq \beta_1 \left\|\mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k\right\| + (1-\beta_1)\left\|\mathrm{s}^k - \mathbf{1}_{N,n}\bar{\mathrm{s}}^k\right\|.$$

The proof of Lemma 2.7 follows by combining (2.155a) and (2.156a) with the triangle inequality.

Lemma 2.8 (Input signal error)**.** *For all $k \geq 0$, it holds*

$$\begin{aligned}\|\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\| \leq{}& \frac{\beta_1\sqrt{N}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{m}}^k\right\| + \frac{\beta_1}{\sqrt{\epsilon}}\left\|\mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k\right\| + \frac{(1-\beta_1)}{\sqrt{\epsilon}}\left\|\mathrm{s}^k - \bar{\mathrm{s}}^k\right\| \\ &+ \frac{(1-\beta_1)\overline{L}}{\sqrt{\epsilon}}\left\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\right\| + \frac{(1-\beta_1)\beta_1\overline{L}\sqrt{N}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{x}}^k - x_\star^k\right\|.\end{aligned}$$

Proof. By using (2.155c) and (2.156c), one has

$$\begin{aligned}\left\|\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\right\| &= \left\|\left(I - \frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right)(\mathrm{V}^{k+1} + \epsilon I)^{-1/2}\mathrm{m}^{k+1}\right\| \\ &\overset{(a)}{\leq} \left\|(\mathrm{V}^{k+1} + \epsilon I)^{-1/2}\right\|\left\|\mathrm{m}^{k+1}\right\| \overset{(b)}{\leq} \frac{1}{\sqrt{\epsilon}}\left\|\mathrm{m}^{k+1}\right\| \\ &\overset{(c)}{\leq} \frac{1}{\sqrt{\epsilon}}\left\|\mathrm{m}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}\right\| + \frac{\sqrt{N}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{m}}^{k+1}\right\|, \qquad (2.161)\end{aligned}$$

where in (a) we apply the Cauchy-Schwarz inequality combined with $\left\|I - \frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right\| \leq 1$, in (b) we use the bound $\left\|(\mathrm{V}^{k+1} + \epsilon I)^{-1/2}\right\| \leq \frac{1}{\sqrt{\epsilon}}$ (justified by the fact that $\mathrm{v}^k \geq 0$ for all $k \geq 0$), in (c) we add and subtract within the norm $\mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}$ and apply the triangle inequality and an algebraic property. The proof follows by using Lemma 2.6 and 2.7 in (2.161). ■

Lemma 2.9 (Tracking error)**.** *Let Assumptions 2.9, 2.10, and 2.11 hold. Then, for all $k \geq 0$, it holds*

$$\begin{aligned}\|\mathrm{s}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{s}}^{k+1}\| \leq{}& \left(\Lambda + \gamma\frac{2(1-\beta_1)\overline{L}}{\sqrt{\epsilon}}\right)\left\|\mathrm{s}^k - \mathbf{1}_{N,n}\bar{\mathrm{s}}^k\right\| \\ &+ \gamma\frac{2\beta_1\overline{L}\sqrt{N}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{m}}^k\right\| + \gamma\frac{2\beta_1\overline{L}}{\sqrt{\epsilon}}\left\|\mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k\right\| \\ &+ \left(\overline{L}\,\|W - I\| + \gamma\frac{2(1-\beta_1)\beta_1 L^2}{\sqrt{\epsilon}}\right)\left\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\right\|\end{aligned}$$

$$+\gamma\frac{(1-\beta_1)(1+\beta_1)\overline{L}^2\sqrt{N}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{x}}^k-x_\star^k\right\|+\sqrt{N}\eta^k.$$

where $\Lambda\in(0,1)$ *is the spectral radius of* $\mathcal{W}-\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}$ *and* η^k *has been defined in* (2.158).

Proof.

By combining (2.155e) and (2.156e) one has

$$\begin{aligned}\|\mathrm{s}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{s}}^{k+1}\| &= \left\|\mathcal{W}\mathrm{s}^k+\mathrm{g}^{k+1}-\mathrm{g}^k-\mathbf{1}_{N,n}\Big(\bar{\mathrm{s}}^k+\frac{1}{N}\sum_{i=1}^N(\mathrm{g}_i^{k+1}-\mathrm{g}_i^k)\Big)\right\|\\ &\overset{(a)}{\leq}\left\|\left(\mathcal{W}-\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right)(\mathrm{s}^k-\mathbf{1}_{N,n}\bar{\mathrm{s}}^k)\right\|+\left\|\left(I-\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right)(\mathrm{g}^{k+1}-\mathrm{g}^k)\right\|\\ &\overset{(b)}{=}\Lambda\left\|\mathrm{s}^k-\mathbf{1}_{N,n}\bar{\mathrm{s}}^k\right\|+\left\|\mathrm{g}^{k+1}-\mathrm{g}^k\right\|,\end{aligned}\tag{2.162}$$

where (a) uses $\mathbf{1}_{N,n}\in\ker\left(\mathcal{W}-\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right)$ and the triangle inequality, and (b) combines the Cauchy-Schwarz inequality with the bounds $\left\|\mathcal{W}-\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right\|\leq\Lambda$ and $\left\|I-\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right\|\leq 1$. Let $\tilde{\mathrm{g}}^k:=\mathrm{col}(\nabla f_1^{k+1}(x_1^k),\dots,\nabla f_N^{k+1}(x_N^k))$ and manipulate the term $\left\|\mathrm{g}^{k+1}-\mathrm{g}^k\right\|$ in (2.162) as

$$\begin{aligned}\|\mathrm{g}^{k+1}-\mathrm{g}^k\| &\leq \|\mathrm{g}^{k+1}-\tilde{\mathrm{g}}^k\|+\left\|\tilde{\mathrm{g}}^k-\mathrm{g}^k\right\|\\ &\overset{(a)}{\leq} L\|\mathrm{x}^{k+1}-\mathrm{x}^k\|+\|\tilde{\mathrm{g}}^k-\mathrm{g}^k\|\overset{(b)}{\leq} L\left\|\mathrm{x}^{k+1}-\mathrm{x}^k\right\|+\sqrt{N}\eta^k\\ &\overset{(c)}{=} L\|\mathcal{W}\mathrm{x}^k-\gamma\mathrm{d}^{k+1}-\mathrm{x}^k\|+\sqrt{N}\eta^k,\end{aligned}\tag{2.163}$$

where in (a) we use the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.9), (b) uses the variable η^k (cf (2.158)), and (c) uses the update (2.155d) of x^{k+1}. Let us manipulate the first term on the right-hand side of (2.163):

$$\begin{aligned}\left\|\mathcal{W}\mathrm{x}^k-\gamma\mathrm{d}^{k+1}-\mathrm{x}^k\right\| &\overset{(a)}{=}\left\|(\mathcal{W}-I)(\mathrm{x}^k-\mathbf{1}_{N,n}\bar{\mathrm{x}}^k)-\gamma\mathrm{d}^{k+1}\right\|\\ &\overset{(b)}{\leq}\|\mathcal{W}-I\|\|\mathrm{x}^k-\mathbf{1}_{N,n}\bar{\mathrm{x}}^k\|+\gamma\|\mathrm{d}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\|+\gamma\|\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\|,\end{aligned}\tag{2.164}$$

where (a) uses the fact that $\ker(\mathcal{W}-I)=\mathrm{span}(\mathbf{1}_{N,n})$ and in (b) we add and subtract the term $\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}$ within the norm and we apply the triangle inequality and the Cauchy-Schwarz inequality. Regarding $\|\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\|$, we use (2.155c) and (2.156c) to write

$$\|\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\|=\left\|\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\mathrm{d}^{k+1}\right\|=\left\|\frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}(\mathrm{V}^{k+1}+\epsilon I)^{-1/2}\mathrm{m}^{k+1}\right\|$$

$$\overset{(a)}{\leq} \frac{1}{\sqrt{\epsilon}} \left\| \mathrm{m}^{k+1} \right\| \overset{(b)}{\leq} \frac{1}{\sqrt{\epsilon}} \left\| \mathrm{m}^{k+1} - \mathbf{1}_{N,n} \bar{\mathrm{m}}^{k+1} \right\| + \frac{\sqrt{N}}{\sqrt{\epsilon}} \left\| \bar{\mathrm{m}}^{k+1} \right\|, \tag{2.165}$$

where in (a) we apply the Cauchy-Schwarz inequality and the bounds $\left\| \frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N} \right\| \leq 1$ and $\left\| (\mathrm{V}^{k+1} + \epsilon)^{-1/2} \right\| \leq \frac{1}{\sqrt{\epsilon}}$, in (b) we add and subtract within the norm the term $\mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}$, apply the triangle inequality, and use an algebraic property. By combining (2.164) and (2.165), we bound (2.163) as

$$\begin{aligned} \|\mathrm{g}^{k+1} - \mathrm{g}^k\| \leq \overline{L}\|\mathcal{W} - I\|\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\| + \gamma\overline{L}\|\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\| \\ + \gamma\frac{\overline{L}}{\sqrt{\epsilon}} \left\| \mathrm{m}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1} \right\| + \gamma\frac{\overline{L}\sqrt{N}}{\sqrt{\epsilon}} \left\| \bar{\mathrm{m}}^{k+1} \right\| + \sqrt{N}\eta^k. \end{aligned} \tag{2.166}$$

Now, by using the bound (2.166) within (2.164), we get

$$\begin{aligned} \left\| \mathrm{s}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{s}}^{k+1} \right\| \leq \Lambda \left\| \mathrm{s}^k - \mathbf{1}_{N,n}\bar{\mathrm{s}}^k \right\| + \overline{L}\|\mathcal{W} - I\|\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\| \\ + \gamma\overline{L}\|\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\| + \gamma\frac{L}{\sqrt{\epsilon}} \left\| \mathrm{m}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1} \right\| \\ + \gamma\frac{\overline{L}\sqrt{N}}{\sqrt{\epsilon}} \left\| \bar{\mathrm{m}}^{k+1} \right\| + \sqrt{N}\eta^k. \end{aligned} \tag{2.167}$$

The proof follows by using Lemma 2.6, 2.7 and 2.8 to bound $\left\| \bar{\mathrm{m}}^{k+1} \right\|$, $\left\| \mathrm{m}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1} \right\|$, and $\left\| \mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1} \right\|$. ■

Lemma 2.10 (Consensus error)**.** *Let Assumptions 2.9, and 2.11 hold. Then, for all $k \geq 1$, it holds*

$$\begin{aligned} \|\mathrm{x}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\| \leq \left(\Lambda + \gamma\frac{(1-\beta_1)\overline{L}}{\sqrt{\epsilon}} \right) \|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\| + \gamma\frac{\beta_1\sqrt{N}}{\sqrt{\epsilon}} \left\| \bar{\mathrm{m}}^k \right\| \\ + \gamma\frac{\beta_1}{\sqrt{\epsilon}} \left\| \mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k \right\| + \gamma\frac{(1-\beta_1)}{\sqrt{\epsilon}} \left\| \mathrm{s}^k - \bar{\mathrm{s}}^k \right\| \\ + \gamma\frac{(1-\beta_1)\beta_1\overline{L}\sqrt{N}}{\sqrt{\epsilon}} \left\| \bar{\mathrm{x}}^k - x_\star^k \right\|. \end{aligned}$$

Proof.

By combining (2.155d) and (2.156d), we have

$$\begin{aligned} \|\mathrm{x}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\| &= \|\mathcal{W}\mathrm{x}^k - \gamma\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k + \gamma\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\| \\ &\overset{(a)}{\leq} \|\mathcal{W}\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\| + \gamma\|\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\| \\ &\overset{(b)}{\leq} \Lambda\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\| + \gamma\|\mathrm{d}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\|, \end{aligned}$$

where in (a) we apply the triangle inequality and (b) follows by $\left\|\mathcal{W} - \frac{\mathbf{1}_{N,n}\mathbf{1}_{N,n}^\top}{N}\right\| \leq \Lambda$. The proof follows by Lemma 2.8. ■

Lemma 2.11 (Tracking momentum difference magnitude)**.** *Let Assumptions 2.9, 2.10, and 2.11 hold. Then, for all $k \geq 0$, it holds*

$$\begin{aligned}\|\bar{\mathrm{s}}^{k+1} - \bar{\mathrm{m}}^{k+1}\| \leq \beta_1\|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| + \gamma\frac{\beta_1\overline{L}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{m}}^k\right\| + \gamma\frac{2\beta_1\overline{L}}{\sqrt{\epsilon}\sqrt{N}}\|\mathrm{m}^k - \mathbf{1}_{N,n}\bar{\mathrm{m}}^k\| \\ + \left(\Lambda\frac{\overline{L}}{\sqrt{N}} + \frac{\overline{L}}{\sqrt{N}} + \gamma\frac{(1-\beta_1)L^2}{\sqrt{\epsilon}\sqrt{N}}\right)\left\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\right\| \\ + \gamma\frac{\overline{L}}{\sqrt{\epsilon}\sqrt{N}}\left\|\mathrm{s}^k - \mathbf{1}_{N,n}\bar{\mathrm{s}}^k\right\| + \gamma\frac{(1-\beta_1)\overline{L}^2}{\sqrt{\epsilon}}\left\|\bar{\mathrm{x}}^k - x_\star^k\right\| + \frac{1}{\sqrt{N}}\eta^k.\end{aligned}$$

Proof. From the updates of $\bar{\mathrm{s}}^{k+1}$ and $\bar{\mathrm{m}}^{k+1}$ (cf. (2.156e), (2.156a)), we get

$$\begin{aligned}\|\bar{\mathrm{s}}^{k+1} - \bar{\mathrm{m}}^{k+1}\| &= \left\|\bar{\mathrm{s}}^k + \frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\mathrm{x}_i^{k+1}) - \frac{1}{N}\sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) - \beta_1\bar{\mathrm{m}}^k - (1-\beta_1)\bar{\mathrm{s}}^k\right\| \\ &\overset{(a)}{\leq} \beta_1\|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| + \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\mathrm{x}_i^{k+1}) - \frac{1}{N}\sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k)\right\|,\end{aligned}$$

where (a) uses the triangle inequality. By adding and subtracting within the second norm $\frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\bar{\mathrm{x}}^{k+1})$ and $\frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\mathrm{x}_i^k)$, we use the triangle inequality to obtain

$$\begin{aligned}\|\bar{\mathrm{s}}^{k+1} - \bar{\mathrm{m}}^{k+1}\| \leq \beta_1\|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| + \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\mathrm{x}_i^{k+1}) - \frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\bar{\mathrm{x}}^{k+1})\right\| \\ + \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\mathrm{x}_i^k) - \frac{1}{N}\sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k)\right\| \\ + \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i^{k+1}(\bar{\mathrm{x}}^{k+1}) - \frac{1}{N}\sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k)\right\| \\ \overset{(a)}{\leq} \beta_1\|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| + \frac{\overline{L}}{\sqrt{N}}\|\mathrm{x}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\| + \frac{1}{\sqrt{N}}\eta^k \\ + \frac{\overline{L}}{\sqrt{N}}\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\|, \qquad (2.168)\end{aligned}$$

where in (a) we use the Lipschitz continuity of the gradients of the cost functions (cf. Assumptions 2.9) for the second and the third norm, and we use η^k (cf. (2.158)). Now,

we replace $\bar{\mathrm{x}}^{k+1}$ with its update (2.156d) within the last term of (2.168) obtaining

$$\begin{aligned}\|\bar{\mathrm{s}}^{k+1}-\bar{\mathrm{m}}^{k+1}\| &\leq \beta_1\|\bar{\mathrm{s}}^k-\bar{\mathrm{m}}^k\|+\frac{L}{\sqrt{N}}\|\mathrm{x}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\|+\frac{\overline{L}}{\sqrt{N}}\eta^k\\ &\quad+\frac{\overline{L}}{\sqrt{N}}\|\mathbf{1}_{N,n}\bar{\mathrm{x}}^k-\gamma\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}-\mathrm{x}^k\|\\ &\overset{(a)}{\leq}\beta_1\|\bar{\mathrm{s}}^k-\bar{\mathrm{m}}^k\|+\frac{\overline{L}}{\sqrt{N}}\|\mathrm{x}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\|+\frac{\overline{L}}{\sqrt{N}}\eta^k+\frac{\overline{L}}{\sqrt{N}}\|\mathrm{x}^k-\mathbf{1}_{N,n}\bar{\mathrm{x}}^k\|\\ &\quad+\gamma\frac{\overline{L}}{\sqrt{\epsilon}\sqrt{N}}\left\|\mathrm{m}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}\right\|+\gamma\frac{\overline{L}}{\sqrt{\epsilon}}\left\|\bar{\mathrm{m}}^{k+1}\right\|,\end{aligned}\tag{2.169}$$

where in (a) we use (2.165) to bound $\left\|\mathbf{1}_{N,n}\bar{\mathrm{d}}^{k+1}\right\|$. The proof follows by using Lemma 2.10, 2.6, and 2.7 to bound $\left\|\mathrm{x}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{x}}^{k+1}\right\|$, $\left\|\bar{\mathrm{m}}^{k+1}\right\|$, and $\left\|\mathrm{m}^{k+1}-\mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}\right\|$, respectively. ■

Lemma 2.12 (Solution error)**.** *Let Assumptions 2.9, 2.10, and 2.11 hold. Then, for all $k\geq 0$, it holds*

$$\begin{aligned}\|\bar{\mathrm{x}}^{k+1}-x_\star^{k+1}\| &\leq (1-\gamma\delta)\|\bar{\mathrm{x}}^k-x_\star^k\|+\gamma\frac{\beta_1}{\sqrt{\epsilon}}\|\bar{\mathrm{s}}^k-\bar{\mathrm{m}}^k\|+\gamma\frac{\overline{L}}{\sqrt{\epsilon}\sqrt{N}}\|\mathrm{x}^k-\mathbf{1}_{N,n}\bar{\mathrm{x}}^k\|\\ &\quad+\gamma\frac{\beta_1}{\sqrt{\epsilon}\sqrt{N}}\left\|\mathrm{m}^k-\mathbf{1}_{N,n}\bar{\mathrm{m}}^k\right\|+\gamma\frac{(1-\beta_1)}{\sqrt{\epsilon}\sqrt{N}}\left\|\mathrm{s}^k-\mathbf{1}_{N,n}\bar{\mathrm{s}}^k\right\|+\zeta^k,\end{aligned}$$

where ζ^k is defined in (2.158) *and $\delta:=\min\left\{\frac{\mu}{\sqrt{\epsilon+G}},\frac{\overline{L}}{\sqrt{\epsilon}}\right\}$.*

Proof. By using (2.156d), one has

$$\left\|\bar{\mathrm{x}}^{k+1}-x_\star^{k+1}\right\|=\left\|\bar{\mathrm{x}}^k-\gamma\bar{\mathrm{d}}^{k+1}-x_\star^{k+1}\right\|\overset{(a)}{\leq}\left\|\bar{\mathrm{x}}^k-\gamma\bar{\mathrm{d}}^{k+1}-x_\star^k\right\|+\zeta^k,$$

where in (a) we add and subtract within the norm $x_\star^k$, use the triangle inequality, and use ζ^k (cf. (2.158)). Now, we add and subtract within the norm $\gamma\frac{\mathbf{1}_{N,n}^\top(\mathrm{V}^{k+1}+\epsilon I)^{-1/2}\mathbf{1}_{N,n}}{N^2}\nabla f^k(\bar{\mathrm{x}}^k)$ and we use the triangle inequality to write

$$\begin{aligned}\|\bar{\mathrm{x}}^{k+1}-x_\star^{k+1}\| &\leq\left\|\bar{\mathrm{x}}^k-\gamma\frac{\mathbf{1}_{N,n}^\top(\mathrm{V}^{k+1}+\epsilon I)^{-1/2}\mathbf{1}_{N,n}}{N^2}\nabla f^k(\bar{\mathrm{x}}^k)-x_\star^k\right\|\\ &\quad+\gamma\left\|\frac{\mathbf{1}_{N,n}^\top(\mathrm{V}^{k+1}+\epsilon I)^{-1/2}\mathbf{1}_{N,n}}{N^2}\nabla f^k(\bar{\mathrm{x}}^k)-\bar{\mathrm{d}}^{k+1}\right\|+\zeta^k.\end{aligned}\tag{2.170}$$

Consider the second term of (2.170) and use (2.156c) to write

$$\gamma\left\|\frac{\mathbf{1}_{N,n}^\top(\mathrm{V}^{k+1}+\epsilon I)^{-1/2}\mathbf{1}_{N,n}}{N}\frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N}-\bar{\mathrm{d}}^{k+1}\right\|$$

$$
\begin{aligned}
&= \gamma \left\| \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1} + \epsilon I)^{-1/2} \mathbf{1}_{N,n}}{N} \frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N} - \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1} + \epsilon I)^{-1/2}}{N} \mathrm{m}^{k+1} \right\| \\
&\overset{(a)}{\leq} \gamma \left\| \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1} + \epsilon I)^{-1/2} \mathbf{1}_{N,n}}{N} \left(\frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N} - \bar{\mathrm{m}}^{k+1} \right) \right\| \\
&\quad + \gamma \left\| \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1} + \epsilon I)^{-1/2}}{N} (\mathrm{m}^{k+1} - \mathbf{1}_{N,n} \bar{\mathrm{m}}^{k+1}) \right\| \\
&\overset{(b)}{\leq} \frac{\gamma}{\sqrt{\epsilon}} \left\| \frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N} - \bar{\mathrm{m}}^{k+1} \right\| + \frac{\gamma}{\sqrt{\epsilon}\sqrt{N}} \left\| \mathrm{m}^{k+1} - \mathbf{1}_{N,n} \bar{\mathrm{m}}^{k+1} \right\|, \qquad (2.171)
\end{aligned}
$$

where in (a) we add and subtract within the norm the term $\frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1}+\epsilon I)^{-1/2} \mathbf{1}_{N,n}}{N} \bar{\mathrm{m}}^{k+1}$ and we apply the triangle inequality, in (b) we apply the Cauchy-Schwarz inequality combined with the bounds $\left\| \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1}+\epsilon I)^{-1/2} \mathbf{1}_{N,n}}{N} \right\| \leq \frac{1}{\sqrt{\epsilon}}$ and $\left\| \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1}+\epsilon I)^{-1/2}}{N} \right\| \leq \frac{1}{\sqrt{\epsilon}\sqrt{N}}$. Now, we add and subtract the term $\frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k)$ and then we use the triangle inequality to rewrite the first term of the second member of (2.171) as

$$
\begin{aligned}
&\gamma \frac{1}{\sqrt{\epsilon}} \left\| \frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N} - \bar{\mathrm{m}}^{k+1} \right\| \\
&= \gamma \frac{1}{\sqrt{\epsilon}} \left\| \frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) - \bar{\mathrm{m}}^{k+1} \right\| + \gamma \frac{1}{\sqrt{\epsilon}} \left\| \frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N} - \frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) \right\| \\
&\overset{(a)}{=} \gamma \frac{1}{\sqrt{\epsilon}} \left\| \frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) - \beta_1 \bar{\mathrm{m}}^k - (1-\beta_1) \bar{\mathrm{s}}^k \right\| + \gamma \frac{1}{\sqrt{\epsilon}} \left\| \frac{\nabla f^k(\bar{\mathrm{x}}^k)}{N} - \frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) \right\| \\
&\overset{(b)}{\leq} \gamma \frac{\beta_1}{\sqrt{\epsilon}} \|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| + \gamma \frac{\overline{L}}{\sqrt{\epsilon}\sqrt{N}} \|\mathrm{x}^k - \mathbf{1}_{N,n} \bar{\mathrm{x}}^k\|, \qquad (2.172)
\end{aligned}
$$

where in (a) we use (2.156a), (b) uses the relation $\bar{\mathrm{s}}^k = \frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k)$, and the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.9) combined with $\frac{1}{N} \sum_{i=1}^N \nabla f_i^k(\mathrm{x}_i^k) = \sum_{i=1}^N \nabla f_i^k(\bar{\mathrm{x}}^k)$. Next, in order to bound the right-hand side of (2.170), first notice that $\frac{1}{\sqrt{G+\epsilon}} < \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1}+\epsilon I)^{-1/2} \mathbf{1}_{N,n}}{N} < \frac{1}{\sqrt{\epsilon}}$. Moreover, being f^k μ-strongly convex for all $k \geq 0$ (cf. Assumption 2.10) and having $\overline{L}$-Lipschitz continuous gradients (cf. Assumption 2.9), we apply Lemma B.2 (in Appendix B) to write

$$
\left\| \bar{\mathrm{x}}^k - \gamma \frac{\mathbf{1}_{N,n}^\top (\mathrm{V}^{k+1} + \epsilon I)^{-1/2} \mathbf{1}_{N,n}}{N} \nabla f^k(\bar{\mathrm{x}}^k) - x_\star^k \right\| \leq \phi \left\| \bar{\mathrm{x}}^k - x_\star^k \right\|, \qquad (2.173)
$$

where $\phi := \max\left\{ \left| 1 - \frac{\gamma}{\sqrt{\epsilon+G}} \mu \right|, \left| 1 - \frac{\gamma}{\sqrt{\epsilon}} \overline{L} \right| \right\}$. If we take $\gamma < \min\left\{ \frac{\sqrt{\epsilon+G}}{\mu}, \frac{\sqrt{\epsilon}}{\overline{L}} \right\}$, then it holds $\phi = 1 - \gamma\delta$, where δ is defined in the statement of Theorem 2.7. By combining the

latter with (2.172) and (2.173), it is possible to upper bound (2.170) as

$$\|\bar{\mathrm{x}}^{k+1} - x_\star^{k+1}\| \leq (1-\gamma\delta)\|\bar{\mathrm{x}}^k - x_\star^k\| + \gamma\frac{\beta_1}{\sqrt{\epsilon}}\|\bar{\mathrm{s}}^k - \bar{\mathrm{m}}^k\| + \frac{\gamma}{\sqrt{\epsilon}\sqrt{N}}\left\|\mathrm{m}^{k+1} - \mathbf{1}_{N,n}\bar{\mathrm{m}}^{k+1}\right\| + \frac{\gamma\overline{L}}{\sqrt{\epsilon}\sqrt{N}}\left\|\mathrm{x}^k - \mathbf{1}_{N,n}\bar{\mathrm{x}}^k\right\| + \zeta^k. \tag{2.174}$$

The proof follows by invoking Lemma 2.7 to bound $\left\|\mathrm{m}^{k+1} - \bar{\mathrm{m}}^{k+1}\right\|$ within (2.174). ■

Dynamic Regret

Once the necessary preparatory results have been provided, the main result of this paper is stated as follows.

Theorem 2.7. *Consider GTAdam as given in Algorithm 4. Let Assumptions 2.9, 2.10, and 2.11 hold. Then, for a sufficiently small step-size $\gamma > 0$, there exists a constant $0 < \tilde{\rho} < 1$, such that*

$$R_T \leq \frac{\overline{L}\lambda^2}{2}\left(\frac{\left\|\mathrm{y}^0\right\|^2}{1-\tilde{\rho}^2} + 2\left\|\mathrm{y}^0\right\| S_T + Q_T\right), \tag{2.175}$$

where R_T is defined in (2.154)*, the constant λ is defined in the proof (cf.* (2.185)*) and*

$$S_T := \sum_{k=1}^{T}\sum_{\tau=0}^{k-1}\tilde{\rho}^{k+q}\left(\frac{N+1}{\sqrt{N}}\left\|\eta^{k-\tau-1}\right\| + \left\|\zeta^{k-\tau-1}\right\|\right) \tag{2.176a}$$

$$Q_T := \sum_{k=1}^{T}\left(\sum_{\tau=0}^{k-1}\tilde{\rho}^{k}\left(\frac{N+1}{\sqrt{N}}\left\|\eta^{k-\tau-1}\right\| + \left\|\zeta^{k-\tau-1}\right\|\right)\right)^2, \tag{2.176b}$$

where η^k, ζ^k are defined in (2.158) *and we assume that are finite. Moreover, it holds*

$$\lim_{T\to\infty}\sum_{i=1}^{N}\left\|\mathrm{x}_i^T - \bar{\mathrm{x}}^T\right\|^2 \leq \frac{\lambda^2}{(1-\tilde{\rho})^2}\max_k\left\{\frac{N^2+1}{N}\eta^k + \zeta^k\right\}. \tag{2.177}$$

Proof. By recalling the definition of y^k given in (2.157) and combining Lemma 2.6, 2.7, 2.8, 2.9, 2.10, 2.11, 2.12, it is possible to write

$$\mathrm{y}^{k+1} \leq A(\gamma)\mathrm{y}^k + u^k, \tag{2.178}$$

where $u^k := \text{col}\left(0, \frac{1}{\sqrt{N}}\eta^k, 0, \sqrt{N}\eta^k, 0, \zeta^k\right)$. The matrix $A(\gamma)$ can be decomposed in $A(\gamma) := A_0 + \gamma E$, with

$$A_0 := \begin{bmatrix} \beta_1 & 0 & 0 & 0 & \beta_1 c_1 & (1-\beta_1)\overline{L} \\ 0 & \beta_1 & 0 & 0 & \Lambda c_1 + c_1 & 0 \\ 0 & 0 & \beta_1 & 1-\beta_1 & 0 & 0 \\ 0 & 0 & 0 & \Lambda & c_2 & 0 \\ 0 & 0 & 0 & 0 & \Lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$E := \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{\beta_1\overline{L}}{\sqrt{\epsilon}} & 0 & \frac{2\beta_1 c_1}{\sqrt{\epsilon}} & \frac{c_1}{\sqrt{\epsilon}} & \frac{(1-\beta_1)c_1\overline{L}}{\sqrt{\epsilon}} & \frac{(1-\beta_1)\overline{L}^2}{\sqrt{\epsilon}} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{2\beta_1\overline{L}\sqrt{N}}{\sqrt{\epsilon}} & 0 & \frac{2\beta_1\overline{L}}{\sqrt{\epsilon}} & \frac{2(1-\beta_1)\overline{L}}{\sqrt{\epsilon}} & c_3 & c_4 \\ \frac{\beta_1\sqrt{N}}{\sqrt{\epsilon}} & 0 & \frac{\beta_1}{\sqrt{\epsilon}} & \frac{1-\beta_1}{\sqrt{\epsilon}} & 0 & c_5 \\ 0 & \frac{\beta_1}{\sqrt{\epsilon}} & \frac{\beta_1}{\sqrt{\epsilon}\sqrt{N}} & 0 & \frac{(1-\beta_1)}{\sqrt{\epsilon}\sqrt{N}} & -\delta \end{bmatrix},$$

where we used the following shorthands

$$c_1 := \frac{\overline{L}}{\sqrt{N}}, \quad c_2 := \overline{L}\|\mathcal{W} - I\|, \quad c_3 := \frac{2(1-\beta_1)\beta_1\overline{L}^2}{\sqrt{\epsilon}},$$
$$c_4 := \frac{(1-\beta_1)(1+\beta_1)\overline{L}^2\sqrt{N}}{\sqrt{\epsilon}}, \quad c_5 := \frac{(1-\beta_1)\beta_1\overline{L}\sqrt{N}}{\sqrt{\epsilon}}.$$

Being A_0 triangular, it is easy to see that its spectral radius is 1 since both β_1 and Λ are in $(0, 1)$. We want to study how the perturbation matrix γE affects the simple eigenvalue 1 of A_0. Hence, we denote by $\chi(\gamma)$ such eigenvalue of $A(\gamma)$ as a function of γ. Call w and v respectively the left and right eigenvectors of A_0 associated to the eigenvalue 1, then $w = \text{col}\,(0, 0, 0, 0, 0, 1)$ and $v = \text{col}\left(\overline{L}, 0, 0, 0, 0, 1\right)$. Since the eigenvalue 1 is simple, from Theorem B.1 (in Appendix B) it holds

$$\left.\frac{d\chi(\gamma)}{d\gamma}\right|_{\gamma=0} = \frac{w^\top E v}{w^\top v} = -\delta < 0.$$

Then, by continuity of eigenvalues with respect to the matrix entries, $\chi(\gamma)$ is strictly less than 1 for sufficiently small $\gamma > 0$. Then, it is always possible to choose $\gamma > 0$ so as the remaining eigenvalues stay in the unit circle. Therefore, the spectral radius is $\rho(A(\gamma)) < 1$. Moreover, since $A(\gamma)$ and u^k have only non-negative entries, one can

use (2.178) to write

$$\mathrm{y}^k \leq A(\gamma)^k \mathrm{y}^0 + \sum_{\tau=0}^{k-1} A(\gamma)^{k-1-\tau} u^\tau. \tag{2.179}$$

From [85, Lemma 5.6.10], we have that for any $\iota > 0$, there exists a matrix norm, say $|||\cdot|||_\iota$, such that

$$|||A(\gamma)|||_\iota \leq \rho(A(\gamma)) + \iota. \tag{2.180}$$

Let us pick $\iota \in (0, 1 - \rho(A(\gamma)))$ and define $\tilde{\rho} := \rho(A(\gamma)) + \iota$. Then, in light of (2.180) it holds $|||A(\gamma)|||_\iota \leq \tilde{\rho} < 1$. Moreover, by applying [85, Theorem 5.7.13], there exists a vector norm $\|\cdot\|_\iota$ such that $\|Mv\|_\iota \leq |||M|||_\iota \|v\|_\iota$ for any matrix $M \in \mathbb{R}^{6\times 6}$ and $v \in \mathbb{R}^6$. Hence, we can manipulate (2.179) taking the norm and using the triangle inequality to write

$$\left\|\mathrm{y}^k\right\|_\iota \leq \left\|A(\gamma)^k \mathrm{y}^0\right\|_\iota + \left\|\sum_{\tau=0}^{k-1} A(\gamma)^{k-1-\tau} u^\tau\right\|_\iota \leq \tilde{\rho}^k \left\|\mathrm{y}^0\right\|_\iota + \sum_{\tau=0}^{k-1} \tilde{\rho}^k \left\|u^{k-1-\tau}\right\|_\iota, \tag{2.181}$$

which shows that first term decreases linearly with rate $\tilde{\rho} < 1$ while the second one is bounded. By using the Lipschitz continuity of the gradients of f^k (cf. Assumption 2.9), we have

$$f^k(\bar{\mathrm{x}}^k) - f^k(x_\star^k) \leq \frac{\overline{L}}{2}\|\bar{\mathrm{x}}^k - x_\star^k\|^2 \overset{(a)}{\leq} \frac{\overline{L}}{2}\left\|\mathrm{y}^k\right\|^2, \tag{2.182}$$

where in (a) we use the fact that $\left\|\bar{\mathrm{x}}^k - x_\star^k\right\|$ represents a component of y^k leading to the trivial bound $\left\|\bar{\mathrm{x}}^k - x_\star^k\right\| \leq \left\|\mathrm{y}^k\right\|$. Recalling that all norms are equivalent on finite-dimensional vector spaces, there always exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that

$$\|\cdot\| \leq \lambda_1 \|\cdot\|_\iota \tag{2.183a}$$

$$\|\cdot\|_\iota \leq \lambda_2 \|\cdot\|. \tag{2.183b}$$

Thus, by applying (2.183a), we bound (2.182) as

$$f^k(\bar{\mathrm{x}}^k) - f^k(x_\star^k) \leq \frac{\overline{L}\lambda_1}{2}\left\|\mathrm{y}^k\right\|_\iota^2,$$

which, combined with the definition of R_T (cf. (2.154)) and the result (2.181), leads to

$$R_T \leq \frac{\overline{L}\lambda_1^2}{2}\left(\sum_{k=1}^{T} \tilde{\rho}^{2k}\left\|\mathrm{y}^0\right\|_\iota^2 + 2\left\|\mathrm{y}^0\right\|_\iota \sum_{k=1}^{T}\sum_{\tau=0}^{k-1} \tilde{\rho}^{k+\tau}\left\|u^{k-1-\tau}\right\|_\iota + \sum_{k=1}^{T}\left(\sum_{\tau=0}^{k-1}\tilde{\rho}^\tau\left\|u^{k-1-\tau}\right\|_\iota\right)\right)^2$$

$$\overset{(a)}{\leq} \frac{\overline{L}\lambda_1^2\lambda_2^2}{2}\left(\frac{\left\|\mathrm{y}^0\right\|^2}{1-\tilde{\rho}^2}+2\left\|\mathrm{y}^0\right\|\sum_{k=1}^{T}\sum_{\tau=0}^{k-1}\tilde{\rho}^{k+\tau}\left\|u^{k-1-\tau}\right\|+\sum_{k=1}^{T}\left(\sum_{\tau=0}^{k-1}\tilde{\rho}^{\tau}\left\|u^{k-1-\tau}\right\|\right)^2\right), \tag{2.184}$$

where in (a) we use the geometric series property and the relation (2.183b). The proof follows by using the definitions of U_T and Q_T (cf. (2.176)) and by setting

$$\lambda := \lambda_1\lambda_2. \tag{2.185}$$

Finally, in order to prove (2.177), we notice that $\sum_{i=1}^{N}\left\|\mathrm{x}_i^T-\bar{\mathrm{x}}^T\right\|^2 \leq \left\|\mathrm{y}^T\right\|^2 \leq \lambda_1^2\left\|\mathrm{y}^T\right\|_\iota^2$, in which we apply (2.183a). By applying the bound (2.181) for $k=T$, we get

$$\left\|\mathrm{y}^T\right\|_\iota \leq \tilde{\rho}^T\left\|\mathrm{y}^0\right\|_\iota + \sum_{\tau=0}^{T-1}\tilde{\rho}^{\tau}\left\|u^{T-\tau-1}\right\|_\iota.$$

The first term of the latter inequality vanishes as $T\to\infty$, while the second one can be bounded by relying on geometric series property and $\max_k\{\left\|u^k\right\|^2\}$. By exploiting these arguments, we can write

$$\lim_{T\to\infty}\sum_{i=1}^{N}\left\|\mathrm{x}_i^T-\bar{\mathrm{x}}^T\right\|^2 \leq \frac{\lambda_1^2}{(1-\tilde{\rho})^2}\max_k\left\{\left\|u^k\right\|_\iota^2\right\} \overset{(a)}{\leq} \frac{\lambda^2}{(1-\tilde{\rho})^2}\max_k\left\{\left\|u^k\right\|^2\right\}. \tag{2.186}$$

where in (a) we apply (2.183b) and the definition (2.185) of λ. The result (2.177) follows by noting that

$$\max_k\left\{\left\|u^k\right\|^2\right\} = \max_k\left\{\frac{N^2+1}{N}\eta^k+\zeta^k\right\}.$$

■

There is evidence in the literature, see, e.g., [48, 105, 107, 132, 140, 169], that the bound on the dynamic regret cannot be sublinear with respect to T. As stated, e.g., in [105], when the objective functions are strongly convex and have bounded gradients, the bound on dynamic regret is $O(1+\eta^k)$. Our work does not assume gradient boundedness and, thus, our bound has additional terms due to variations over time of the gradients. Specifically, Theorem 2.7 shows that R_T is upper bounded by a constant depending on the initial conditions and by other two terms. The latters involve S_T and Q_T, which capture the time-varying nature of the problem itself. Indeed, suppose that the problem varies linearly, i.e., there exists $C>0$ so that $\eta^k,\zeta^k\leq C$ for all $k\geq 0$. Then,

being $\tilde{\rho} \in (0,1)$, we can exploit the geometric series properties to write the following

$$S_T \leq \frac{(N+\sqrt{N}+1)(\tilde{\rho}-\tilde{\rho}^{k+1})(1-\tilde{\rho})C}{\sqrt{N}(1-\tilde{\rho})^2}, \quad Q_T \leq \frac{(N+\sqrt{N}+1)^2(1-\tilde{\rho}^T)^2C^2T}{N(1-\tilde{\rho})^2}.$$

In this case, (2.175) ensures that the average regret R_T/T asymptotically approaches a constant when $T \to \infty$, specifically

$$\lim_{T\to\infty} \frac{R_T}{T} \leq \frac{\overline{L}\lambda^2(N+\sqrt{N}+1)^2C^2}{2N(1-\tilde{\rho})^2}.$$

The key point of the proof consists in showing that the error vector y^k (see (2.157)) evolves according to a linear system with state matrix $A(\gamma)$ (whose entries depend on the problem parameters such, e.g., the strong convexity function or the network connectivity) which is perturbed by an input u^k related to the variations of the problem over time (see (2.178)). Notice that the parameter $\tilde{\rho}$ is related to the spectral radius of $A(\gamma)$ and, thus, depends also on the network topology.

Agent Regret

We may also consider a regret for each agent i defined as $R_{T,i} := \sum_{k=1}^T f^k(\mathrm{x}_i^k) - \sum_{k=1}^T f^k(x_\star^k)$.

Corollary 2.1. *Under the same assumptions of Theorem 2.7, for all $i \in \{1,\dots,N\}$, it holds*

$$R_{T,i} \leq 2\overline{L}\lambda^2 \left(\frac{\left\|\mathrm{y}^0\right\|^2}{1-\tilde{\rho}^2} + 2\left\|\mathrm{y}^0\right\| S_T + Q_T \right),$$

where λ, $\tilde{\rho}$, S_T, and Q_T are defined as in Theorem 2.7.

Proof. We add and subtract $f^k(\bar{\mathrm{x}}^k)$ to $f^k(\mathrm{x}_i^k) - f^k(x_\star^k)$, obtaining

$$\begin{aligned} f^k(\mathrm{x}_i^k) - f^k(x_\star^k) &= f^k(\mathrm{x}_i^k) - f^k(\bar{\mathrm{x}}^k) + f^k(\bar{\mathrm{x}}^k) - f^k(x_\star^k) \\ &\overset{(a)}{\leq} f^k(\mathrm{x}_i^k) - f^k(\bar{\mathrm{x}}^k) + \frac{\overline{L}}{2}\left\|\bar{\mathrm{x}}^k - x_\star^k\right\|^2 \\ &\overset{(b)}{\leq} \nabla f^k(\bar{\mathrm{x}}^k)^\top(\mathrm{x}_i^k - \bar{\mathrm{x}}^k) + \frac{\overline{L}}{2}\left\|\mathrm{x}_i^k - \bar{\mathrm{x}}^k\right\|^2 + \frac{\overline{L}}{2}\left\|\bar{\mathrm{x}}^k - x_\star^k\right\|^2, \end{aligned} \tag{2.187}$$

where in (a) we apply (2.182) and in (b) we use the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.9). Being $\nabla f^k(x_\star^k) = 0$, we rewrite (2.187) as

$$\begin{aligned} f^k(\mathrm{x}_i^k) - f^k(x_\star^k) \leq{}& (\nabla f^k(\bar{\mathrm{x}}^k) - \nabla f^k(x_\star^k))^\top(\mathrm{x}_i^k - \bar{\mathrm{x}}^k) \\ &+ \frac{\overline{L}}{2}\left\|\mathrm{x}_i^k - \bar{\mathrm{x}}^k\right\|^2 + \frac{\overline{L}}{2}\left\|\bar{\mathrm{x}}^k - x_\star^k\right\|^2 \end{aligned}$$

$$\overset{(a)}{\leq} \overline{L} \left\| \bar{\mathrm{x}}^k - x_\star^k \right\| \left\| \mathrm{x}_i^k - \bar{\mathrm{x}}^k \right\| + \frac{\overline{L}}{2} \left\| \mathrm{x}_i^k - \bar{\mathrm{x}}^k \right\|^2 + \frac{\overline{L}}{2} \left\| \bar{\mathrm{x}}^k - x_\star^k \right\|^2, \quad (2.188)$$

where in (a) we use the Cauchy-Schwarz inequality and the Lipschitz continuity of the gradients of the cost functions (cf. Assumption 2.9). Now, we notice that both $\left\| \bar{\mathrm{x}}^k - x_\star^k \right\|$ and $\left\| \mathrm{x}_i^k - \bar{\mathrm{x}}^k \right\|$ represent a component of the vector y^k defined in (2.157), and thus, can be both upper bounded by $\left\| \mathrm{y}^k \right\|$. Hence, the inequality (2.188) can be elaborated as

$$f^k(\mathrm{x}_i^k) - f^k(x_\star^k) \leq 2\overline{L} \left\| \mathrm{y}^k \right\|^2. \quad (2.189)$$

By summing over k the inequality in (2.189), we bound $R_{T,i}$ as

$$R_{T,i} \leq 2\overline{L} \sum_{k=1}^{T} \left\| \mathrm{y}^k \right\|^2 \overset{(a)}{\leq} 2\overline{L}\lambda_1^2 \sum_{k=1}^{T} \left\| \mathrm{y}^k \right\|_\iota^2, \quad (2.190)$$

where in (a) we apply (2.183a). As done above to prove (2.175), the proof follows by combining (2.190), (2.181), and (2.183b). ■

Static setup

We provide an additional corollary of Theorem 2.7 asserting theoretical guarantees in a static scenario. Specifically, for this special case the GTAdam distributed algorithm converges to the optimal solution with a linear rate.

Corollary 2.2 (Static setup)**.** *Under the same assumptions of Theorem 2.7, if additionally holds $f^k = f$ for all $k \geq 0$, then, for a sufficiently small step-size $\gamma > 0$, there exists a constant $0 < \tilde{\rho} < 1$ such that*

$$f(\bar{\mathrm{x}}^k) - f(x_\star^k) \leq \tilde{\rho}^{2k} \frac{\overline{L}\lambda^2}{2} \left\| \mathrm{y}^0 \right\|^2, \quad (2.191)$$

where the constant λ is defined in (2.185)*.*

Proof. Using the same arguments of Theorem 2.7 we start from (2.181). Differently from the dynamic case, in the static setup we have $\nabla f_i^k(x) = \nabla f_i(x)$ for all k and i, leading to $x_\star^k = x_\star$ for all k. Thus, we can combine (2.181) with $u^k \equiv 0$, the Lipschitz continuity of the gradient of the cost function (cf. Assumption 2.9) and (2.183a), to write $f(\bar{\mathrm{x}}^k) - f(x_\star^k) \leq \tilde{\rho}^{2k} \frac{\overline{L}\lambda_1^2}{2} \left\| \mathrm{y}^0 \right\|_\iota^2 \leq \tilde{\rho}^{2k} \frac{\overline{L}\lambda_1^2\lambda_2^2}{2} \left\| \mathrm{y}^0 \right\|^2$, in which we use (2.183b). The proof follows by using the definition (2.185) of λ. ■

2.6.2 Numerical Simulations

In this section we consider three multi-agent distributed learning problems to show the effectiveness of GTAdam. The first scenario regards the computation of a linear classifier via a regularized logistic regression function for a set of points that change over time.

The second scenario involves the localization of a moving target. The third example is a stochastic optimization problem arising in a distributed image classification task. In all the examples, the parameters of GTAdam are chosen as $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Moreover, we compare GTAdam with the Gradient Tracking distributed algorithm (GT) (cf. (2.5) in Section 2.2), the distributed gradient descent (DGD) (see [137]), and the distributed Adam (DAdam) (see [134]) described by

$$\begin{aligned}
\mathrm{m}_i^{k+1} &= \beta_1 \mathrm{m}_i^k + (1-\beta_1)\nabla f_i^k(\mathrm{x}_i^k) \\
\mathrm{v}_i^{k+1} &= \beta_2 \mathrm{v}_i^k + (1-\beta_2)\nabla f_i^k(\mathrm{x}_i^{k+1}) \odot \nabla f_i^k(\mathrm{x}_i^{k+1}) \\
\tilde{\mathrm{v}}_i^{k+1} &= \beta_3 \tilde{\mathrm{v}}_i^k + (1-\beta_3)\max\{\tilde{\mathrm{v}}_i^k, \mathrm{v}_i^{k+1}\} \\
\mathrm{x}_i^{k+1} &= \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{x}_j^k + \gamma^k \frac{\mathrm{m}_i^{k+1}}{\tilde{\mathrm{v}}_i^{k+1}},
\end{aligned}$$

for all $i \in \{1,\dots,N\}$. As suggested in [134], we set $\beta_1 = \beta_3 = 0.9$, $\beta_2 = 0.999$, and a diminishing step-size $\gamma^k = (\frac{\gamma}{k})^{-1/2}$, for some $\gamma > 0$.

Distributed classification via logistic regression

Here, we consider an online instance of the distributed classification problem already presented in Section 1.2.2 and addressed in Section 2.4.3. In particular, we consider a network of agents that want to cooperatively train a linear classifier for a set of (moving) points in a given feature space. At time $k \geq 0$, each agent i is equipped with $m_i \in \mathbb{N}$ points $p_{i,1}^k, \dots, p_{i,m_i}^k \in \mathbb{R}^n$ with binary labels $l_{i,k} \in \{-1, 1\}$ for all $k \in \{1,\dots,m_i\}$. The problem consists of building a linear classification model from the given points, also called training samples. In particular, we look for a separating hyperplane described by a pair $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ given by $\{p \in \mathbb{R}^d \mid w^\top p + b = 0\}$. This *online* classification problem can be posed at each time $k \geq 0$, as a minimization problem described by

$$\min_{w,b} \sum_{i=1}^{N}\sum_{q=1}^{m_i} \log\left(1 + e^{-l_{i,q}(w^\top p_{i,q}^k + b)}\right) + \frac{C}{2}\left(\|w\|^2 + b^2\right), \tag{2.192}$$

where $C > 0$ is the so-called regularization parameter. Each point $p_{i,q}^k \in \mathbb{R}^2$ moves along a circle of radius $r = 1$ according to the following law

$$p_{i,q}^k = p_{i,q}^c + r\begin{bmatrix}\cos(k/100)\\ \sin(k/100)\end{bmatrix},$$

where $p_{i,q}^c \in \mathbb{R}^2$ represents the randomly generated center of the considered circle. We consider a network of $N = 50$ agents and pick $m_i = 5$ (for all i). We performed an experimental tuning to optimize the step-sizes to enhance the convergence properties of

each algorithm. In particular, we selected $\gamma = 0.1$ for GTAdam, $\gamma = 0.05$ for Gradient Tracking, $\gamma = 0.1$ for DGD, and $\gamma = 0.1$ for DAdam. We performed Monte Carlo simulations consisting of 100 trials, in which we alternatively consider an undirected, connected Erdős-Rényi graph with connectivity parameter 0.5, and a ring graph. In Figure 2.14, we plot the average across the trials of the relative cost error, namely $\frac{f^k(\bar{\mathrm{x}}^k) - f^k(x_\star^k)}{f^k(x_\star^k)}$, with $x_\star^k$ being the minimum of f^k for all k.

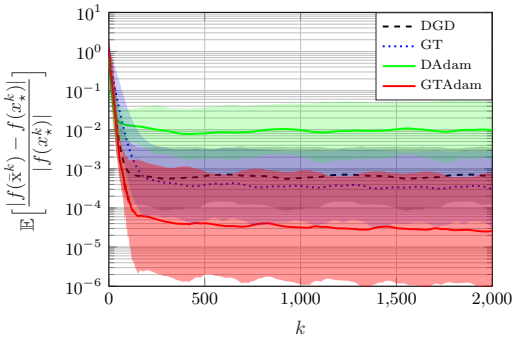


Figure 2.14: Distributed classification via logistic regression. Mean of the relative cost errors and 1-standard deviation band obtained with Monte Carlo simulations consisting of 100 trials in which each of the $N = 50$ agents is equipped with $m = 5$ points.

The plot highlights that GTAdam exhibits a faster convergence compared to the other algorithms, and achieves a smaller tracking error.

Finally, we consider a static instance of problem (2.192), i.e., with fixed objective function $f_i^k = f_i$ for all $k \geq 0$ and $i \in \{1, \dots, N\}$. We consider a network of $N = 50$ agents in a ring topology. We take $\gamma = 0.001$ for GTAdam, $\gamma = 0.01$ for GT, $\gamma = 0.1$ for DGD, and $\gamma = 0.5$ for DAdam. In Figure 2.15, we plot the error $\left\|\bar{\mathrm{x}}^k - x^\star\right\|$ achieved by the considered methods, where $x^\star \in \mathbb{R}^n$ is the (fixed) optimal solution of the problem. Figure 2.15 clearly shows the benefit of the tracking mechanism, which allows GTAdam and GT to achieve the exact problem solution. The plot also shows that GTAdam is faster than Gradient Tracking.

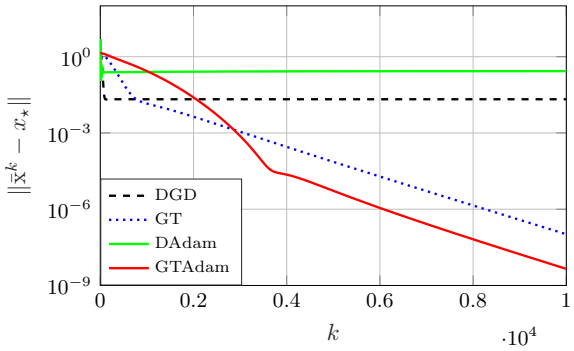


Figure 2.15: Distributed classification via logistic regression. Static setup in which each of the $N = 50$ agents is equipped with $m = 5$ points.

Distributed source localization in smart sensor networks

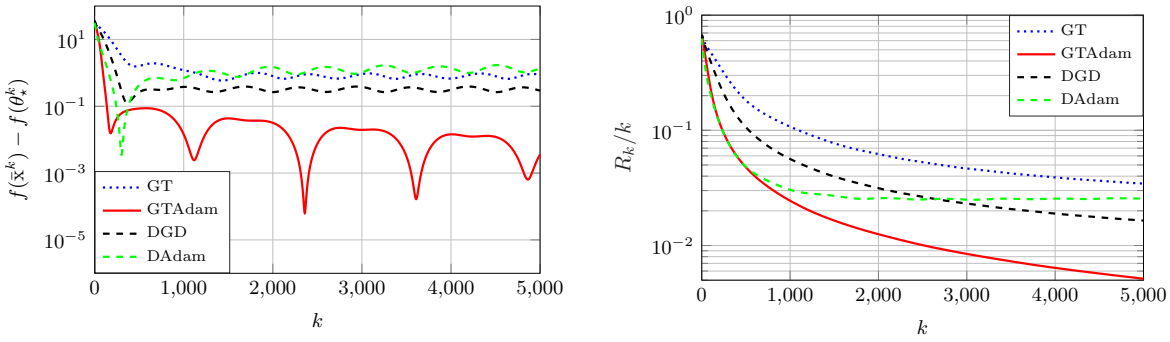
The estimation of the exact position of a source is a key task in several applications in multi-agent distributed estimation and learning. Here, we consider an online version of the static localization problem considered in [160, Section 4.2]. An acoustic source is positioned at an unknown and time-varying location $\theta^k_{\text{target}} \in \mathbb{R}^2$. A network of N sensors is capable to measure an isotropic signal related to such location and aims at cooperatively estimating θ^t_{target}. Each sensor is placed at a fixed location $c_i \in \mathbb{R}^2$ and takes, at each time instant, a noisy measurement according to an isotropic propagation model $\omega^k_i := \frac{A}{\|\theta^k_{\text{target}} - c_i\|^\omega} + \epsilon^k_i$, where $A > 0, \omega \geq 1$ describes the attenuation characteristics of the medium through which the signal propagates, and ϵ^k_i is a zero-mean Gaussian noise with variance σ^2. With this data, each node i at each time $k \geq 0$ addresses a nonlinear least-squares online problem

$$\min_{\mathrm{x}} \sum_{i=1}^{N} \left(\omega^k_i - \frac{A}{\|\mathrm{x} - c_i\|^\omega}\right)^2.$$

We consider a network of $N = 50$ agents randomly located according to a two-dimensional Gaussian distribution with zero mean and variance $a^2 I_2 = 100 I_2$. The agents want to track the location of a moving target which starts at a random location $\theta^0_{\text{target}} \in \mathbb{R}^2$ generated according to the same distribution of the agents. The target moves along a circle of radius $r = 0.5$ according to the following law

$$\theta^k_{\text{target}} = \theta_{\text{center}} + r \begin{bmatrix} \cos(k/200) \\ \sin(k/200) \end{bmatrix},$$

where $\theta_{\text{center}} \in \mathbb{R}^2$ represents the randomly generated circle center. We pick $\omega = 1$, $A = 100$ and a noise variance $\sigma^2 = 0.001$. We take $\gamma = 0.05$ for GTAdam, $\gamma = 0.02$ for GT, $\gamma = 0.05$ for DGD, and $\gamma = 0.0725$ for DAdam. The agents communicate according to a ring graph. In Figure 2.16a we compare the algorithm performance in terms of the (instantaneous) cost function evolution. Figure 2.16b shows that the best performance in terms of average dynamic regret is obtained by GTAdam. GTAdam seems to achieve a smaller error with respect to the other algorithms. We make these comparisons by using θ^k_{target} as the optimal estimate associated to the iteration k, but we note that the actual optimal solution may be slightly different since the noise ϵ^k_i affects the measurement of each agent.



(a) Relative cost error.

(b) Average dynamic regret.

Figure 2.16: Distributed source localization over a network of $N = 50$ agents.

Distributed image classification via neural networks

In this example, we consider an image classification problem in which N nodes have to cooperatively learn how to correctly classify images. We pick the Fashion-MNIST dataset [197] consisting of black-and-white 28×28-pixels images of clothes belonging to 10 different classes. Each agent i has a local dataset $\mathcal{D}_i = \{(p_{i,q}, y_{i,q})\}_{q=1}^{m_i}$ consisting of m_i images $p_{i,\ell} \in \mathbb{R}^{28\times 28}$ and their associated labels $y_{i,q} \in \{1, \dots, 10\}$. The goal of the agents is to learn the parameters $\mathrm{x}_\star$ of a function $h(p; \mathrm{x}_\star)$ so that $h(p_{i,q}; \mathrm{x}_\star)$ gives the correct label for $p_{i,q}$. The resulting optimization problem is

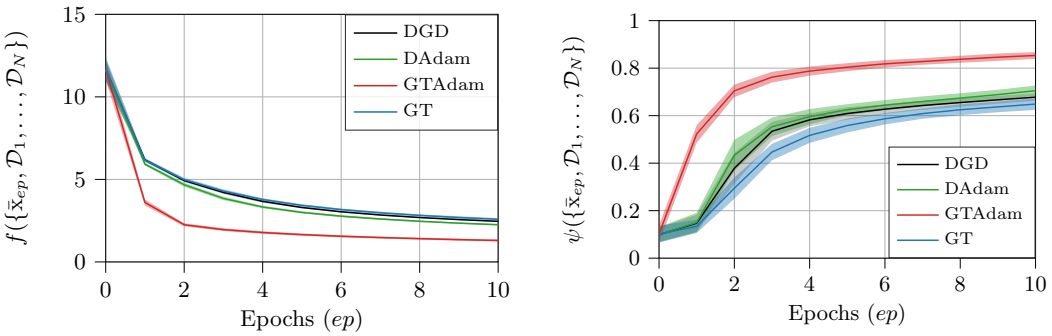
$$\min_x \sum_{i=1}^{N} \frac{1}{m_i} \sum_{q=1}^{m_i} V(y_{i,q}, h(p_{i,q}, x)) + C\|x\|^2,$$

where $V(\cdot)$ is the categorical cross-entropy loss, and $C > 0$ is a regularization parameter. The local cost function is

$$f_i(x \mid \mathcal{D}_i) := \mathbb{E}_{\mathcal{D}_i}[\ell_i(x)] = \frac{1}{m_i} \sum_{q=1}^{m_i} V(y_{i,q}, h(p_{i,q}, x)) + \frac{C}{N}\|x\|^2.$$

We represent $h(\cdot)$ by a neural network with one hidden layer (with 300 units with ReLU activation function) and an output layer with 10 units. Moreover, we pick $N = 16$ agents and associate each of them $m_i = 3750$ labeled images for all i. We performed Monte Carlo simulations consisting of 100 trials and each trial lasts 10 epochs over the local datasets. The results are reported In Figure 2.17a and Figure 2.17b in terms of the global training loss $f(\{\bar{\mathrm{x}}_{ep}, \mathcal{D}_1, \dots, \mathcal{D}_N\}) := \sum_{i=1}^{N} f_i(\bar{\mathrm{x}}_{ep} \mid \mathcal{D}_i)$, with $\bar{\mathrm{x}}_{ep} := \frac{1}{N}\sum_{i=1}^{N} \mathrm{x}_{i,ep}$, and the average training accuracy $\psi(\{\bar{\mathrm{x}}_{ep}, \mathcal{D}_1, \dots, \mathcal{D}_N\}) := \frac{1}{N}\sum_{i=1}^{N} \psi_i(\bar{\mathrm{x}}_{ep} \mid \mathcal{D}_i)$, where $\psi_i(\bar{\mathrm{x}}_{ep} \mid \mathcal{D}_i)$ is the accuracy achieved with $\bar{\mathrm{x}}_{ep}$ on the local dataset of the agent i at the end of epoch ep. We take $\gamma = 0.001$ for GTAdam, and $\gamma = 0.1$ for DGD, GT, and DAdam.

As it can be appreciated from Figure 2.17a and Figure 2.17b, in both cases GTAdam outperforms the other algorithms.



(a) Training loss.

(b) Training accuracy.

Figure 2.17: Distributed image classification. Mean and 3−standard deviation band of the training loss and the training accuracy.

Chapter 3

Tracking-Based Algorithms for Distributed Aggregative Optimization

In this chapter, we focus on distributed aggregative optimization problems, i.e., on the optimization scenario (already introduced in Section 1.3) in which a network of N agents aim to cooperatively solve problems in the form

$$\min_{(x_1,\dots,x_N)\in X} \sum_{i=1}^{N} f_i(x_i, \sigma(x)), \tag{3.1}$$

where each $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is the cost function known to agent i only, and the so-called aggregative variable $\sigma(x) \in \mathbb{R}^d$ is given by

$$\sigma(x) = \frac{1}{N} \sum_{i=1}^{N} \phi_i(x_i), \tag{3.2}$$

where each aggregation rule $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}^d$ is a function modeling the contribution of the i-th agent to the aggregative variable.

In Section 3.2, we start by addressing the online version of problem (3.1), namely the one in which all the cost functions, aggregation rules, and feasible sets vary over time. In this context, we propose Projected Aggregative Tracking, i.e., a distributed optimization algorithm that optimizes the overall cost function by combining (i) a distributed implementation of the projected gradient descent, and (ii) dynamic consensus techniques to reconstruct the global information that are not locally available. The algorithm generalizes an already existing method by allowing for more general online setups and for the use of constant step-sizes. Thanks to refined steps in the algorithm evolution, we improve the existing performance results. In detail, we provide tight

bounds for the dynamic regret and linear convergence in case of time-invariant optimization problems. Then, in Section 3.3, we consider a "personalized" setup in which each local function is given by the sum of a known term and an unknown one capturing the user's dissatisfaction. In this setting, we interlace the previous algorithm with a Recursive Least Squares (RLS) scheme to take advantage of users' noisy feedback to learn the parameters of the unknown function concurrently with the optimization steps. We prove an upper bound for the dynamic regret related to (i) the initial conditions, (ii) the temporal variations of the objective functions, and (iii) the learning errors. Moreover, by considering the average dynamic regret, we prove that both initial conditions and learning errors do not affect the asymptotic performance of the algorithm. Subsequently, in Section 3.4, we present Aggregative Tracking Feedback, i.e., a novel distributed feedback optimization law to steer network systems to a steady-state minimizing an aggregative optimization problem with (possibly) nonconvex objective function. The key feature of Aggregative Tracking Feedback is that it directly implements an optimization algorithm in closed-loop with a set of physical systems with nonlinear dynamics. We perform a system theoretical analysis to show that Aggregative Tracking Feedback steers the network to a stationary point of the optimization problem. Finally, we consider the case with single integrator dynamics and strongly convex objective function. In this case, we adapt Aggregative Tracking Feedback to get a closed loop system that exponentially converges to a configuration corresponding to the optimal solution of the problem. The results of this chapter are based on [26, 29, 30, 32].

3.1 Literature Review

Distributed aggregative optimization is a recently emerged framework in which a network of agents must cooperatively minimize the sum of local cost functions that depend both on a local optimization variable and on a global variable obtained by performing some kind of aggregation of all the local variables (as, e.g., the mean). This framework stems from distributed aggregative games (see Chapter 4) where however the objective is to compute a (generalized) Nash equilibrium rather than an optimal solution cooperatively. Some analogies can be also found with the so-called constraint-coupled framework. Indeed, the latter is a cooperative optimization framework where each local cost function depends on a local decision variable, but all the variables are coupled through separable coupling constraints [22–24, 38, 60, 61, 110, 123, 139, 173, 212].

Distributed aggregative optimization has been introduced in the pioneering work [104], where a static, unconstrained instance of (3.1) is tackled. Constrained, online version of the problem is investigated in [106], where the performance of the proposed algorithm is analyzed in terms of dynamic regret. The authors of [42] design a distributed algorithm for this setting to also handle communication with finite bits. In [192], the aggregative

framework is addressed through a distributed algorithm based on the Franke-Wolfe update to reduce the computational effort.

Personalized Optimization

In several domains there are devices with computation and communication capabilities directly involving end-users. For this reason, users' dissatisfaction needs to be taken into account together with engineering-oriented goals in many tasks. Human preferences are taken into account, e.g., for demand response tasks in the electric field [39, 145], to design robot trajectories [98, 120, 214], or for rehabilitation robots [126]. In this context, while engineering goals can be assessed by using well-known metrics, users' dissatisfaction is usually described through synthetic models. However, the complexity of human preferences leads to scarce and biased optimization outcomes for this kind of models. Hence, personalized strategies relying on users' feedback may improve the outcome. Firsts attempts in this direction are given in [146, 172], while recently users' feedback has been used in the context of distributed optimization [141].

Feedback Optimization

Feedback optimization techniques represent an emerging class of control laws aiming at steering dynamic systems toward steady-states while minimizing an associated optimization problem, see the recent surveys [80, 96] for an overview. The key feature of feedback optimization controllers is that they only rely on real-time gradient measurements, thus avoiding the knowledge of the objective function of the optimization problem. Applications for such a control paradigm can be found in several fields ranging from real-time optimal power flow in electrical networks, see [47, 180], to congestion control in communication networks, [118]. First attempts for the design of these controllers leverage the so-called extremum seeking techniques. In this context, the estimate of the gradient of an unknown objective function is obtained and used to steer the system toward its minimizer, [5, 97, 178, 182, 195]. In [128], a feedback optimization law has been designed and applied to a power system setup. In [78, 81, 83] feedback optimization has been used to implement model-free optimization algorithms with constraint handling. In [144], algebraic systems are controlled by relying on gradient information affected by random errors modelled as Sub-Weibull distributions. In [45], a feedback optimization technique is designed for linear time-invariant systems. The approach is based on gradient flow dynamics augmented with learning methods to estimate the cost function based on infrequent and possibly noisy data. A distributed feedback optimization law has been proposed in [183] to address a partition-based optimization scenario over a network of communicating systems.

3.2 Distributed Online Aggregative Optimization

In this section, we address online instances of problem (3.1)

$$\min_{(x_1,\dots,x_N)\in X^k} \sum_{i=1}^{N} f_i^k(x_i, \sigma^k(x)), \quad k \geq 0 \tag{3.3}$$

in which $x := \text{col}(x_1, \dots, x_N) \in \mathbb{R}^n$ is the global decision vector, with each $x_i \in \mathbb{R}^{n_i}$ and $n = \sum_{i=1}^N n_i$. The global decision vector at iteration index k is constrained to belong to a set $X^k \subseteq \mathbb{R}^n$ that can be written as $X^k = (X_1^k \times \ldots \times X_N^k)$, where each $X_i^k \subseteq \mathbb{R}^{n_i}$. The functions $f_i^k : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}$ represent the local objective functions at iteration k, while the *aggregation function* $\sigma^k(x)$ has the form

$$\sigma^k(x) := \frac{\sum_{i=1}^N \phi_i^k(x_i)}{N}, \tag{3.4}$$

where each $\phi_i^k : \mathbb{R}^{n_i} \to \mathbb{R}^d$ is the i-th contribution to the aggregative variable at iteration k. We compactly denote the cost function of problem (3.3) as $f^k(x, \sigma^k(x)) := \sum_{i=1}^N f_i^k(x_i, \sigma^k(x))$. In problem (3.3), $f^k(\cdot, \sigma^k(\cdot))$ is not known to any agent: each of them can only privately access f_i^k, X_i^k, and ϕ_i^k. We remark that each agent i accesses its private information f_i^k, and ϕ_i^k only once its estimate x_i^k has been computed.

The goal is to design distributed algorithms to seek a minimum for problem (3.3). Next, we will denote as $\nabla_1 f_i^k(\cdot,\cdot)$ and as $\nabla_2 f_i^k(\cdot,\cdot)$ the gradient of f_i^k with respect to respectively the first argument and the second argument. Moreover, we also introduce $G^k : \mathbb{R}^n \times \mathbb{R}^{Nd} \to \mathbb{R}^n$ defined as $G^k(x,s) := G_1^k(x,s) + \nabla\phi(x)\frac{\mathbf{1}_{N,d}}{N}\sum_{i=1}^N f_i^k(x_i, s_i)$, where $x := \text{col}(x_1, \dots, x_N) \in \mathbb{R}^n$, $s := \text{col}(s_1, \dots, s_N) \in \mathbb{R}^{Nd}$ with each $x_i \in \mathbb{R}^{n_i}$, $s_i \in \mathbb{R}^d$ for all $i \in \{1, \dots, N\}$, $G_1^k(x,s) := \text{col}(\nabla_1 f_{1,t}(x_1, s_1), \dots, \nabla_1 f_{N,t}(x_N, s_N))$, and $\nabla\phi(x) := \text{blkdiag}(\nabla\phi_1(x_1), \dots, \nabla\phi_N(x_N)) \in \mathbb{R}^{n\times Nd}$.

Let x_i^k be the solution estimate of the problem at iteration k maintained by agent i, and let $x_\star^k$ be the (unique) minimizer of $f^k(x, \sigma^k(x))$ over the set X^k. Indeed, as we will formalize within Assumption 3.2, strong convexity of $f^k(x, \sigma^k(x))$ guarantees existence (and uniqueness) of $x_\star^k$ (cf. Proposition A.2 in Appendix A). As in Section 2.6, given a finite value $T > 1$, the agents want to minimize the dynamic regret:

$$R_T := \sum_{t=1}^{T} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{t=1}^{T} f^k(x_\star^k, \sigma^k(x_\star^k)). \tag{3.5}$$

Another popular metric is the so-called static regret [86]. However, as done in most of the literature, we focus on (3.5), which is more challenging to handle.

3.2.1 Projected Aggregative Tracking: Algorithm Description and Analysis

In this section, we propose and analyze Projected Aggregative Tracking, i.e., a novel distributed algorithm to address problem (3.3). Each agent i maintains for iteration k an estimate x_i^k of the component i of a minimum $x_\star^k$ of problem (3.3). In order to reconstruct the descent direction and use it to update the estimate x_i^k, agent i needs to reconstruct the global information $\sum_{i=1}^N \frac{\phi_i^k(\mathrm{x}_i^k)}{N}$ and $\sum_{i=1}^N \nabla_2 f_i^k \left(\mathrm{x}_i^k, \sum_{j=1}^N \frac{\phi_j^k(w_{ij})}{N}\right)$, which are not locally available. To overcome this lack of information, agent i maintains auxiliary variables s_i^k and y_i^k and iteratively updates them according to a perturbed consensus mechanism. A pseudo-code of the Projected Aggregative Tracking algorithm is reported in Algorithm 5 from the perspective of agent i, in which γ is a positive constant step-size, $\delta \in (0,1)$ is a constant algorithm parameter, and each element w_{ij} represents the (i,j) entry of the weighted adjacency matrix $\mathcal{W}_\mathcal{G}$ of the network.

Algorithm 5 Projected Aggregative Tracking (Agent i)

initialization:

$$\mathrm{x}_i^0 \in X_i^0, \quad \mathrm{s}_i^0 = \phi_i^0(\mathrm{x}_i^0), \quad \mathrm{y}_i^0 = \nabla_2 f_i^0(\mathrm{x}_i^0, \mathrm{s}_i^0)$$

for $k = 0, 1, \dots$ **do**

$$\begin{aligned}
\tilde{\mathrm{x}}_i^k &= P_{X_i^k}\left[\mathrm{x}_i^k - \gamma(\nabla_1 f_i^k(\mathrm{x}_i^k, \mathrm{s}_i^k) + \nabla\phi_i^k(\mathrm{x}_i^k)\mathrm{y}_i^k)\right] \\
\mathrm{x}_i^{k+1} &= \mathrm{x}_i^k + \delta(\tilde{\mathrm{x}}_i^k - \mathrm{x}_i^k) \\
\mathrm{s}_i^{k+1} &= \sum_{j=1}^N w_{ij}\mathrm{s}_j^k + \phi_i^{k+1}(\mathrm{x}_i^{k+1}) - \phi_i^k(\mathrm{x}_i^k) \\
\mathrm{y}_i^{k+1} &= \sum_{j=1}^N w_{ij}\mathrm{y}_j^k + \nabla_2 f_i^{k+1}(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}) - \nabla_2 f_i^k(\mathrm{x}_i^k, \mathrm{s}_i^k)
\end{aligned}$$

end for

In order to analyze the convergence properties of the proposed scheme, we rewrite Algorithm 5 in a stacked vector form as

$$\tilde{\mathrm{x}}^k = P_{X^k}\left[\mathrm{x}^k - \gamma(\nabla_1 f^k(\mathrm{x}^k, \mathrm{s}^k) + \nabla\phi^k(\mathrm{x}^k)\mathrm{y}^k)\right] \tag{3.6a}$$

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k) \tag{3.6b}$$

$$\mathrm{s}^{k+1} = \mathcal{W}\mathrm{s}^k + \phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k) \tag{3.6c}$$

$$\mathrm{y}^{k+1} = \mathcal{W}\mathrm{y}^k + G_2^{k+1}(\mathrm{x}^{k+1}, \mathrm{s}^{k+1}) - G_2^k(\mathrm{x}^k, \mathrm{s}^k), \tag{3.6d}$$

where we introduced the symbols $\mathcal{W} := \mathcal{W}_{\mathcal{G}} \otimes I_d$ and

$$\mathrm{x}^k := \begin{bmatrix} \mathrm{x}_1^k \\ \dots \\ \mathrm{x}_N^k \end{bmatrix}, \quad \mathrm{s}^k := \begin{bmatrix} \mathrm{s}_1^k \\ \vdots \\ \mathrm{s}_N^k \end{bmatrix}, \quad \mathrm{y}^k := \begin{bmatrix} \mathrm{y}_1^k \\ \vdots \\ \mathrm{y}_N^k \end{bmatrix}, \quad G_2^k(\mathrm{x}^k, \mathrm{s}^k) := \begin{bmatrix} \nabla_2 f_1^k(\mathrm{x}_1^k, \mathrm{s}_1^k) \\ \vdots \\ \nabla_2 f_N^k(\mathrm{x}_N^k, \mathrm{s}_N^k) \end{bmatrix}.$$

In order to perform the convergence analysis, we derive bounds for the quantities $\left\|\mathrm{x}^{k+1} - \mathrm{x}^k\right\|$, $\left\|\mathrm{x}^{k+1} - x_\star^{k+1}\right\|$, $\left\|\mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1}\right\|$, and $\left\|\mathrm{s}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{s}}^{k+1}\right\|$, in which $\bar{\mathrm{y}}^k := \frac{1}{N}\sum_{i=1}^N \mathrm{y}_i^k$ and $\bar{\mathrm{s}}^k := \frac{1}{N}\sum_{i=1}^N \mathrm{s}_i^k$ denote the mean vectors of y^k and s^k, respectively. Let z^k be the vector staking the above quantities

$$\mathrm{z}^k := \begin{bmatrix} \left\|\mathrm{x}^k - x_\star^k\right\| \\ \left\|\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\| \\ \left\|\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\| \end{bmatrix}. \tag{3.7}$$

Moreover, also the following variables will be useful to provide the main result of the paper, namely

$$\eta^k := \sup_{x\in\mathbb{R}^n, z\in\mathbb{R}^{Nd}} \left\|G_2^{k+1}(x,z) - G_2^k(x,z)\right\| \tag{3.8a}$$

$$\omega^k := \sup_{x\in\mathbb{R}^n} \left\|\phi^{k+1}(x) - \phi^k(x)\right\| \tag{3.8b}$$

$$\alpha^k := \sup_{x\in\mathbb{R}^n} \left|\mathrm{dist}(x, X^{k+1}) - \mathrm{dist}(x, X^k)\right| \tag{3.8c}$$

$$\zeta^k := \|x_\star^{k+1} - x_\star^k\|, \tag{3.8d}$$

where we recall that $x_\star^k$ is the optimal solution of $f^k(\cdot, \sigma^k(\cdot))$. Next, we state the assumptions of our framework.

Assumption 3.1 (Communication graph)**.** *The graph $\mathcal{G}$ is undirected and connected and $\mathcal{W}_{\mathcal{G}}$ is doubly stochastic.* △

Assumption 3.2 (Convexity)**.** *For all $i \in \{1, \dots, N\}$ and all $k \geq 0$, $X_i^k \subseteq \mathbb{R}^{n_i}$ is nonempty, closed and convex, while the global objective function $f^k(x, \sigma^k(x))$ is μ-strongly convex.* △

Assumption 3.3 (Function Regularity)**.** *For all $k \geq 0$, the function $f^k(x, \sigma^k(x))$ is differentiable with $\overline{L}_1$-Lipschitz continuous gradients, and $G^k(x,s)$, $G_2^k(x,s)$ are Lipschitz continuous with constants $\overline{L}_1, \overline{L}_2 > 0$, respectively. For all $i \in \{1, \dots, N\}$ and $k \geq 0$, the aggregation function $\phi_i^k(x_i)$ is differentiable and $\overline{L}_3$-Lipschitz continuous, and η^k and ω^k are finite.* △

We start by noting that

$$\bar{\mathrm{s}}^{k+1} = \bar{\mathrm{s}}^k + \frac{\mathbf{1}_{N,d}^\top}{N}(\phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k)) \tag{3.9a}$$

$$\bar{\mathrm{y}}^{k+1} = \bar{\mathrm{y}}^k + \frac{\mathbf{1}_{N,d}^\top}{N}(G_2^{k+1}(\mathrm{x}^{k+1}, \mathrm{s}^{k+1}) - G_2^k(\mathrm{x}^k, \mathrm{s}^k)). \tag{3.9b}$$

Then, if we initialize σ and y as $\sigma^0 := \phi^0(\mathrm{x}^0)$ and $y^0 := G_2^0(x^0, s^0)$, from (3.9a) and (3.9b), it holds for all $k \geq 0$

$$\bar{\mathrm{s}}^k = \frac{1}{N}\sum_{i=1}^N \phi^k(\mathrm{x}_i^k) = \sigma^k(\mathrm{x}^k) \tag{3.10a}$$

$$\bar{\mathrm{y}}^k = \frac{1}{N}\sum_{i=1}^N \nabla_2 f_i^k(\mathrm{x}_i^k, \mathrm{s}_i^k). \tag{3.10b}$$

Now, we present four preparatory Lemmas that we need to prove the main result of this section, i.e., Theorem 3.1. For brevity, we will use d^k to denote the descent direction used within the update (3.41a), i.e.,

$$\mathrm{d}^k := \nabla_1 f^k(\mathrm{x}^k, \mathrm{s}^k) + \nabla\phi^k(\mathrm{x}^k)\mathrm{y}^k. \tag{3.11}$$

Lemma 3.1. *Let Assumptions 3.1, 3.2, and 3.3 hold. If $\gamma \leq \frac{1}{\overline{L}_1}$, then*

$$\left\|\mathrm{x}^{k+1} - x_\star^{k+1}\right\| \leq (1-\delta\mu\gamma)\left\|\mathrm{x}^k - x_\star^k\right\| + \delta\gamma\overline{L}_1\left\|\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\| + \delta\gamma\overline{L}_3\left\|\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\| + \zeta^k.$$

Proof. We begin by using (3.41b), which leads to

$$\begin{aligned}\left\|\mathrm{x}^{k+1} - x_\star^{k+1}\right\| &= \left\|\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k) - x_\star^{k+1}\right\| \\ &\overset{(a)}{\leq} \left\|\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k) - x_\star^k\right\| + \left\|x_\star^{k+1} - x_\star^k\right\| \\ &\overset{(b)}{\leq} \left\|\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k) - x_\star^k\right\| + \zeta^k,\end{aligned} \tag{3.12}$$

where in *(a)* we add and subtract the term $x_\star^k$ and use the triangle inequality, and in *(b)* we use ζ^k (cf (3.8d)). Being $x_\star^k$ the minimizer of f^k over X^k, then it holds $P_{X^k}\left[x_\star^k - \gamma f^k(x_\star^k, \sigma^k(x_\star^k))\right] = x_\star^k$. Then, we add the null term

$$\delta\left(P_{X^k}\left[x_\star^k - \gamma\nabla f^k(x_\star^k, \sigma^k(x_\star^k))\right] - x_\star^k\right)$$

in the first norm of (3.12) and we apply the triangle inequality and (3.6a) to write

$$\begin{aligned}\left\|\mathrm{x}^{k+1} - x_\star^{k+1}\right\| &\leq (1-\delta)\left\|\mathrm{x}^k - x_\star^k\right\| + \delta\left\|P_{X^k}[\mathrm{x}^k - \gamma\mathrm{d}^k] - P_{X^k}\left[x_\star^k - \gamma\nabla f^k(x_\star^k, \sigma^k(x_\star^k))\right]\right\| \\ &\quad + \zeta^k \\ &\overset{(a)}{\leq} (1-\delta)\left\|\mathrm{x}^k - x_\star^k\right\| + \delta\left\|\mathrm{x}^k - \gamma\mathrm{d}^k - (x_\star^k - \gamma\nabla f^k(x_\star^k, \sigma^k(x_\star^k)))\right\| + \zeta^k,\end{aligned} \tag{3.13}$$

where *(a)* uses the non-expansiveness of the projection, see [14]. Add and subtract within the second norm $\gamma\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))$ and apply the triangle inequality to rewrite (3.13) as

$$\begin{aligned}\left\|\mathrm{x}^{k+1}-x_\star^{k+1}\right\| &\leq (1-\delta)\left\|\mathrm{x}^k-x_\star^k\right\| \\ &\quad+\delta\left\|\mathrm{x}^k-\gamma\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))-\left(x_\star^k-\gamma f^k(x_\star^k,\sigma^k(x_\star^k))\right)\right\| \\ &\quad+\delta\gamma\left\|\mathrm{d}^k-\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))\right\|+\zeta^k \\ &\overset{(a)}{\leq}(1-\delta\mu\gamma)\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\left\|\mathrm{d}^k-\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))\right\|+\zeta^k,\end{aligned}$$

where *(a)* uses [104, Lemma 3]. Add and subtract into the second norm the term $\nabla\phi^k(\mathrm{x}^k)\mathbf{1}_{N,d}\frac{1}{N}\sum_{i=1}^N\nabla_2 f_i^k(\mathrm{x}_i^k,\mathrm{s}_i^k)$, and rearrange as

$$\begin{aligned}\left\|\mathrm{x}^{k+1}-x_\star^{k+1}\right\| &\leq (1-\delta\mu\gamma)\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))\right\| \\ &\quad+\delta\gamma\left\|\nabla\phi^k(\mathrm{x}^k)\left(\mathrm{y}^k-\mathbf{1}_{N,d}\frac{1}{N}\sum_{i=1}^N\nabla_2 f_i^k(\mathrm{x}_i^k,\mathrm{s}_i^k)\right)\right\|+\zeta^k \\ &\overset{(a)}{=}(1-\delta\mu\gamma)\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))\right\| \\ &\quad+\delta\gamma\left\|\nabla\phi^k(\mathrm{x}^k)\left(\mathrm{y}^k-\mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right)\right\|+\zeta^k, \qquad (3.14)\end{aligned}$$

where in *(a)* we use (3.10b). Consider the term $\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))\|$. The definition of G^k and (3.10a) gives

$$\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(\mathrm{x}^k,\sigma^k(\mathrm{x}^k))\right\|=\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(\mathrm{x}^k,\bar{\mathrm{s}}^k)\right\|\overset{(a)}{\leq}\overline{L}_1\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\|, \qquad (3.15)$$

where *(a)* uses the Lipschitz continuity of G^k (cf. Assumption 3.3). The proof follows by (3.14), (3.15), and $\left\|\nabla\phi^k(x)\right\|\leq\overline{L}_3$ for all $x\in\mathbb{R}^n$ (which is derived from Assumption 3.3). ■

Lemma 3.2. *Let Assumptions 3.1, 3.2, and 3.3 hold. Then*

$$\left\|\mathrm{x}^{k+1}-\mathrm{x}^k\right\|\leq\delta(2+\gamma\overline{L}_1+\gamma L_1L_3)\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\overline{L}_1\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\|+\delta\gamma\overline{L}_3\left\|\mathrm{y}^k-\mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\|.$$

Proof. We can use (3.6b) to write

$$\left\|\mathrm{x}^{k+1}-\mathrm{x}^k\right\|=\left\|\mathrm{x}^k+\delta(\tilde{\mathrm{x}}^k-\mathrm{x}^k)-\mathrm{x}^k\right\|=\delta\left\|\tilde{\mathrm{x}}^k-\mathrm{x}^k\right\|\overset{(a)}{=}\delta\left\|P_{X^k}[\mathrm{x}^k-\gamma\mathrm{d}^k]-\mathrm{x}^k\right\|,$$

where in *(a)* we have used the update (3.41a). By adding the null quantity

$$\left(P_{X^k}\left[x_\star^k-\gamma\nabla f^k(x_\star^k,\sigma^k(x_\star^k))\right]-x_\star^k\right)$$

within the norm and applying the triangle inequality, we get

$$\begin{aligned}\left\|\mathrm{x}^{k+1}-\mathrm{x}^k\right\| &\leq \delta\left\|P_{X^k}[\mathrm{x}^k-\gamma\mathrm{d}^k]-P_{X^k}\left[x_\star^k-\gamma\nabla f^k(x_\star^k,\sigma^k(x_\star^k))\right]\right\|+\delta\left\|\mathrm{x}^k-x_\star^k\right\|\\ &\overset{(a)}{\leq} 2\delta\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\left\|\mathrm{d}^k-\nabla f^k(x_\star^k,\sigma^k(x_\star^k))\right\|,\end{aligned}$$

where in *(a)* we use a projection property and the triangle inequality. We add and subtract within the norm the term $\nabla\phi^k(\mathrm{x}^k)\mathbf{1}_{N,d}\sum_{i=1}^N\nabla_2 f_i^k(\mathrm{x}_i^k,\mathrm{s}_i^k)$ and use the expression of d^k and G^k and the triangle inequality to write

$$\begin{aligned}\left\|\mathrm{x}^{k+1}-\mathrm{x}^k\right\| &\leq 2\delta\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(x_\star^k,\sigma^k(x_\star^k))\right\|\\ &\quad+\delta\gamma\left\|\nabla\phi^k(\mathrm{x}^k)\left(\mathrm{y}^k-\mathbf{1}_{N,d}\sum_{i=1}^N\nabla_2 f_i^k(\mathrm{x}_i^k,\mathrm{s}_i^k)\right)\right\|\\ &\overset{(a)}{=} 2\delta\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(x_\star^k,\sigma^k(x_\star^k))\right\|\\ &\quad+\delta\gamma\overline{L}_3\left\|\mathrm{y}^k-\mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\|,\end{aligned}\tag{3.16}$$

where in *(a)* we use (3.10b) and $\left\|\nabla\phi^k(x)\right\|\leq\overline{L}_3$. The definition of G^k and its Lipschitz continuity (cf. Assumption 3.3) imply

$$\left\|G^k(\mathrm{x}^k,\mathrm{s}^k)-\nabla f^k(x_\star^k,\sigma^k(x_\star^k))\right\|\leq\overline{L}_1\left\|\mathrm{x}^k-x_\star^k\right\|+\overline{L}_1\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\sigma^k(x_\star^k)\right\|.\tag{3.17}$$

By combining (3.16) with (3.17), we get

$$\begin{aligned}\left\|\mathrm{x}^{k+1}-\mathrm{x}^k\right\| &\leq \delta(2+\gamma\overline{L}_1)\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\overline{L}_1\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\sigma^k(x_\star^k)\right\|+\delta\gamma\overline{L}_3\left\|\mathrm{y}^k-\mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\|\\ &\overset{(a)}{\leq}\delta(2+\gamma\overline{L}_1)\left\|\mathrm{x}^k-x_\star^k\right\|+\delta\gamma\overline{L}_1\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\|\\ &\quad+\delta\gamma\overline{L}_1\left\|\mathbf{1}_{N,d}\bar{\mathrm{s}}^k-\mathbf{1}_{N,d}\sigma^k(x_\star^k)\right\|+\delta\gamma\overline{L}_3\left\|\mathrm{y}^k-\mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\|,\end{aligned}\tag{3.18}$$

where in *(a)* we add and subtract $\mathbf{1}_{N,d}\bar{\mathrm{s}}^k$ and we apply the triangle inequality. Now, consider the term $\left\|\mathbf{1}_{N,d}\bar{\mathrm{s}}^k-\mathbf{1}_{N,d}\sigma^k(x_\star^k)\right\|$. By using the definition of σ^k and the Lipschitz continuity of ϕ_i^k (cf. Assumption 3.3), it can be seen that (see also [104]) ,

$$\left\|\mathbf{1}_{N,d}\bar{\mathrm{s}}^k-\mathbf{1}_{N,d}\sigma^k(x_\star^k)\right\|\leq\overline{L}_3\left\|\mathrm{x}^k-x_\star^k\right\|.\tag{3.19}$$

By combining the results (3.18) and (3.19), the proof is given. ■

Lemma 3.3. *Let Assumptions 3.1, 3.2, and 3.3 hold. Then*

$$\left\|\mathrm{s}^{k+1}-\mathbf{1}_{N,d}\bar{\mathrm{s}}^{k+1}\right\|\leq\Lambda\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\|+\delta\gamma L_1L_3\left\|\mathrm{s}^k-\mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\|$$

$$+ \delta(2\overline{L}_3 + \gamma L_1 L_3 + \gamma L_1 L_3^2) \left\|\mathrm{x}^k - x_\star^k\right\| + \delta\gamma \overline{L}_3^2 \left\|\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\| + \omega^k,$$

where Λ is the maximum eigenvalue of the matrix $\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$.

Proof.

By applying (3.6c) and (3.9a), we can write

$$\begin{aligned}\left\|\mathrm{s}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{s}}^{k+1}\right\| &= \left\|A\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k + \tilde{I}(\phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k))\right\| \\ &\overset{(a)}{\leq} \left\|\left(A - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}\right)\left(\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right)\right\| + \left\|\tilde{I}(\phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k))\right\|,\end{aligned}$$

where *(a)* applies the triangle inequality, introduces $\tilde{I} := I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$, and uses the fact that $\mathbf{1}_{N,d} \in \ker\left(\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}\right)$. Now, we add and subtract within the second norm the term $\phi^{k+1}(\mathrm{x}^k)$ and we apply the triangle inequality obtaining

$$\begin{aligned}\left\|\mathrm{s}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{s}}^{k+1}\right\| &\leq \left\|\left(A - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}\right)\left(\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right)\right\| \\ &\quad + \left\|\tilde{I}(\phi^{k+1}(\mathrm{x}^{k+1}) - \phi^{k+1}(\mathrm{x}^k))\right\| + \left\|\tilde{I}(\phi^{k+1}(\mathrm{x}^k) - \phi^k(\mathrm{x}^k))\right\| \\ &\overset{(a)}{\leq} \Lambda\left\|\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\| + \overline{L}_3\left\|\mathrm{x}^{k+1} - \mathrm{x}^k\right\| + \omega^k,\end{aligned}$$

where in *(a)* we use the maximum eigenvalue Λ of the matrix $\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$, Assumption 3.3, ω^k (cf (3.8b)), and $\left\|\tilde{I}\right\| = 1$. By using Lemma 3.2 to bound $\left\|\mathrm{x}^{k+1} - \mathrm{x}^k\right\|$, the proof follows. ■

Lemma 3.4. *Let Assumptions 3.1, 3.2, and 3.3 hold. Then*

$$\begin{aligned}\left\|\mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1}\right\| &\leq \Lambda\left\|\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\| + \delta\gamma\overline{L}_3(\overline{L}_2 + L_2L_3)\left\|\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\| \\ &\quad + \delta(2 + \gamma\overline{L}_1 + \gamma\overline{L}_1)(\overline{L}_2 + L_2L_3)\left\|\mathrm{x}^k - x_\star^k\right\| \\ &\quad + \delta\gamma\overline{L}_1(\overline{L}_2 + L_2L_3)\left\|\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\| + 2L_2\left\|\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\| + \overline{L}_2\omega^k + \eta^k,\end{aligned}$$

where Λ is the maximum eigenvalue of the matrix $\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$.

Proof. We use (3.6d) and (3.9b) to write

$$\left\|\mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1}\right\| \leq \left\|\mathcal{W}\mathrm{y}^k - \bar{\mathrm{y}}^k\right\|$$

$$
\begin{aligned}
&+ \left\| \left(I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} \right) (G_2^{k+1}(\mathrm{x}^{k+1}, \mathrm{s}^{k+1}) - G_2^k(\mathrm{x}^k, \mathrm{s}^k)) \right\| \\
&\overset{(a)}{\leq} \left\| \left(\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} \right) (\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k) \right\| \\
&\quad + \left\| \left(I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} \right) (G_2^{k+1}(\mathrm{x}^{k+1}, \mathrm{s}^{k+1}) - G_2^{k+1}(\mathrm{x}^k, \mathrm{s}^k)) \right\| \\
&\quad + \left\| \left(I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} \right) (G_2^{k+1}(\mathrm{x}^k, \mathrm{s}^k) - G_2^k(\mathrm{x}^k, \mathrm{s}^k)) \right\|,
\end{aligned}
$$

where *(a)* uses $\mathbf{1}_{N,d} \in \ker(\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N})$ and applies the triangle inequality after adding and subtracting $(I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N})G_2^{k+1}(\mathrm{x}^k, \mathrm{s}^k)$ within the norm. By using the maximum eigenvalue Λ of $\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$, Assumption 3.3, and η^k (cf. (3.8a)), we get

$$
\left\| \mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1} \right\| \quad \leq \Lambda \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \overline{L}_2 \left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| + \overline{L}_2 \left\| \mathrm{s}^{k+1} - \mathrm{s}^k \right\| + \eta^k.
$$

Now, we can use (3.6d) to get

$$
\begin{aligned}
\left\| \mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1} \right\| &\leq \Lambda \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \overline{L}_2 \left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| \\
&\quad + \overline{L}_2 \left\| (\mathcal{W} - I)\mathrm{s}^k + \phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k) \right\| + \eta^k \\
&\overset{(a)}{\leq} \Lambda \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \overline{L}_2 \left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| + \overline{L}_2 \left\| (\mathcal{W} - I)(\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k) \right\| \\
&\quad + \overline{L}_2 \left\| \phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k) \right\| + \eta^k,
\end{aligned}
$$

where in *(a)* we apply the triangle inequality and the fact that $\mathbf{1}_{N,d} \in \ker\left(\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}\right)$. We add and subtract within the norm the term $\phi^{k+1}(\mathrm{x}^k)$ and apply the triangle inequality, obtaining

$$
\begin{aligned}
\left\| \mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1} \right\| &\leq \Lambda \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \overline{L}_2 \left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| + \overline{L}_2 \left\| (\mathcal{W} - I)(\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k) \right\| \\
&\quad + \overline{L}_2 \left\| \phi^{k+1}(\mathrm{x}^{k+1}) - \phi^{k+1}(\mathrm{x}^k) \right\| + \overline{L}_2 \left\| \phi^{k+1}(\mathrm{x}^k) - \phi^k(\mathrm{x}^k) \right\| + \eta^k \\
&\overset{(a)}{\leq} \rho \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \overline{L}_2 \left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| + \overline{L}_2 \left\| (\mathcal{W} - I)(\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k) \right\| \\
&\quad + L_2 L_3 \left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| + \overline{L}_2 \omega^k + \eta^k,
\end{aligned}
$$

where *(a)* uses Assumption 3.3 and ω^k (cf. (3.8b)). The proof follows by $\|\mathcal{W} - I\| \leq 2$, and by applying Lemma 3.2. ■

Now, we state the main theoretical results of the section. Next theorem provides

a bound on the dynamic regret of the iterates generated by the Projected Aggregative Tracking distributed algorithm in the general, online setup (3.3).

Theorem 3.1. *Consider Projected Aggregative Tracking as given in Algorithm 5. Let Assumptions 3.1, 3.2, and 3.3 hold. Then, there exists $\lambda, \bar{\delta} > 0$ and $\tilde{\rho} \in (0,1)$ so that, if $\gamma \leq \frac{1}{\overline{L}_1}$ and $\delta \in (0, \bar{\delta})$, it holds*

$$R_T \leq \frac{\overline{L}_1 \lambda^2}{2} \left(\frac{\left\| z^0 \right\|^2}{1-\tilde{\rho}^2} + 2 \left\| z^0 \right\| W_T + Q_T \right), \tag{3.20}$$

where R_T is defined as in (3.5) *and*

$$W_T := \sum_{k=1}^{T} \sum_{q=0}^{k-1} \tilde{\rho}^{k+q} \left(\left\| \zeta^{k-q-1} \right\| + 2 \left\| \eta^{k-q-1} \right\| + (1+\overline{L}_2) \left\| \omega^{k-q-1} \right\| \right), \tag{3.21a}$$

$$Q_T := \sum_{k=1}^{T} \left(\sum_{q=0}^{k-1} \tilde{\rho}^q \left(\zeta^{k-q-1} + 2\eta^{k-q-1} + (1+\overline{L}_2)\omega^{k-q-1} \right) \right)^2. \tag{3.21b}$$

Moreover, if α^k (cf. (3.8c)*) is finite for all $k \geq 0$, then the constraint violation is bounded by*

$$\sum_{k=1}^{T} \operatorname{dist}(\mathrm{x}^k, X^k) \leq \frac{1}{1-(1-\delta)^T} \operatorname{dist}(\mathrm{x}^0, X^0) + \sum_{k=1}^{T} \sum_{q=0}^{k-1} (1-\delta)^q \alpha^{k-q-1}. \tag{3.22}$$

Proof. Let us introduce u^k to denote $u^k := \textsc{col}(\zeta^k, \eta^k, \omega^k)$. Then, by combining Lemma 3.1, 3.3, and 3.4, we bound the evolution of z^k (defined in (3.7)) through the following dynamical system

$$\mathrm{z}^{k+1} \leq M(\delta)\mathrm{z}^k + Bu^k, \tag{3.23}$$

in which

$$M(\delta) := M_0 + \delta E, \quad B := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & \overline{L}_2 \end{bmatrix},$$

where

$$M_0 := \begin{bmatrix} 1 & 0 & 0 \\ 0 & \Lambda & 0 \\ 0 & 2L_2 & \Lambda \end{bmatrix}, \quad E := \begin{bmatrix} -\mu\gamma & \gamma\overline{L}_1 & \gamma\overline{L}_3 \\ E_{21} & \gamma L_1 L_3 & \gamma\overline{L}_3^2 \\ E_{31} & E_{32} & E_{33} \end{bmatrix},$$

with $E_{21} := 2\overline{L}_3 + \gamma L_1 L_3 + \gamma L_1 L_3^2$, $E_{31} := (2 + \gamma\overline{L}_1 + \gamma\overline{L}_1)(\overline{L}_2 + L_2 L_3)$, $E_{32} := \gamma\overline{L}_1(\overline{L}_2 + L_2 L_3)$, and $E_{33} := \gamma\overline{L}_3(\overline{L}_2 + L_2 L_3)$. Being M_0 triangular, its spectral radius is 1 since $\Lambda \in (0,1)$ as implied by Assumption 3.1. Denote by $\chi(\delta)$ the eigenvalues of $M(\delta)$ as a

function of δ. Call v and w respectively the right and left eigenvectors of M_0 associated to 1. Then, $v = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$, $w = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$. Being 1 a simple eigenvalue of $M(0)$, from Theorem B.1 (in Appendix B) it holds

$$\left.\frac{d\chi(\delta)}{d\delta}\right|_{\chi=1,\delta=0} = \frac{w^\top E v}{w^\top v} = -\mu\gamma < 0.$$

Then, by continuity of eigenvalues with respect to the matrix entries, there exists $\bar{\delta} > 0$ so that $\rho_{\max}(M(\delta)) < 1$ for any $\delta \in (0, \bar{\delta})$. From now on we will omit the dependency of M and its eigenvalues from δ. Since $\mathrm{z}^k \geq 0$ for all k, and M and Bu^k have only non-negative entries, one can use (3.23) to write

$$\mathrm{z}^k \leq M^k \mathrm{z}^0 + \sum_{q=0}^{k-1} M^q B u^q. \tag{3.24}$$

Pick $\theta \in (0, 1 - \rho_{\max}(M))$ and define $\tilde{\rho} := \rho_{\max}(M) + \theta$. Then, by [85, Lemma 5.6.10], there exists a matrix norm[1], which we denote as $\|\cdot\|_\iota$, such that $\|M\|_\iota \leq \rho_{\max}(M) + \theta < 1$. Moreover, by applying [85, Theorem 5.7.13], there exists a vector norm, which we denote by $\|\cdot\|_\iota$, which is compatible with the corresponding matrix norm, i.e., such that $\|Mv\|_\iota \leq \|M\|_\iota \|v\|_\iota$ for any matrix $M \in \mathbb{R}^{3\times 3}$ and $v \in \mathbb{R}^3$. Using this fact, we use the norm $\|\cdot\|_\iota$ on both sides of (3.24) and we apply the triangle inequality to get

$$\begin{aligned} \left\|\mathrm{z}^k\right\|_\iota &\leq \left\|M^k \mathrm{z}^0\right\|_\iota + \left\|\sum_{q=0}^{k-1} M^q B u^{k-q-1}\right\|_\iota \\ &\leq \tilde{\rho}^k \left\|\mathrm{z}^0\right\|_\iota + \sum_{q=0}^{k-1} \tilde{\rho}^q \left\|B u^{k-q-1}\right\|_\iota. \end{aligned} \tag{3.25}$$

Being ∇f^k Lipschitz continuous (cf. Assumption 3.3), it holds

$$f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - f^k(x_\star^k, \sigma^k(x_\star^k)) \leq \frac{\overline{L}_1}{2}\|\mathrm{x}^k - x_\star^k\|^2 \overset{(a)}{\leq} \frac{\overline{L}_1}{2}\|\mathrm{z}^k\|^2, \tag{3.26}$$

where in *(a)* uses the fact that $\left\|\mathrm{x}^k - x_\star^k\right\|$ is a component of z^k. Recalling that all norms are equivalent on finite-dimensional vector spaces, there always exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\|\cdot\| \leq \lambda_1 \|\cdot\|_\iota$ and $\|\cdot\|_\iota \leq \lambda_2 \|\cdot\|$. Thus, by exploiting the square norm and combining the results (3.25) with the equivalence of the norms, we can bound (3.26) as

$$f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - f^k(x_\star^k, \sigma^k(x_\star^k))$$

[1] An expression of $\|\cdot\|_\iota$ can be found in the proof of [85, Lemma 5.6.10].

$$\leq \frac{\overline{L}_1\lambda_1^2}{2}\left(\tilde{\rho}^{2k}\left\|\mathrm{z}^0\right\|_\iota^2 + 2\tilde{\rho}^k\left\|\mathrm{z}^0\right\|_\iota \sum_{q=0}^{k-1}\tilde{\rho}^q\left\|Bu^{k-q-1}\right\|_\iota + \left(\sum_{q=0}^{k-1}\tilde{\rho}^q\left\|Bu^{k-q-1}\right\|_\iota\right)^2\right),$$

which, combined with the definitions of R_T (cf. (3.5)), W_T, and Q_T (cf. (3.21)), and the equivalence of the norms, leads to

$$R_T \leq \frac{\overline{L}_1\lambda^2}{2}\left(\sum_{k=1}^{T}\tilde{\rho}^{2k}\left\|z^0\right\| + 2\left\|z^0\right\| W_T + Q_T\right)$$

$$\overset{(a)}{\leq} \frac{\overline{L}_1\lambda^2}{2}\left(\frac{\left\|z^0\right\|^2}{1-\tilde{\rho}^2} + 2\left\|z^0\right\| W_T + Q_T\right), \tag{3.27}$$

where $\lambda = \lambda_1\lambda_2$ and *(a)* uses the geometric series property.

As regards the result (3.22), we use (3.13) to write

$$\begin{aligned}\mathrm{dist}(\mathrm{x}^{k+1}, X^{k+1}) &= \mathrm{dist}(\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k), X^{k+1}) \\ &\overset{(a)}{\leq} \mathrm{dist}(\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k), X^k) + \alpha^k,\end{aligned} \tag{3.28}$$

where in *(a)* we add and subtract the term $\mathrm{dist}(\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k), X^k)$ and we introduce α^k (cf. (3.8c)). Now, we recall that

$$\mathrm{dist}(\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k), X^k) = \min_{y\in X^k}\left\|\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k) - y\right\|.$$

Thus, by adding and subtracting within the norm the term $(1-\delta)v^k$ with $v^k \in X^k$ so that $\left\|\mathrm{x}^k - v^k\right\| = \mathrm{dist}(\mathrm{x}^k, X^k)$, we can use the triangle inequality and the definition of $\min$ to get

$$\begin{aligned}\mathrm{dist}(\mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k), X^k) &\leq (1-\delta)\left\|\mathrm{x}^k - v^k\right\| + \min_{y\in X^k}\left\|v^k + \delta(\tilde{\mathrm{x}}^k - v^k) - y\right\| \\ &= (1-\delta)\mathrm{dist}(\mathrm{x}^k, X^k) + \mathrm{dist}(v^k + \delta(\tilde{\mathrm{x}}^k - v^k), X^k),\end{aligned}$$

which allows us to rewrite (3.28) as

$$\mathrm{dist}(\mathrm{x}^{k+1}, X^{k+1}) \leq (1-\delta)\mathrm{dist}(\mathrm{x}^k, X^k) + \mathrm{dist}(v_t + \delta(\tilde{\mathrm{x}}^k - v^k), X^k) + \alpha^k. \tag{3.29}$$

Notice that $v^k, \tilde{\mathrm{x}}^k \in X^k$ and $0 < \delta < 1$, then $v_t + \delta(\tilde{\mathrm{x}}^k - v^k) \in X^k$ and the second term of (3.29) is null and (3.29) becomes

$$\mathrm{dist}(\mathrm{x}^{k+1}, X^{k+1}) \leq (1-\delta)\mathrm{dist}(\mathrm{x}^k, X^k) + \alpha^k. \tag{3.30}$$

Both members of (3.30) are always positive, then (3.30) leads to

$$\mathrm{dist}(\mathrm{x}^k, X^k) \leq (1-\delta)^k \mathrm{dist}(x_0, X_0) + \sum_{q=0}^{k-1} (1-\delta)^q \alpha^{k-q-1}.$$

By summing the latter for $k = 1$ up to $k = T$ and using the geometric series property the proof follows. ■

Operatively, in order to choose an appropriate value of the parameter δ, it is necessary to first estimate the upper bound $\bar{\delta}$. As it emerges from the proof of Theorem 3.1, this can be done as follows: (i) compute a matrix $M(\delta)$ (cf. (3.23)), which depends on the various problem constants and on δ, (ii) compute $\bar{\delta}$ as the maximum value of δ such that all the eigenvalues of $M(\delta)$ are strictly in the unit circle. We observe that Theorem 3.1 improves the dynamic regret bound provided in [106], which demonstrates a bound of the type $O(T) + O(\sqrt{T}V_T)$ (where V_T is a term capturing variations of the problem). The authors also show that there exists a particular, constant step-size that allows to tighten the first term to $O(\sqrt{T})$. However, the choice of the the step-size requires a prior knowledge of T and V_T. In both cases, we improve the first term, which is replaced by the constant $\frac{\overline{L}_1 \lambda^2}{2} \frac{\left\|z^0\right\|^2}{1-\tilde{\rho}^2}$, while in our terms W_T and Q_T (cf. (3.21)) the variations of the problem are scaled by $\tilde{\rho}^k$, i.e., an exponentially decaying quantity since $\tilde{\rho} \in (0,1)$. △

Remark 3.1 (Average Regret)**.** Let us consider the case in which the problem variations are bounded by a constant, i.e., suppose there exists $C > 0$ so that $\zeta^k, \eta^k, \omega^k \leq C$ for all $t \geq 0$. In this case, by using the definitions of W_T and Q_T (cf. (3.21)) and recalling that $\tilde{\rho} \in (0,1)$, we can use the geometric series property to get

$$W_T \leq \frac{(4+\overline{L}_2)C}{1-\tilde{\rho}^T}, \quad \text{and} \quad Q_T \leq \frac{(4+\overline{L}_2)^2 C^2 T}{1-\tilde{\rho}^2}.$$

In this case, the average regret approaches a constant value,

$$\lim_{T\to\infty} R_T/T = \frac{\overline{L}_1 \lambda^2 (4+\overline{L}_2)^2 C^2}{2(1-\tilde{\rho}^2)^2}.$$ △

Remark 3.2 (Inequality constraints)**.** Consider the case in which X_i^k can be expressed in terms of inequality constraints, namely

$$X_i^k := \{x_i \in \mathbb{R}^{n_i} \mid h_i^k(x_i) \leq 0_{m_i}\},$$

with $h_i^k : \mathbb{R}^{n_i} \to \mathbb{R}^{m_i}$ for all $i \in \{1, \ldots, N\}$ and $k \geq 0$. In this case, in place of the distance function $\mathrm{dist}(x, X^k)$, one can use $\|[h^k(x)]^+\|$ as a metric to characterize the constraint violation, where $h^k(x) := \mathrm{col}(h_1^k(x_1), \ldots, h_N^k(x_N))$. By repeating similar

arguments as in the proof of Theorem 3.1, one obtains similarly that $\sum_{k=1}^{T} \|[h^k(x^k)]^+\| \leq \frac{1}{1-(1-\delta)^T}\|[h^k(x_0)]^+\| + \sum_{k=1}^{T}\sum_{q=0}^{k-1}(1-\delta)^q\alpha_{k-q-1}$. △

In the following corollary, we assess that in the static case the Projected Aggregative Tracking distributed algorithm converges to the (fixed) optimal solution $x^\star$ with a linear rate.

Corollary 3.1 (Static setup)**.** *Under the same assumptions of Theorem 3.1, if it holds $f^k = f$, $\phi^k = \phi$, and $X_i^k = X_i$ for all $i \in \{1,\dots,N\}$ and all $k \geq 0$, then there exists $\lambda, \bar{\delta} > 0$ and $\tilde{\rho} \in (0,1)$ so that, if $\gamma \leq \frac{1}{\overline{L}_1}$ and $\delta \in (0, \bar{\delta})$, it holds*

$$f(\mathrm{x}^k, \sigma(\mathrm{x}^k)) - f(x^\star, \sigma(x^\star)) \leq \tilde{\rho}^{2k}\frac{\overline{L}_1\lambda^2}{2}\left\|\mathrm{z}^0\right\|^2.$$

Proof. Here, for all $k \geq 0$, it holds $u^k \equiv 0$ and $x_\star^k = x^\star$. By the same arguments of Theorem 3.1, we use the Lipschitz continuity of ∇f^k (cf. Assumption 3.3), (3.25) with $u^k \equiv 0$, and the equivalence of the norms to get $f(\mathrm{x}^k, \sigma(\mathrm{x}^k)) - f(x^\star, \sigma(x^\star)) \leq \tilde{\rho}^{2k}\frac{\overline{L}_1\lambda_1^2}{2}\left\|\mathrm{z}^0\right\|_\iota^2$. The proof follows by using again the equivalence of the norms and setting $\lambda := \lambda_1\lambda_2$. ■

3.2.2 Numerical Simulations

In this section we show the effectiveness of Projected Aggregative Tracking on an online version of the multi-robot surveillance scenario already presented in Section 1.3.2.

Online setup

Let us consider a network of cooperating robots that aim to protect a target with location $b^k \in \mathbb{R}^2$ at iteration k from some intruders. The optimization variables $\mathrm{x}_i^k \in \mathbb{R}^2$ represent the position of robots at each iteration k and each robot i is able to move from x_i^k to x_i^{k+1} using a local controller. We associate to each robot i an intruder located at $p_i^k \in \mathbb{R}^2$ at iteration k. The dynamic protection strategy applied by each robot consists of staying simultaneously close to the protected target and to the associated intruder. Meanwhile, the whole team of robots tries to keep its weighted center of mass rotating close to the target. A concept of this scenario is given in Figure (3.1).

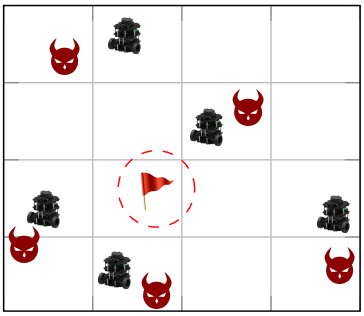


Figure 3.1: Multi- robot surveillance scenario - Robot icons denote agents, devil icons denote intruders, while the flag is the target to be protected.

This strategy is obtained by solving problem (3.3) with the cost functions $f_i^k(x_i, \sigma^k(x)) = \frac{1}{2}\left\|x_i - p_i^k\right\|^2 + \frac{\alpha_1}{2}\left\|x_i - b^k\right\|^2 + \frac{\alpha_2}{2N}\left\|\sigma^k(x) - b^k\right\|^2$, with $\alpha_1 = 1$, $\alpha_2 = 10$ and the aggregation rules $\phi_i^k(x_i) = \beta_i x_i + a^k$, where $\beta_i > 0$ and $a_t \in \mathbb{R}^2$ represents a time-varying offset which follows the law $a^k = r\text{col}(\cos(k/(2\pi\tau)), \sin(k/(2\pi\tau)))$ for some $r, \tau > 0$. In this way, the center of mass $\frac{1}{N}\sum_{i=1}^N x_i^k$ is forced to rotate around the target position b_t.

We address a scenario with $N = 50$ agents and intruders. As regards the constraints, we consider a common time-varying box $X_i^k = \{x \in \mathbb{R}^2 \mid 0 \leq x \leq x_{\max}^k\}$ for all i, where $x_{\max}^k \in \mathbb{R}^2$ starts from $[20, 20]$ and linearly increases at each iteration. In this way, the agents initially stay closer to the target and then they move toward the associated intruders. Each intruder i moves along a circle of radius $r = 1$ according to the law $p_i^k = p_{i,c} + r\text{col}(\cos(k/100), \sin(k/100))$, where $p_{i,c} \in \mathbb{R}^2$ is randomly generated. The target b^k and the offset a^k follow similar laws. In this setup, being the sinusoidal functions bounded, the constants η^k and ω^k introduced in (3.8) can be uniformly bounded as $\eta^k \leq \alpha_2\sqrt{N}r$ and $\omega^k \leq \sqrt{N}r$ for all $k \geq 0$. Moreover, the vector $x_{\max}^k$ defining the box X_i^k changes linearly with respect time and, thus, also the constant α^k (cf. (3.8c)) can be uniformly bounded. As regards the algorithm parameters, we set $\gamma = 1$ and $\delta = 0.5$. We performed 100 Monte Carlo trials that differ in the problem parameters and agents' initial conditions. Figure 3.2 shows that the behavior of the algorithm does not depend on the generated instances. Indeed, the achieved average dynamic regret, as predicted in Remark 3.1, converges asymptotically to a constant.

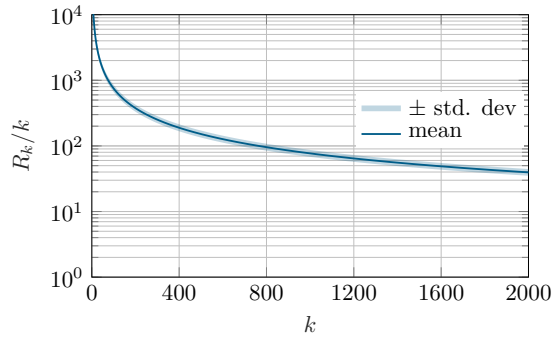


Figure 3.2: Online case – Mean of the average Dynamic regret and 1-standard deviation band over 100 Monte Carlo trials.

Static setup

Now we address a static instance of the problem. Namely, we fix X^k and the positions of the intruders and of the target. We perform a Monte Carlo simulation consisting of 100 trials on the same network of $N = 50$ agents with the same algorithm parameters. As predicted by Corollary 3.1, Figure 3.3 shows an exponential decay of $\frac{\|\mathbf{x}^k - x^\star\|}{\|x^\star\|}$.

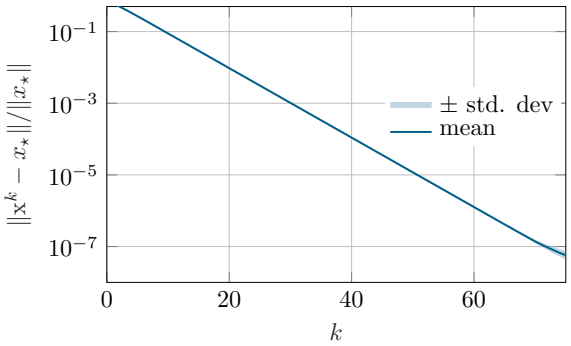


Figure 3.3: Static case – Mean of the relative error and 1-standard deviation band obtained with 100 Monte Carlo trials.

3.3 Distributed Personalized Aggregative Optimization

In this section, we consider "personalized" instances of problem (3.3), i..e, an optimization framework described by

$$\min_{(x_1,\dots,x_N)\in X} \sum_{i=1}^{N} \underbrace{V_i^k(x_i, \sigma^k(x)) + U_i(x_i, \sigma^k(x))}_{f_i^k(x_i,\sigma^k(x))}, \tag{3.31}$$

in which $x := \text{col}(x_1, \dots, x_N) \in \mathbb{R}^n$ is the global decision vector, with each $x_i \in \mathbb{R}^{n_i}$ and $n = \sum_{i=1}^N n_i$. Each agent i is equipped with the known time-varying engineering cost

$V_i^k : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}$ and with the unknown user's dissatisfaction function $U_i : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}$. For each agent i, these two contributions define the local cost function

$$f_i^k(\cdot, \sigma^k(\cdot)) := V_i^k(\cdot, \sigma^k(\cdot)) + U_i(\cdot, \sigma^k(\cdot)).$$

The global decision vector is constrained to belong to the set $X \subseteq \mathbb{R}^n$ given by $X := X_1 \times \cdots \times X_N$, with each $X_i \subseteq \mathbb{R}^{n_i}$. The *aggregation function* $\sigma^k(x)$ still has the form (3.2). As in Section 3.2, we denote the cost function of problem (3.31) more compactly as $f^k(x, \sigma^k(x))$, where $f^k : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is defined as $f^k(x, w) := \sum_{i=1}^N f_i^k(x_i, w)$ with $\mathrm{x}_i \in \mathbb{R}^{n_i}$, $w \in \mathbb{R}^d$. The same reasoning applies also for the functions $V^k(\mathrm{x}, w) := \sum_{i=1}^N V_i^k(x_i, w)$ and $U(x, w) := \sum_{i=1}^N U_i(x_i, w)$. Further, given $x := \mathrm{col}(x_1, \ldots, x_N) \in \mathbb{R}^n$ and $w := \mathrm{col}(w_1, \ldots, w_N) \in \mathbb{R}^{Nd}$ with $x_i \in \mathbb{R}^{n_i}$, $w_i \in \mathbb{R}^d$ for all $\in \{1, \ldots, N\}$, we will use the symbols

$$\phi^k(x) := \begin{bmatrix} \phi_1^k(x_1) \\ \ldots \\ \phi_N^k(x_N) \end{bmatrix}, G_{1,V}^k(x, w) := \begin{bmatrix} \nabla_1 V_1^k(x_1, w_1) \\ \ldots \\ \nabla_1 V_N^k(x_N, w_N) \end{bmatrix}, G_{2,V}^k(x, w) := \begin{bmatrix} \nabla_2 V_1^k(x_1, w_1) \\ \ldots \\ \nabla_2 V_N^k(x_N, y_N) \end{bmatrix},$$

and

$$G_V^k(x, w) := G_{1,V}^k(x, w) + \nabla \phi^k(x) \mathbf{1}_{N,d} \frac{1}{N} \sum_{i=1}^N \nabla_2 V_i^k(x_i, w_i).$$

The idea is to solve problem (3.31) in a distributed way over a network of N agents communicating according to an undirected graph $\mathcal{G} := (\{1, \ldots, N\}, \mathcal{E})$. The features of problem (3.31) are summarized within the next assumptions.

Assumption 3.4 (Unknown Function)**.** *The function U_i has a quadratic structure*

$$U_i(x_i, s_i) = \begin{bmatrix} x_i^\top s_i^\top \end{bmatrix} P_i \begin{bmatrix} x_i \\ s_i \end{bmatrix} + q_i^\top \begin{bmatrix} x_i \\ s_i \end{bmatrix} + r_i, \tag{3.32}$$

where $x_i \in \mathbb{R}^{n_i}$, $s_i \in \mathbb{R}^d$, $P_i = P_i^\top \in \mathbb{R}^{(n_i+d)\times(n_i+d)}$ has eigenvalues in $[\mu_U, \overline{L}_U]$ with $\overline{L}_U > \mu_U > 0$, $q_i \in \mathbb{R}^{n_i+d}$, and $r_i \in \mathbb{R}$ for all $i \in \{1, \ldots, N\}$. The parameters of each function are unknown, but each agent i can access the noisy measurement $\mathrm{z}_i \in \mathbb{R}$ defined as

$$\mathrm{z}_i = U_i(x_i, s_i) + \epsilon_i,$$

with ϵ_i a generic scalar zero-mean noise with finite variance.

Assumption 3.5 (Set)**.** *The set X^k is closed and convex.* △

Assumption 3.6 (Engineering Function)**.** *The global engineering function $V^k(x, \sigma^k(x))$ is μ-strongly convex for all $k \geq 0$ and $\left\|x_\star^k\right\|$ is bounded, where $x_\star^k \in \mathbb{R}^n$ denotes its minimizer*

at time $k \geq 0$ over the set X. Further, it is differentiable with $\overline{L}_{1,V}$-Lipschitz continuous gradient. Moreover, the function G_V^k s $\overline{L}_{1,V}$-Lipschitz, namely it holds

$$\left\| G_V^k(x,w) - G_V^k(x',w') \right\| \leq \overline{L}_{1,V} \left\| \text{col}(x-x', w-w') \right\|,$$

for all $x, x' \in \mathbb{R}^n$ and $w, w' \in \mathbb{R}^{Nd}$ and $k \geq 0$. In addition, the functions $G_{2,V}^k(x,y)$ and $\phi_i^k(\cdot)$ are $\overline{L}_{2,V}$-Lipschitz and $\overline{L}_3$-continuous, respectively, for all $k \geq 0$ and $i \in \{1,\dots,N\}$. △

Assumption 3.7 (Communication graph)**.** *$\mathcal{G}$ is connected. Moreover, the adjacency matrix $\mathcal{W}_\mathcal{G}$ is doubly stochastic.* △

Assumptions 3.4 and 3.6 imply that, for all $k \geq 0$, the gradients ∇f^k and $\nabla_1 f^k$ are Lipschitz continuous functions with parameter $\overline{L}_U + \overline{L}_{1,V}$, while the Lipschitz parameter of $G_2^k(x,y) := \text{col}(\nabla_2 f_1^k(x_1,y_1), \dots, \nabla_2 f_N^k(x_N,y_N))$ is $\overline{L}_U + \overline{L}_{2,V}$.

We assume that each agent i can only privately access V_i^k, ϕ_i^k, X_i, and a noisy user feedback $\mathrm{z}_i^k = U_i(\mathrm{x}_i^k, \mathrm{s}_i^k) + \epsilon_i^k$ where x_i^k and s_i^k are its local estimates of the solution and the aggregative variable at iteration k, respectively, and ϵ_i^k is a noise term. It is important to remark that each agent i accesses its private information only once the estimate x_i^k of the i-th component of the minimizer of problem (3.31) has been computed. We denote the minimizer of $f^k(x, \sigma^k(x))$ over X as $x_\star^k$ (which is unique in light of Assumption 3.5, cf. Proposition A.2 in Appendix A). We will present RLS Projected Aggregative Tracking, i.e., a distributed scheme to address problem (3.31) and, as in Section 3.2, we will evaluate its performance in terms of dynamic regret that recall as follows

$$R_T := \sum_{k=1}^{T} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=1}^{T} f^k(x_\star^k, \sigma^k(x_\star^k)), \tag{3.33}$$

given a finite value $T > 1$, Next, we provide a preliminary section to the present Recursive Least Squares scheme, used within the learning part of RLS Projected Aggregative Tracking.

Recursive Least Squares

The distributed algorithm proposed in this paper relies on a learning part driven by users' feedback. Specifically, at each $q \geq 0$, let agent i have some local states $\mathrm{x}_i^q \in \mathbb{R}^{n_i}$ and $\mathrm{s}_i^q \in \mathbb{R}^d$, and suppose it can measure

$$z_i^q = U_i(\mathrm{x}_i^q, \mathrm{s}_i^q) + \epsilon_i^q \tag{3.34}$$

to learn the parameters of the unknown function U_i by using a Least Squares (LS) method. That is, it computes

$$\hat{\xi}_i^k := \arg\min_{\xi_i} \sum_{q=1}^{k} \left(\xi_i^\top \chi_i^q - \mathrm{z}_i^q\right)^2, \tag{3.35}$$

in which, for all $q \in \{1, \dots, k\}$, the regressor vector $\chi_i^q \in \mathbb{R}^{n_{\mathrm{ls}}}$, $n_{\mathrm{ls}} := 1 + n_i + d + (n_i + d)^2$, is given by

$$\chi_i^q := \mathrm{col}\left(1, \begin{bmatrix}\mathrm{x}_i^q \\ \mathrm{s}_i^q\end{bmatrix}, \frac{1}{2}\begin{bmatrix}\mathrm{x}_i^q \\ \mathrm{s}_i^q\end{bmatrix}_1 \begin{bmatrix}\mathrm{x}_i^q \\ \mathrm{s}_i^q\end{bmatrix}, \dots, \frac{1}{2}\begin{bmatrix}\mathrm{x}_i^q \\ \mathrm{s}_i^q\end{bmatrix}_{n_i+d} \begin{bmatrix}\mathrm{x}_i^q \\ \mathrm{s}_i^q\end{bmatrix}\right). \tag{3.36}$$

The asymptotic properties of the LS scheme are summarized in the next lemma, which requires the following assumption on persistent excitation of data.

Assumption 3.8 (Persistent Excitation). *For any $i \in \{1, \dots, N\}$ and $k \geq 0$, let the sequence $\{\mathrm{x}_i^q, \mathrm{s}_i^q\}_{0 \leq q \leq k}$ be such that*

- *$\{(\chi_i^q, \mathrm{z}_i^q)\}_{0 \leq q \leq k}$ in (3.34) and (3.36) is a realization of a jointly stationary ergodic process;*
- *the matrix $\Sigma := \mathbb{E}[\chi_i^q (\chi_i^q)^\top]$ is non-singular;*
- *the sequence $\{\chi_i^q \epsilon_i^q\}_{0 \leq q \leq k}$ is a martingale difference sequence with finite second moments (cf. [82, Assumption 2.5]).* △

Lemma 3.5 (Estimation error). *Let Assumption 3.8 holds and, for any $i \in \{1, \dots, N\}$, let $\xi_i^\star := \mathrm{col}(1, q_i, [P_i]_1, \dots, [P_i]_{n_{ls}})$. Then, denoting $S := \mathbb{E}\left[\chi_i^q \xi_i^q (\chi_i^q \xi_i^q)^\top\right]$, it holds*

$$\lim_{k \to \infty} \sqrt{k}(\hat{\xi}_i^k - \xi_{i,\star}^k) \xrightarrow{D} \mathcal{N}(0, \Sigma^{-1} S \Sigma^{-1}), \tag{3.37}$$

where the notation $\xrightarrow{D}$ stands for convergence in distribution. Moreover, the estimated $\hat{U}_i^k(x, s)$ is bounded for any finite x and s, for all $i \in \{1, \dots, N\}$, and $k \geq 0$. Further, for any $\nu \in (0, 1]$ and $\epsilon > 1$, there exists a finite $\bar{k} \geq 0$, for which the estimated $\hat{P}_i^k$ is symmetric and it has eigenvalues in the set $[0, \epsilon \overline{L}_U]$ with probability $1 - \nu$, i.e., it holds

$$\epsilon \overline{L}_U I_{n+Nd} \geq \hat{P}_i^k = (\hat{P}_i^k)^\top \geq 0. \tag{3.38}$$

The result (3.37) can be derived from, e.g., [117, Chapters 8, 9, 11] and [82, Proposition 2.1]. As regards the result (3.38), see [141, Appendix A.3].

Recursive Least Squares (RLS) (see, e.g., [117, Chapter 11]) is an efficient scheme to iteratively solve (3.35) as soon as new data arrive. Asymptotic properties of RLS

coincide with the ones of non-recursive LS, thus in our scheme we will use RLS and require Assumption 3.8.

Remark 3.3. Among the possibilities, we exploited an RLS scheme to estimate a quadratic unknown function. However, especially in the case of more general functions instead of the quadratic ones, other strategies may be investigated such as the exploitation of, e.g., Gaussian processes [145, 146, 172] or neural networks [46]. △

3.3.1 RLS Projected Aggregative Tracking: Algorithm Description and Analysis

In this section, we present RLS Projected Aggregative Tracking, namely a distributed algorithm extending Algorithm 5 to address problem (3.31). RLS Projected Aggregative Tracking includes a Recursive Least Squares mechanism (see, e.g., [82, 117, 163]) driven by noisy user feedback z_i^k and providing estimates $\hat{f}_i^k$ of the local objective function f_i^k.

Each agent i, at each iteration $k \geq 0$, maintains an estimate $\mathrm{x}_i^k \in \mathbb{R}^{n_i}$ of the component i of the minimizer $x_\star^k$ of problem (3.31), and two auxiliary variables $\mathrm{s}_i^k, \mathrm{y}_i^k \in \mathbb{R}^d$. The estimate x_i^k is updated by manipulating the learned $\hat{f}_i^k$ and the variables s_i^k and y_i^k. A pseudo-code of the RLS Projected Aggregative Tracking algorithm is reported in Algorithm 6 from the perspective of agent i, in which $r_i^0 > 0$ is an initialization parameter, $\gamma > 0$ is a step-size, $\delta \in (0, 1)$ represents a convex combination constant, and each element w_{ij} represents the (i, j)-entry of the weighted adjacency matrix $\mathcal{W}_\mathcal{G}$ of the network communication graph.

Algorithm 6 RLS Projected Aggregative Tracking (Agent i)

Initialization: set $R_i^0 = r_i^0 I_{n_{\text{ls}}}$, $\hat{\xi}_i^0 = 0$, $\mathrm{x}_i^0 \in X_i$, $s_i^0 = \phi_i^0(x_i^0)$, $\mathrm{y}_i^0 = \nabla_2 \hat{f}_i^0(x_i^0, s_i^0)$

for $k = 0, 1, \dots$ **do**

Optimization

$$\begin{aligned}
\tilde{\mathrm{x}}_i^k &= P_{X_i}\left[\mathrm{x}_i^k - \gamma(\nabla_1 \hat{f}_i^k(\mathrm{x}_i^k, \mathrm{s}_i^k) + \nabla \phi_i^k(\mathrm{x}_i^k)\mathrm{y}_i^k)\right] \\
\mathrm{x}_i^{k+1} &= \mathrm{x}_i^k + \delta(\tilde{\mathrm{x}}_i^k - \mathrm{x}_i^k) \\
\mathrm{s}_i^{k+1} &= \sum_{j=1}^N w_{ij}\mathrm{s}_j^k + \phi_i^{k+1}(\mathrm{x}_i^{k+1}) - \phi_i^k(\mathrm{x}_i^k)
\end{aligned}$$

Measurement

$\mathrm{z}_i^{k+1} = U_i(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}) + \epsilon_i^{k+1}$

Learning

$$\begin{aligned}
\psi_i^{k+1} &= \frac{R_i^k \chi_i^{k+1}}{1 + (\chi_i^{k+1})^\top R_i^k \chi_i^{k+1}} \\
R_i^{k+1} &= R_i^k - (1 + (\chi_i^{k+1})^\top R_i^k \chi_i^{k+1})\psi_i^{k+1}(\psi_i^{k+1})^\top \\
\hat{\xi}_i^{k+1} &= \hat{\xi}_i^k + (\mathrm{z}_i^{k+1} - (\chi_i^{k+1})^\top \hat{\xi}_i^{k+1})\psi_i^{k+1} \\
(\hat{P}_i^{k+1}, \hat{q}_i^{k+1}, \hat{r}_{i,k+1}) &= \textsc{unpack}(\hat{\xi}_i^{k+1}) \\
\hat{P}_i^{k+1} &\leftarrow (\hat{P}_i^{k+1} + (\hat{P}_i^{k+1})^\top)/2
\end{aligned}$$

$$\mathrm{y}_i^{k+1} = \sum_{j=1}^N w_{ij}\mathrm{y}_j^k + \nabla_2 \hat{f}_i^{k+1}(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}) - \nabla_2 \hat{f}_i^k(\mathrm{x}_i^k, \mathrm{s}_i^k)$$

end for

In Algorithm 6, in order to move x_i^k toward the minimizer of problem (3.31), as in Algorithm 5, each agent i employs two trackers to reconstruct the unavailable global quantities $\nabla_1 f_i^k\left(\mathrm{x}_i^k, \sum_{j=1}^N \frac{\phi_j^k(\mathrm{x}_j^k)}{N}\right)$ and $\frac{1}{N}\nabla \phi_i^k(\mathrm{x}_i^k)\sum_{i=1}^N \nabla_2 f_i^k\left(\mathrm{x}_i^k, \sum_{j=1}^N \frac{\phi_j^k(\mathrm{x}_j^k)}{N}\right)$. Moreover, as for the not completely known local function f_i^k, here each agent i needs to replace it by the estimate $\hat{f}_i^k$ provided by the learning part of the scheme. Indeed, in Algorithm 6, the optimization steps (inspired by Algorithm 5) are interlaced with a RLS performed by using the measurements z_i^{k+1}. By relying on this outcome, agent i can manipulate the updated estimated cost function $\hat{f}_i^{k+1}$. In fact, once x_i^{k+1}, s_i^{k+1}, and $\hat{\xi}_i^{k+1}$

have been computed, it can access

$$\hat{U}_i^{k+1}(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}) := \frac{1}{2}\begin{bmatrix}(\mathrm{x}_i^{k+1})^\top & (\mathrm{s}_i^{k+1})^\top\end{bmatrix} \hat{P}_i^{k+1} \begin{bmatrix}\mathrm{x}_i^{k+1} \\ \mathrm{s}_i^{k+1}\end{bmatrix} + (\hat{q}_i^{k+1})^\top \begin{bmatrix}\mathrm{x}_i^{k+1} \\ \mathrm{s}_i^{k+1}\end{bmatrix} + \hat{r}_i^{k+1}, \tag{3.39}$$

where the estimates $\hat{P}_i^{k+1}$, $\hat{q}_i^{k+1}$, and $\hat{r}_i^{k+1}$ are extracted from the vector $\hat{\xi}_i^{k+1}$ through the unpack operator (see Notation paragraph). The estimate (3.39) can be combined with the known part V_i^{k+1} to get the whole local function estimate $\hat{f}_i^{k+1}$ as

$$\hat{f}_i^{k+1}(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}) := V_i^{k+1}(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}) + \hat{U}_i^{k+1}(\mathrm{x}_i^{k+1}, \mathrm{s}_i^{k+1}). \tag{3.40}$$

Such an estimate is used to compute the update directions used in the optimization steps.

We specify that, in Algorithm 6, $\nabla_1 \hat{f}_i^k$ and $\nabla_2 \hat{f}_i^k$ denote the gradients of $\hat{f}_i^k$ computed by the agent i using its currently available estimates of the U_i parameters, namely $(\hat{P}_i^k, \hat{q}_i^k, \hat{r}_i^k)$.

In order to derive an upper bound for the dynamic regret that RLS Projected Aggregative Tracking can achieve, let us rewrite all the agents' updates of Algorithm 6 in a stacked vector form as

$$\tilde{\mathrm{x}}^k = P_X[\mathrm{x}^k - \gamma(\hat{G}_1^k(\mathrm{x}^k, \mathrm{s}^k) + \nabla\phi^k(\mathrm{x}^k)\mathrm{y}^k)] \tag{3.41a}$$

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta(\tilde{\mathrm{x}}^k - \mathrm{x}^k) \tag{3.41b}$$

$$\mathrm{s}^{k+1} = \mathcal{W}\mathrm{s}^k + \phi^{k+1}(\mathrm{x}^{k+1}) - \phi^k(\mathrm{x}^k) \tag{3.41c}$$

$$\mathrm{y}^{k+1} = \mathcal{W}\mathrm{y}^k + \hat{G}_2^{k+1}(\mathrm{x}^{k+1}, \mathrm{s}^{k+1}) - \hat{G}_2^k(\mathrm{x}^k, \mathrm{s}^k), \tag{3.41d}$$

in which we have collected all the local quantities through the symbols x^k, s^k, y^k, $\phi(\mathrm{x}^k)$ with same meaning as in (3.6) and

$$\hat{G}_1^k(\mathrm{x}^k, \mathrm{s}^k) := \begin{bmatrix}\nabla_1 \hat{f}_1^k(\mathrm{x}_1^k, \mathrm{s}_1^k) \\ \vdots \\ \nabla_1 \hat{f}_N^k(\mathrm{x}_N^k, \mathrm{s}_N^k)\end{bmatrix}, \quad \hat{G}_2^k(\mathrm{x}^k, \mathrm{s}^k) := \begin{bmatrix}\nabla_2 \hat{f}_1^k(\mathrm{x}_1^k, \mathrm{s}_1^k) \\ \vdots \\ \nabla_2 \hat{f}_N^k(\mathrm{x}_N^k, \mathrm{s}_N^k)\end{bmatrix}.$$

We prove the performance properties of RLS Projected Aggregative Tracking by properly defining an error vector and using Lemmas 3.1, 3.2, 3.3, and 3.4 suitably adapted for the setting of this section. Indeed, we notice that, according to the result (3.38), for any $\nu \in (0, 1]$, there exists $\bar{k} \geq 0$ such that, for any $k \geq \bar{k}$, with probability $1 - \nu$, the function $\hat{f}^k(\cdot, \sigma_k(\cdot)) := \sum_{i=1}^N \hat{f}_i^k(\cdot, \sigma^k(\cdot))$ is μ-strongly convex. Hence, for any $k \geq \bar{k}$, with probability $1 - \nu$, the cost $\hat{f}^k(\cdot, \sigma_k(\cdot))$ over X has a unique minimizer $\hat{x}_\star^k \in \mathbb{R}^n$ (cf. Proposition A.2 in Appendix A). Further, $\nabla \hat{f}^k$ and $\hat{G}_1^k$ are $\overline{L}_1$-Lipschitz continuous with

$\overline{L}_1 := \epsilon\overline{L}_U + \overline{L}_{1,V}$, and $\hat{G}_2^k$ is $\overline{L}_2$-Lipschitz continuous with $\overline{L}_2 := \epsilon\overline{L}_U + \overline{L}_{2,V}$. Thus, by suitably replacing f^k and $x_\star^k$ with the estimated $\hat{f}^k$ and its minimizer $\hat{x}_\star^k$ over X, we can state the useful bounds relying on the same arguments of Lemmas 3.1, 3.2, 3.3, and 3.4. To this end, we introduce four useful variables defined as

$$\beta^k := \sup_i \sup_{x\in X_i, z\in\mathbb{R}^d} \left\| U_i(x,z) - \hat{U}_i^k(x,z) \right\| \tag{3.42a}$$

$$\eta^k := \sup_{x\in X, z\in\mathbb{R}^{Nd}} \left\| \hat{G}_2^{k+1}(x,z) - \hat{G}_2^k(x,z) \right\| \tag{3.42b}$$

$$\omega^k := \sup_{x\in X} \left\| \phi^{k+1}(x) - \phi^k(x) \right\| \tag{3.42c}$$

$$\zeta^k := \left\| \hat{x}_\star^{k+1} - \hat{x}_\star^k \right\|. \tag{3.42d}$$

Lemma 3.6. *Let Assumptions 3.4, 3.5, 3.6, 3.7, and 3.8 hold. If $\gamma \leq \frac{1}{\overline{L}_1}$, then, for any $\nu \in (0,1]$, $\exists \bar{k} \geq 0$ such that, for all $k \geq \bar{k}$, it holds*

$$\left\| \mathrm{x}^{k+1} - \hat{x}_\star^{k+1} \right\| \leq (1-\delta\mu\gamma) \left\| \mathrm{x}^k - \hat{x}_\star^k \right\| + \delta\gamma\overline{L}_1 \left\| \mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k \right\| + \delta\gamma\overline{L}_3 \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \zeta^k.$$

Lemma 3.7. *Let Assumptions 3.4, 3.5, 3.6, 3.7, and 3.8 hold. Then, for any $\nu \in (0,1]$, $\exists \bar{k} \geq 0$ such that, for all $k \geq \bar{k}$, it holds*

$$\left\| \mathrm{x}^{k+1} - \mathrm{x}^k \right\| \leq \delta(2 + \gamma\overline{L}_1 + \gamma L_1 L_3) \left\| \mathrm{x}^k - \hat{x}_\star^k \right\| + \delta\gamma\overline{L}_1 \left\| \mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k \right\| + \delta\gamma\overline{L}_3 \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\|.$$

Lemma 3.8. *Let Assumptions 3.4, 3.5, 3.6, 3.7, and 3.8 hold. Then, for any $\nu \in (0,1]$, $\exists \bar{k} \geq 0$ such that, for all $k \geq \bar{k}$, it holds*

$$\begin{aligned}\left\| \mathrm{s}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{s}}^{k+1} \right\| \leq{}& \Lambda \left\| \mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k \right\| + \delta\gamma L_1 L_3 \left\| \mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k \right\| \\ &+ \delta(2\overline{L}_3 + \gamma L_1 L_3 + \gamma L_1 L_3^2) \left\| \mathrm{x}^k - \hat{x}_\star^k \right\| \\ &+ \delta\gamma\overline{L}_3^2 \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \omega^k,\end{aligned}$$

where Λ is the maximum eigenvalue of the matrix $\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$.

Lemma 3.9. *Let Assumptions 3.4, 3.5, 3.6, 3.7, and 3.8 hold. Then, for any $\nu \in (0,1]$, $\exists \bar{k} \geq 0$ such that, for all $k \geq \bar{k}$, it holds*

$$\begin{aligned}\left\| \mathrm{y}^{k+1} - \mathbf{1}_{N,d}\bar{\mathrm{y}}^{k+1} \right\| \leq{}& \Lambda \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| + \delta\gamma\overline{L}_3(\overline{L}_2 + L_2 L_3) \left\| \mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k \right\| \\ &+ \delta(2 + \gamma\overline{L}_1 + \gamma\overline{L}_1)(\overline{L}_2 + L_2 L_3) \left\| \mathrm{x}^k - \hat{x}_\star^k \right\| + 2L_2 \left\| \mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k \right\| \\ &+ \delta\gamma\overline{L}_1(\overline{L}_2 + L_2 L_3) \left\| \mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k \right\| + \overline{L}_2\sqrt{N}\omega^k + \eta^k,\end{aligned}$$

where Λ is the maximum eigenvalue of the matrix $\mathcal{W} - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}$.

Now, we are ready to state the main theoretical results of the paper. Indeed, the next theorem provides a bound on the dynamic regret of the estimates generated by the RLS Projected Aggregative Tracking. To this end, we will use Lemma 3.6, 3.7, 3.8, and 3.9 to bound the components of the vector e^k defined as

$$\mathrm{e}^k := \begin{bmatrix} \left\|\mathrm{x}^k - \hat{x}_\star^k\right\| \\ \left\|\mathrm{s}^k - \mathbf{1}_{N,d}\bar{\mathrm{s}}^k\right\| \\ \left\|\mathrm{y}^k - \mathbf{1}_{N,d}\bar{\mathrm{y}}^k\right\| \end{bmatrix}. \tag{3.43}$$

Theorem 3.2. *Consider RLS Projected Aggregative Tracking as given in Algorithm 6. Let Assumptions 3.4, 3.5, 3.6, 3.7, and 3.8 hold and assume that η^k and ω^k are finite for any $k \geq 0$. If $\gamma \leq \frac{1}{\overline{L}_1}$, then, for any $\nu \in (0,1]$, there exist $C, \lambda, \bar{k} > 0$, and $\bar{\delta} \in (0,1)$ such that, for any $\delta \in (0,\bar{\delta})$, there exists $\tilde{\rho} \in (0,1)$ such that, with probability $1-\nu$, it holds*

$$R_T \leq \frac{\overline{L}_1\lambda^2}{2}\left(\frac{\left\|\mathrm{e}^{\bar{k}}\right\|^2}{1-\tilde{\rho}^2} + 2\left\|\mathrm{e}^{\bar{k}}\right\| W_T + Q_T\right) + \mathcal{B}_T, \tag{3.44}$$

where R_T has been defined in (3.5) *and*

$$W_T := \sum_{k=\bar{k}}^{T}\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{k+k}\left(\zeta^{k-q-1} + 2\eta^{k-q-1} + (1+\overline{L}_2)\omega^{k-q-1}\right) \tag{3.45a}$$

$$Q_T := \sum_{k=\bar{k}}^{T}\left(\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{q}\left(\zeta^{k-q-1} + 2\eta^{k-q-1} + (1+\overline{L}_2)\omega^{k-q-1}\right)\right)^2 \tag{3.45b}$$

$$\mathcal{B}_T := 2N\sum_{k=\bar{k}}^{T}\beta^k + C\bar{k}, \tag{3.45c}$$

where ζ^k, η^k, ω^k, and β^k are defined in (3.42).

Proof. The first steps of the proof of Theorem 3.2 mimics the ones of the proof of Theorem 3.1. Indeed, by relying on Lemma 3.6, 3.7, 3.8, and 3.9, for any $\nu \in (0,1)$ there exists $\bar{k} > 0$ so that, for any $k \geq \bar{k}$, the evolution of e^k (cf. (3.43)) is governed by the same system given in (3.23), namely

$$\mathrm{e}^{k+1} \leq M(\delta)\mathrm{e}^k + Bu^k,$$

where $u^k := \mathrm{col}(\zeta^k, \eta^k, \omega^k)$ (cf. (3.42)) is the input variable, and the matrices $M(\delta) \in \mathbb{R}^{3\times 3}$, $B \in \mathbb{R}^{3\times 3}$ have the same meaning as in (3.23). Thus, as we already proved in the proof of Theorem 3.1, there exist $\bar{\delta} > 0$, $\tilde{\rho} \in (0,1)$, and a norm $\|\cdot\|_\nu$ so that, for any

$\delta \in (0, \bar{\delta})$, it holds

$$\left\|\mathrm{e}^k\right\|_\iota \leq \tilde{\rho}^k \left\|\mathrm{e}^0\right\|_\iota + \sum_{q=0}^{k-1} \tilde{\rho}^q \left\|Bu^{k-q-1}\right\|_\iota . \tag{3.46}$$

Keep this result in mind and recall the definition of dynamic regret given in (3.33) to write

$$\begin{aligned} R_T &= \sum_{k=1}^{T} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=1}^{T} f^k(x_\star^k, \sigma^k(x_\star^k)) \\ &\overset{(a)}{=} \sum_{k=1}^{\bar{k}-1} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=1}^{\bar{k}-1} f^k(x_\star^k, \sigma^k(x_\star^k)) \\ &\quad + \sum_{k=\bar{k}}^{T} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=\bar{k}}^{T} f^k(x_\star^k, \sigma^k(x_\star^k)), \end{aligned} \tag{3.47}$$

where in (a) we isolate the terms of the sum until $k = \bar{k}$. We notice that, in light of Assumptions 3.4 and 3.6, the function f^k is bounded in the case of bounded arguments. Thus, we can introduce some positive constant $C > 0$ to write

$$\sum_{k=1}^{\bar{k}-1} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=1}^{\bar{k}-1} f^k(x_\star^k, \sigma^k(x_\star^k)) \leq C\bar{k},$$

which allows us to upper bound (3.47) as

$$\begin{aligned} R_T &= C\bar{k} + \sum_{k=\bar{k}}^{T} f^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=\bar{k}}^{T} f^k(x_\star^k, \sigma^k(x_\star^k)) \\ &\overset{(a)}{\leq} C\bar{k} + \sum_{k=\bar{k}}^{T} \hat{f}^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=\bar{k}}^{T} f^k(x_\star^k, \sigma^k(x_\star^k)) \\ &\quad + \sum_{k=\bar{k}}^{T} \sum_{i=1}^{N} \left(U_i(\mathrm{x}_i^k, \sigma^k(\mathrm{x}_i^k)) - \hat{U}_i^k(\mathrm{x}_i^k, \sigma^k(\mathrm{x}_i^k)) \right), \end{aligned} \tag{3.48}$$

where (a) uses the definition of $\hat{f}_i^k$ given in (3.40) to keep apart the learning errors terms $\sum_{i=1}^{N}(U_i(\mathrm{x}_i^k, \sigma^k(\mathrm{x}_i^k)) - \hat{U}_i^k(\mathrm{x}_i^k, \sigma^k(\mathrm{x}_i^k)))$. Next, we handle $f^k(x_\star^k, \sigma^k(x_\star^k))$. Indeed, by adding and subtracting $f^k(\hat{x}_\star^k, \sigma^k(\hat{x}_\star^k))$ and $\hat{f}^k(x_\star^k, \sigma^k(x_\star^k))$, we get

$$\begin{aligned} f^k(x_\star^k, \sigma^k(x_\star^k)) &= \hat{f}^k(\hat{x}_\star^k, \sigma^k(\hat{x}_\star^k)) + \hat{f}^k(x_\star^k, \sigma^k(x_\star^k)) - \hat{f}^k(\hat{x}_\star^k, \sigma^k(\hat{x}_\star^k)) \\ &\quad + f^k(x_\star^k, \sigma^k(x_\star^k)) - \hat{f}^k(x_\star^k, \sigma^k(x_\star^k)). \end{aligned} \tag{3.49}$$

Since $\hat{x}_\star^k$ is the unique minimizer of $\hat{f}^k$, then

$$\hat{f}^k(x_\star^k, \sigma^k(x_\star^k)) - \hat{f}^k(\hat{x}_\star^k, \sigma^k(\hat{x}_\star^k)) > 0. \tag{3.50}$$

Moreover, the definitions of f^k and $\hat{f}^k$ lead to

$$f^k(x_\star^k, \sigma^k(x_\star^k)) - \hat{f}^k(x_\star^k, \sigma^k(x_\star^k)) = \sum_{i=1}^N \left(U_i(x_\star^k, \sigma^k(x_\star^k)) - \hat{U}_i^k(x_\star^k, \sigma^k(x_\star^k)) \right).$$

Therefore, by combining the latter with the results (3.49) and (3.50), we can upper bound (3.48) as

$$\begin{aligned} R_T &\leq C\bar{k} + \sum_{k=1}^T \hat{f}^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=1}^T \hat{f}^k(\hat{x}_\star^k, \sigma^k(\hat{x}_\star^k)) \\ &\quad - \sum_{k=\bar{k}}^T \sum_{i=1}^N \left(U_i(x_\star^k, \sigma^k(x_\star^k)) - \hat{U}_i^k(x_\star^k, \sigma^k(x_\star^k)) \right) \\ &\quad + \sum_{k=\bar{k}}^T \sum_{i=1}^N \left(U_i(\mathrm{x}_i^k, \sigma^k(\mathrm{x}_i^k)) - \hat{U}_i^k(\mathrm{x}_i^k, \sigma^k(\mathrm{x}_i^k)) \right) \\ &\overset{(a)}{\leq} C\bar{k} + \sum_{k=\bar{k}}^T \hat{f}^k(\mathrm{x}^k, \sigma^k(\mathrm{x}^k)) - \sum_{k=\bar{k}}^T \hat{f}^k(\hat{x}_\star^k, \sigma^k(\hat{x}_\star^k)) + 2N \sum_{k=\bar{k}}^T \beta^k, \end{aligned} \tag{3.51}$$

where in (a) we use (3.42a) to bound the two sums related to the learning errors. Now, we notice that, in light of Lemma 3.5, $\nabla \hat{f}^k$ is $\overline{L}_1$-Lipschitz continuous with probability $1-\nu$ for all $k \geq \bar{k}$. Thus, by applying the Descent Lemma (cf.[14, Proposition 6.1.2]) on the right-hand side of the inequality (3.51), we have that, with probability $1-\nu$, it holds

$$\begin{aligned} R_T &\leq C\bar{k} + \frac{\overline{L}_1}{2} \sum_{k=\bar{k}}^T \left\| \mathrm{x}^k - \hat{x}_\star^k \right\|^2 + 2N \sum_{k=\bar{k}}^T \beta^k \\ &\overset{(a)}{\leq} C\bar{k} + \frac{\overline{L}_1}{2} \sum_{k=\bar{k}}^T \left\| \mathrm{e}^k \right\|^2 + 2N \sum_{k=\bar{k}}^T \beta^k, \end{aligned} \tag{3.52}$$

where in (a) we use the bound $\left\| \mathrm{x}^k - \hat{x}_\star^k \right\|^2 \leq \left\| \mathrm{e}^k \right\|^2$ which follows by the definition of e^k given in (3.43). Recalling that all norms are equivalent on finite-dimensional vector spaces, there always exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\|\cdot\| \leq \lambda_1 \|\cdot\|_\iota$ and $\|\cdot\|_\iota \leq \lambda_2 \|\cdot\|$. Thus, the inequality (3.52) can be upper bounded as

$$R_T \leq C\bar{k} + \frac{\overline{L}_1 \lambda_1^2}{2} \sum_{k=\bar{k}}^T \left\| \mathrm{e}^k \right\|_\iota^2 + 2N \sum_{k=\bar{k}}^T \beta^k. \tag{3.53}$$

By combining the results (3.46) and (3.53), we can write

$$
\begin{aligned}
R_T \leq \frac{\overline{L}_1\lambda_1^2}{2}\Bigg(\sum_{k=\bar{k}}^{T}\tilde{\rho}^{2(k-\bar{k})}\left\|\mathrm{e}^{\bar{k}}\right\|_\iota + \sum_{k=\bar{k}}^{T}2\rho^{k-\bar{k}}\left\|\mathrm{e}^{\bar{k}}\right\|_\iota\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{q}\left\|Bu^{k-q-1}\right\|_\iota \\
+\sum_{k=\bar{k}}^{T}\left(\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{q}\left\|Bu^{k-q-1}\right\|_\iota\right)^2\Bigg) + C\bar{k} + 2N\sum_{k=1}^{T}\beta^k \\
\overset{(a)}{\leq} \frac{\overline{L}_1\lambda_1^2}{2}\left(\frac{\left\|\mathrm{e}^{\bar{k}}\right\|_\iota^2}{1-\tilde{\rho}^2} + 2\left\|\mathrm{e}^{\bar{k}}\right\|_\iota\sum_{k=\bar{k}}^{T}\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{k-\bar{k}+q}\left\|Bu^{k-q-1}\right\|_\iota + \sum_{k=\bar{k}}^{T}\left(\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{q}\left\|Bu_{k-q-1}\right\|_\iota\right)^2\right) \\
+ C\bar{k} + 2N\sum_{k=1}^{T}\beta^k \\
\overset{(b)}{\leq} \frac{\overline{L}_1\lambda^2}{2}\left(\frac{\left\|\mathrm{e}^{\bar{k}}\right\|^2}{1-\tilde{\rho}^2} + 2\left\|\mathrm{e}^{\bar{k}}\right\|\sum_{k=\bar{k}}^{T}\sum_{k=\bar{k}}^{k-1}\tilde{\rho}^{k-\bar{k}+q}\left\|Bu^{k-q-1}\right\| + \sum_{k=\bar{k}}^{T}\left(\sum_{q=\bar{k}}^{k-1}\tilde{\rho}^{q}\left\|Bu^{k-q-1}\right\|\right)^2\right) \\
+ C\bar{k} + 2N\sum_{k=1}^{T}\beta^k,
\end{aligned}
$$

where in (a) we use the geometric series property, and in (b) we use the relation $\|\cdot\|_\iota \leq \lambda_2\|\cdot\|$ and set $\lambda := \lambda_1\lambda_2$. The proof follows by using the triangle inequality to bound the terms $\left\|Bu^{k-q-1}\right\|$ and by invoking the definitions of W_T, Q_T and $\mathcal{B}_T$ (cf. (3.45)). ■

In the next, we employ Theorem 3.2 to characterize the asymptotic performance of RLS Projected Aggregative Tracking in terms of dynamic average regret. In particular, the next corollary guarantees that, if the variations of the problem over time are bounded by a constant, then the average dynamic regret asymptotically converges to a constant.

Corollary 3.2 (Average Dynamic Regret)**.** *Consider the same assumptions of Theorem 3.2 and assume that the problem variations over time are bounded by a constant, i.e., that there exists $D > 0$ such that $\eta^k, \omega^k, \zeta^k \leq D$ for any $k \geq 0$. Then, for any $\nu \in (0, 1]$, there exists $\lambda, \bar{k} \geq 0$ and $\bar{\delta} \in (0, 1)$ such that, for any $\delta \in (0, \bar{\delta})$, there exists $\tilde{\rho} \in (0, 1)$ such that, with probability $1 - \nu$, it holds*

$$
\lim_{T\to\infty}\frac{R_T}{T} \leq \frac{\overline{L}_1\lambda^2(4+\overline{L}_2)^2D^2}{2(1-\tilde{\rho})}.
$$

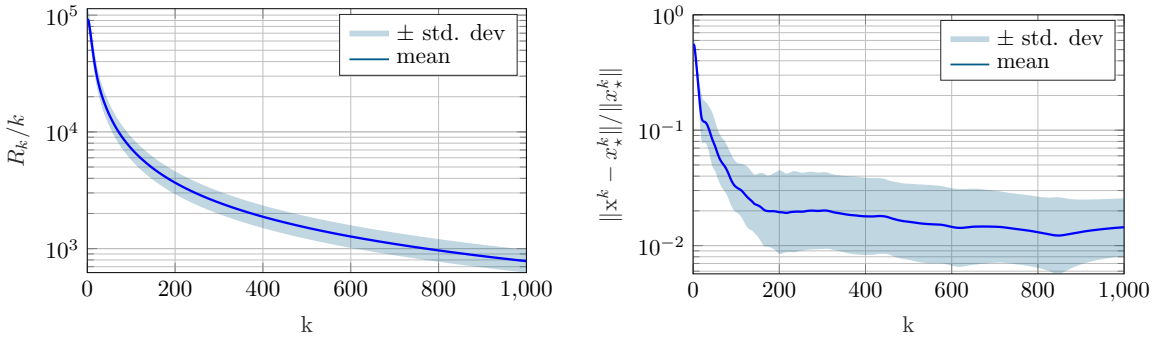
Proof. The result follows by taking the limit of the bound (3.44) and combining it with the result (3.37) of Lemma 3.5 and the properties of the geometric series with $\tilde{\rho} \in (0, 1)$. ■

3.3.2 Numerical Simulations

We consider a network of $N = 10$ agents that want to optimize their opinions about d different topics. In particular, given a time-varying *prejudice* $p_i^k \in \mathbb{R}^d$ for all $k \geq 0$, each agent i wants that its opinion $x_i \in \mathbb{R}^d$ stays as close as possible to this prejudice. Further, each agent i, would also follow the weighted average opinion $\sigma^k(\mathrm{x}) := \frac{1}{N}\sum_{i=1}^N a_i^k \mathrm{x}_i$. Each weight $a_i^k > 0$ represents the social influence of the agent i at time k. This framework can be captured by local engineering functions V_i^k of the form

$$V_i^k(\mathrm{x}_i, \sigma^k(\mathrm{x})) := \frac{1}{2}\left\|\mathrm{x}_i - p_i^k\right\|^2 + \frac{\alpha}{2}\left\|\mathrm{x}_i - \sigma^k(\mathrm{x})\right\|^2,$$

where $\alpha > 0$. Further, we assume that each agent i takes into account the evaluation of a *personalized expert*. Indeed, given the opinion x_i^k of the agent i at iteration k and its estimate s_i^k about the weighted average opinion σ^k of the network, the expert provides a noisy evaluation $\mathrm{z}_i^k = U_i(\mathrm{x}_i^k, \mathrm{s}_i^k) + \epsilon_i^k$ expressing its disagreement with both x_i^k and s_i^k according to a quadratic function as in (3.32). In addition, we include the sets $X_i := [0, 100] \times \cdots \times [0, 100]$ to bound the opinions. The agents communicate according to an undirected, connected Erdős-Rényi graph with connectivity parameter 0.5. We fix $d = 2$, and choose piecewise linear laws for the weights a_i^k, while we pick prejudices p_i^k that vary as $p_i^k = p_{i,c} + r\mathrm{col}(\cos(k/100), \sin(k/100))$, where $r = 1$ and $p_{i,c} \in [0, 100]^2$ is a randomly generated center. Further, we randomly choose each P_i, q_i, and r_i of (3.32) and consider a measurement noise $\epsilon_i^k \sim \mathcal{N}(0, 1)$. We select each component of x_i^0 and p_i^k from the interval $[0, 100]$ with a uniform random distribution. As for the algorithm parameters, we set $r_1^0 = \cdots = r_N^0 = 50$, $\gamma = 0.5$, and $\delta = 0.1$. We perform 20 Monte Carlo trials whose results are provided in Figure 3.4a in terms of average dynamic regret R_T/T. Finally, in Figure 3.4b, we show the achieved relative error $\frac{\|\mathrm{x}^k - x_\star^k\|}{x_\star^k}$.



(a) Mean of the average dynamic regret.

(b) Mean of the relative solution error.

Figure 3.4: Numerical results with 1-standard deviation band over 20 Monte Carlo trials.

3.4 Distributed Feedback Aggregative Optimization

In this section, we present and address the distributed feedback aggregative optimization framework. We consider a system of $N \in \mathbb{N}$ agents. The dynamics of the i-th agent is described by

$$\dot{x}_i = p_i(x_i, u_i), \tag{3.54}$$

where $p_i : \mathbb{R}^{n_i} \times \mathbb{R}^{m_i} \to \mathbb{R}^{n_i}$, and $x_i \in \mathbb{R}^{n_i}$, $u_i \in \mathbb{R}^{m_i}$ denote the state and the control of the i-th agent.

The following assumption is customary in the literature.

Assumption 3.9 (Steady-State map)**.** *For all $i \in \{1, \dots, N\}$ and for any $u_i \in \mathbb{R}^{m_i}$, there exists $h_i : \mathbb{R}^{m_i} \to \mathbb{R}^{n_i}$ such that $h_i(u_i) \in \mathbb{R}^{n_i}$ represents a unique globally exponentially stable equilibrium point for* (3.54)*. Moreover, there exist $L_h, L_p > 0$ such that*

$$\begin{aligned} \|h_i(u_i) - h_i(u_i')\| &\leq L_h \|u_i - u_i'\| \\ \|p_i(x_i, u_i) - p_i(x_i', u_i')\| &\leq L_p \|\textsc{col}(x_i, u_i) - \textsc{col}(x_i', u_i')\|, \end{aligned}$$

for any $x_i, x_i' \in \mathbb{R}^{n_i}$, $u_i, u_i' \in \mathbb{R}^{m_i}$, and all $i \in \{1, \dots, N\}$. Furthermore, $\ker(\nabla h_i(u_i)) = 0$ for any $u_i \in \mathbb{R}^{m_i}$. △

The agents cooperate with the aim of reaching a configuration which represents a stationary point with respect to an unconstrained instance of (1.4), i.e., the optimization problem described by

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{N} f_i(x_i, \sigma(x)), \tag{3.55}$$

in which $x := \textsc{col}(x_1, \dots, x_N) \in \mathbb{R}^n$ is the global decision vector with each $x_i \in \mathbb{R}^{n_i}$ with $n := \sum_{i=1}^{N} n_i$, and $\sigma : \mathbb{R}^n \to \mathbb{R}^d$ is the *aggregation function* defined as

$$\sigma(x) = \frac{\sum_{i=1}^{N} \phi_i(x_i)}{N}, \tag{3.56}$$

where $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}^d$ be the i-th contribution. In the following, we will also use the shorthand

$$F(x, \sigma(x)) := \sum_{i=1}^{N} f_i(x_i, \sigma(x)), \tag{3.57}$$

and the operator $G : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^n$ defined as

$$G(x) := \nabla F(v, \sigma(v)) \mid_{v=x} .$$

According to the distributed computation paradigm, we assume that the global information $\sigma(x)$ and $F(x, \sigma(x))$ are not locally available for the single agent i. Further, we also satisfy the feedback optimization paradigm in the following sense. The analytic expression of the local objective functions and aggregation rules are not available for the agents, they can be only measured according to current local variables. In particular, each agent i can only access $\nabla_1 f_i(x_i, s_i)$, $\nabla_2 f_i(x_i, s_i)$, $\phi_i(x_i)$, and $\nabla\phi_i(x_i)$, where x_i is its current state, while $s_i \in \mathbb{R}^d$ is its local estimate of the aggregative variable.

Assumption 3.10 (Function Regularity)**.** *The global objective function $F(x)$ is radially unbounded and differentiable. Moreover, there exist $L_0, L_1, L_2 > 0$ such that*

$$\begin{aligned}
\|G(x) - G(x')\| &\leq L_0 \|x - x'\| \\
\left\|\nabla_1 f_i(x_i, y_i) - \nabla_1 f_i(x_i', y_i')\right\| &\leq L_1 \left\| \begin{bmatrix} x_i - x_i' \\ y_i - y_i' \end{bmatrix} \right\| \\
\|\nabla_2 f_i(x_i, y_i) - \nabla_2 f_i(x_i', y_i')\| &\leq L_2 \left\| \begin{bmatrix} x_i - x_i' \\ y_i - y_i' \end{bmatrix} \right\|,
\end{aligned}$$

for any $x, x' \in \mathbb{R}^n$, $y, y' \in \mathbb{R}^{Nd}$, $x_i, x_i' \in \mathbb{R}^{n_i}$, $y_i, y_i' \in \mathbb{R}^d$, and all $i \in \{1, \dots, N\}$. Further, the aggregation functions ϕ_i are differentiable and there exists $L_3 > 0$ such that

$$\left\|\phi_i(x_i) - \phi_i(x_i')\right\| \leq L_3 \left\|x_i - x_i'\right\|,$$

for any $x_i, x_i' \in \mathbb{R}^{n_i}$ and all $i \in \{1, \dots, N\}$. △

The communication among the agents is performed according to a directed graph $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ with $\mathcal{E} \subset \{1, \dots, N\} \times \{1, \dots, N\}$ being the edge set. If an edge (j, i) belongs to $\mathcal{E}$, then agent i can receive information from agent j, otherwise not. The set of (in-)neighbors of agent i is defined as $\mathcal{N}_i := \{j \in \{1, \dots, N\} \mid (j, i) \in \mathcal{E}\}$. We associate to the graph $\mathcal{G}$ a weighted adjacency matrix $\mathcal{A} \in \mathbb{R}^{N \times N}$ whose entries satisfy $a_{ij} > 0$ whenever $(j, i) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. The weighted in-degree and out-degree of agent i are defined as $d_i^{\text{in}} = \sum_{j \in \mathcal{N}_i} a_{ij}$ and $d_i^{\text{out}} = \sum_{j \in \mathcal{N}_i} a_{ji}$, respectively. Finally, we associate to $\mathcal{G}$ the so-called Laplacian matrix defined as $\mathcal{L} := \mathcal{D}^{\text{in}} - \mathcal{A}$, where $\mathcal{D}^{\text{in}} := \text{diag}(d_1^{\text{in}}, \dots, d_N^{\text{in}}) \in \mathbb{R}^{N \times N}$.

Assumption 3.11 (Communication graph)**.** *The graph $\mathcal{G}$ is strongly connected and weight-balanced, namely $d_i^{in} = d_i^{out}$ for all $i \in \{1, \dots, N\}$.* △

Let $X := \{x \in \mathbb{R}^n \mid \nabla F(x, \sigma(x)) = 0\}$ be the set of stationary points of problem (3.55). Then, the aim of the paper is to design a *distributed feedback optimization* law $u := \text{col}(u_1, \dots, u_n)$ steering $\|x\|_X$ to zero.

3.4.1 Aggregative Tracking Feedback: Distributed Control Law Description and Analysis

Now, we introduce Aggregative Tracking Feedback, i.e., a distributed feedback optimization law designed to steer the agents' states, whose local dynamics are given in (3.54), to a configuration corresponding to a stationary point of problem (3.55).

To introduce the proposed law, given any $u_i \in \mathbb{R}^{m_i}$, let us study the optimization problem when $x_i = h_i(u_i)$ for all $i \in \{1, \dots, N\}$, i.e., when each agent has already reached its steady-state configuration (see Assumption 3.9). Let us define $u := \text{col}(u_1, \dots, u_N) \in \mathbb{R}^m$, with $m := \sum_{i=1}^N m_i$, and $h(u) := \text{col}(h_1(u_1), \dots, h_N(u_N)) \in \mathbb{R}^m$. Then the optimization problem (3.55) becomes

$$\min_{u \in \mathbb{R}^m} \sum_{i=1}^N f_i(h_i(u_i), \sigma(h(u))). \tag{3.58}$$

It is well-known that (3.58) can be addressed by adopting the continuous-time gradient method (see, e.g., [17]), which, for all $i \in \{1, \dots, N\}$, reads as

$$\begin{aligned} \dot{u}_i &= -\frac{\partial}{\partial u_i} F(h(u), \sigma(h(u))) \\ &= -\nabla h_i(u_i) \left(\nabla_1 f_i(h_i(u_i), \sigma(h(u))) + \frac{\nabla \phi_i(h_i(u_i))}{N} \sum_{j=1}^N \nabla_2 f_j(h_i(u_j), \sigma(h(u))) \right). \end{aligned} \tag{3.59}$$

However, agent i does not analytically know the functions appearing in (3.59). It can only accesses related measurements evaluated in its current state x_i, thus (3.59) needs to be modified as

$$\dot{u}_i = -\nabla h_i(u_i) \left(\nabla_1 f_i(x_i, \sigma(x)) + \frac{\nabla \phi_i(x_i)}{N} \sum_{j=1}^N \nabla_2 f_j(x_j, \sigma(x)) \right). \tag{3.60}$$

In turn, the control law in (3.60) cannot be implemented in a distributed fashion because $\sigma(x)$ and $\sum_{j=1}^N \nabla_2 f_j(x_j, \sigma(x))$ need a centralized information. To overcome this limitation, let

$$\begin{aligned} \pi_i^w(x) &:= -\phi_i(x_i) + \sigma(x) \\ \pi_i^z(x) &:= -\nabla_2 f_i(x_i, \sigma(x)) + \sum_{j=1}^N \nabla_2 f_j(x_j, \sigma(x))/N, \end{aligned}$$

and modify (3.60) as

$$\begin{aligned}\dot{u}_i = &-\nabla h_i(u_i)\left(\nabla_1 f_i(x_i, \pi_i^w(x) + \phi_i(x_i)) + \nabla\phi_i(x_i)\pi_i^z(x)\right)\\ &- \nabla h_i(u_i)\nabla\phi_i(x_i)\nabla_2 f_i(x_i, \pi_i^w(x) + \phi_i(x_i)).\end{aligned} \tag{3.61}$$

The strategy is that of designing estimations for π_i^w and π_i^z, namely $w_i, z_i \in \mathbb{R}^d$, such that

$$\begin{aligned}\lim_{t\to\infty} \|w_i(t) - \pi_i^w(x(t))\| = 0\\ \lim_{t\to\infty} \|z_i(t) - \pi_i^z(x(t))\| = 0,\end{aligned}$$

for all $i \in \{1, \dots, N\}$. To this end, inspired by the continuous-time compensation dynamics of the auxiliary variables in (2.47), we embed two consensus-based mechanisms giving rise to the distributed feedback optimization law termed Aggregative Tracking Feedback and resumed in Algorithm 7. The parameters $\alpha_1, \alpha_2 > 0$ tune the system dynamics. The role of the initialization $w_i(0) = z_i(0) = 0$ for all $i \in \{1, \dots, N\}$ will be detailed in the analysis of the scheme. Fig. 3.5 describes the closed-loop system (3.62) in terms of block-diagrams.

Algorithm 7 Aggregative Tracking Feedback

Agent i perspective
initialization: $x_i(0), u_i(0) \in \mathbb{R}^{n_i}$, $w_i(0) = z_i(0) = 0$

$$\dot{x}_i = p_i(x_i, u_i) \tag{3.62a}$$

$$\dot{u}_i = -\alpha_1 \nabla h_i(u_i)\left(\nabla_1 f_i(x_i, w_i + \phi_i(x_i)) + \nabla\phi_i(x_i)\left(z_i + \nabla_2 f_i(x_i, w_i + \phi_i(x_i))\right)\right) \tag{3.62b}$$

$$\dot{w}_i = -\frac{\alpha_1}{\alpha_2}\sum_{j\in\mathcal{N}_i} a_{ij}\left(w_i + \phi_i(x_i) - w_j - \phi_j(x_j)\right) \tag{3.62c}$$

$$\dot{z}_i = -\frac{\alpha_1}{\alpha_2}\sum_{j\in\mathcal{N}_i} a_{ij}\left(z_i + \nabla_2 f_i(x_i, w_i + \phi_i(x_i)) - z_j - \nabla_2 f_j(x_j, w_j + \phi_j(x_j))\right) \tag{3.62d}$$

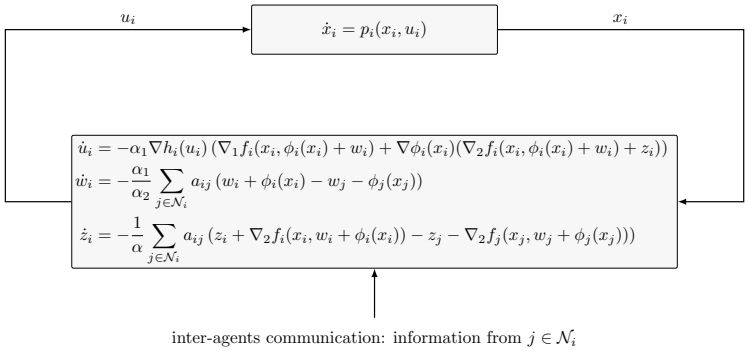


Figure 3.5: Block diagram describing (3.62).

Theorem 3.3. *Consider the closed-loop system* (3.62) *and let Assumptions 3.9, 3.10, and 3.11 hold. Then, there exist* $\bar{\alpha}_1 > 0$ *and* $\bar{\alpha}_2 > 0$ *such that, for any* $\alpha_1 \in (0, \bar{\alpha}_1)$, $\alpha_2 \in (0, \bar{\alpha}_2)$ *and* $\text{col}(x_i(0), u_i(0), w_i(0), z_i(0)) \in \mathbb{R}^{n_i+m_i+2d}$ *such that* $z_i = w_i = 0$ *for all* $i \in \{1, \dots, N\}$*, it holds*

$$\lim_{t \to \infty} \|x(t)\|_X = 0.$$

The proof of Theorem 3.3 will be carried out in the next after some preparatory results. Theorem 3.3 guarantees that Aggregative Tracking Feedback asymptotically steers the network state $x(t)$ into the set x of stationary points of problem (3.55).

We now give an overview of the steps needed to prove Theorem 3.3:

(i) We reformulate (3.62) as the interconnection of three dynamic subsystems describing the evolution of all the states, the control inputs, and (a suitable transformation of) the auxiliary variables.

(ii) Within three separate lemmas we give suitable properties of the time-derivative of three different Lyapunov-like functions. Specifically, each one of these lemmas assesses the convergence of one of the three subsystems identified within step (i) when the convergence of the other subsystems has already occurred.

(iii) To conclude, we define a candidate Lyapunov function for the whole system and, relying on the lemmas of step (ii) and LaSalle arguments, we study its time-derivative to prove Theorem 3.3.

According to step (i), we reformulate (3.62) by leveraging the initialization of w and z and the consensus properties of their dynamics. To this end, we start by defining $L = \mathcal{L} \otimes I_d$, $w = \text{col}(w_1, \dots, w_N)$, $z = \text{col}(z_1, \dots, z_N)$, and by introducing the operators

$G_1 : \mathbb{R}^n \times \mathbb{R}^{Nd} \to \mathbb{R}^n$ and $G_2 : \mathbb{R}^n \times \mathbb{R}^{Nd} \to \mathbb{R}^n$ given by

$$G_1(x, \mathrm{s}) = \begin{bmatrix} \nabla_1 f_1(x_1, \mathrm{s}_1) \\ \vdots \\ \nabla_1 f_N(x_N, \mathrm{s}_N) \end{bmatrix}, G_2(x, \mathrm{s}) = \begin{bmatrix} \nabla_2 f_1(x_1, \mathrm{s}_1) \\ \vdots \\ \nabla_2 f_N(x_N, \mathrm{s}_N) \end{bmatrix},$$

where we used the decomposition $x = \mathrm{col}(x_1, \dots, x_N)$ and $\mathrm{s} = \mathrm{col}(\mathrm{s}_1, \dots, \mathrm{s}_N)$ with $x_i \in \mathbb{R}^{n_i}$ and $\mathrm{s}_i \in \mathbb{R}^d$ for all $i \in \{1, \dots, N\}$. Then, the stacked column form of (3.62) reads as

$$\dot{x} = p(x, u) \tag{3.63a}$$

$$\dot{u} = -\alpha_1 \nabla h(u) G_1(x, w + \phi(x)) - \alpha_1 \nabla h(u) \nabla \phi(x) \left(z + G_2(x, w + \phi(x))\right) \tag{3.63b}$$

$$\dot{w} = -\frac{\alpha_1}{\alpha_2} L \left(w + \phi(x)\right) \tag{3.63c}$$

$$\dot{z} = -\frac{\alpha_1}{\alpha_2} L \left(z + G_2(x, w + \phi(x))\right). \tag{3.63d}$$

Next, we rewrite (3.63) in order to highlight the average dynamics of w and z and their orthogonal ones. To this end, we investigate the effect of the initialization $w_i(0) = z_i(0) = 0$ for all $i \in \{1, \dots, N\}$. Let

$$\mathcal{S} := \{\mathrm{col}(x, u, w, z) \in \mathbb{R}^{n+m+2Nd} \mid \mathbf{1}_{N,d}^\top w = 0, \mathbf{1}_{N,d}^\top z = 0\},$$

and note that $\mathcal{S}$ is invariant for (3.63) because $\mathbf{1}_{N,d}^\top L = 0$ (cf. Assumption 3.11). Hence, we can exploit a change of coordinates to take advantage from this property. To this end, let us introduce $R_d \in \mathbb{R}^{Nd \times (N-1)d}$ such that $R_d^\top R_d = I$, $R_d^\top \mathbf{1}_{N,d} = 0$, and $\|R_d\| = 1$ and the matrix $T \in \mathbb{R}^{2Nd \times 2Nd}$ defined as

$$T := \begin{bmatrix} R_d^\top \\ \mathbf{1}_{N,d}^\top / N \end{bmatrix}.$$

The matrix T is invertible and we define η, $\zeta \in \mathbb{R}^{(N-1)d}$, η_{avg}, $\zeta_{\mathrm{avg}} \in \mathbb{R}^d$ as

$$\begin{bmatrix} \eta \\ \eta_{\mathrm{avg}} \end{bmatrix} := Tw, \quad \begin{bmatrix} \zeta \\ \zeta_{\mathrm{avg}} \end{bmatrix} := Tz. \tag{3.64}$$

Then, by using (3.63c)-(3.63d), we note that

$$\dot{\eta}_{\mathrm{avg}} = 0, \quad \dot{\zeta}_{\mathrm{avg}} = 0.$$

Therefore the initialization $w(0) = z(0) = 0$ guarantees that $\eta_{\mathrm{avg}}(t) = \zeta_{\mathrm{avg}}(t) =$

$0, \qquad \forall\, t \geq 0$. Thus, combining this result with (3.64) it follows

$$w = R_d\, \eta, \quad z = R_d\, \zeta. \tag{3.65}$$

As a consequence, defining $\psi := \text{col}(\eta, \zeta)$ and using (3.64)-(3.65) we can restrict the dynamics (3.63c)-(3.63d) to

$$\dot{\psi} = \frac{\alpha_1}{\alpha_2} \begin{bmatrix} -R_d^\top L R_d & 0 \\ 0 & -R_d^\top L R_d \end{bmatrix} \psi + \frac{\alpha_1}{\alpha_2} \begin{bmatrix} -R_d^\top L & 0 \\ 0 & -R_d^\top L \end{bmatrix} \begin{bmatrix} \phi(x) \\ G_2(x, \begin{bmatrix} R_d & 0 \end{bmatrix} \psi + \phi(x)) \end{bmatrix}. \tag{3.66}$$

Note that

$$\bar{\psi}(x) := - \begin{bmatrix} R_d^\top & 0 \\ 0 & R_d^\top \end{bmatrix} \begin{bmatrix} \phi(x) \\ G_2(x, \mathbf{1}_{N,d}\sigma(x)) \end{bmatrix} \tag{3.67}$$

represents an equilibrium for (3.66) for any $x \in \mathbb{R}^n$. Based on this observation, let us introduce the error coordinate $\xi \in \mathbb{R}^{2(N-1)d}$ defined as

$$\xi := \psi - \bar{\psi}(x). \tag{3.68}$$

As a consequence, using (3.65), the definition of ψ, (3.67), and (3.68), we have

$$w = \begin{bmatrix} R_d & 0 \end{bmatrix} \xi - R_d R_d^\top \phi(x) \tag{3.69a}$$

$$z = \begin{bmatrix} 0 & R_d \end{bmatrix} \xi - R_d R_d^\top G_2(x, \mathbf{1}_{N,d}\sigma(x)). \tag{3.69b}$$

Let use introduce the selection matrices $\mathcal{R}_1, \mathcal{R}_2 \in \mathbb{R}^{Nd \times 2(N-1)d}$ defined as

$$\mathcal{R}_1 := \begin{bmatrix} R & 0 \end{bmatrix} \quad \mathcal{R}_2 := \begin{bmatrix} 0 & R \end{bmatrix}. \tag{3.70}$$

Then, by exploiting (3.56), (3.66), (3.68), (3.69), (3.70), and $I - R_d R_d^\top = \mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top / N$, we rewrite (3.63) as the equivalent, restricted dynamics

$$\dot{x} = p(x, u) \tag{3.71a}$$

$$\begin{aligned} \dot{u} = &-\alpha_1 \nabla h(u) \left(G_1(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) + \nabla\phi(x) \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} G_2(x, \mathbf{1}_{N,d}\sigma(x)) \right) \\ &- \alpha_1 \nabla h(u) \left(G_2(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x)) + \nabla\phi(x)\mathcal{R}_2\xi \right) \end{aligned} \tag{3.71b}$$

$$\begin{aligned} \dot{\xi} = &\frac{\alpha_1}{\alpha_2} \begin{bmatrix} -R_d^\top L R_d & 0 \\ 0 & -R_d^\top L R_d \end{bmatrix} \xi + \frac{\alpha_1}{\alpha_2} \begin{bmatrix} 0 \\ R_d^\top L (G_2(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x))) \end{bmatrix} \\ &- \nabla\bar{\psi}(x)\, p(x, u). \end{aligned} \tag{3.71c}$$

With (3.71) at hand, we provide three preparatory results needed to prove Theorem 3.3.

Lemma 3.10. *There exists a function $W : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ such that, along the trajectories of* (3.71a) *and* (3.71b)*, it holds*

$$c_1 \|x - h(u)\|^2 \leq W(x,u) \leq c_2 \|x - h(u)\|^2 \tag{3.72a}$$

$$\begin{aligned}\dot{W}(x,u) \leq &-(c_3 - \alpha_1 c_4) \|x - h(u)\|^2 + \alpha_1 c_5 \|x - h(u)\| \|\nabla h(u) G(h(u))\| \\ &+ \alpha_1 c_5 c_6 \|x - h(u)\| \|\xi\|,\end{aligned} \tag{3.72b}$$

for some $c_1, c_2, c_3, c_4, c_5, c_6 > 0$.

Proof. By using the Converse Lyapunov Theorem (cf. [166, Theorem 5.17]), the exponential stability of $h(u)$, and the Lipschitz continuity of h (cf. Assumption 3.9), there exists $W : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ such that

$$c_1 \|x - h(u)\|^2 \leq W(x,u) \leq c_2 \|x - h(u)\|^2 \tag{3.73a}$$

$$\nabla_1 W(x,u)\, p(x,u) \leq -c_3 \|x - h(u)\|^2 \tag{3.73b}$$

$$\nabla_2 W(x,u) \leq c_5 \|x - h(u)\|, \tag{3.73c}$$

for some positive constant $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, and $c_5 > 0$. In light of (3.73a), we need only to show (3.72b). To this end, we evaluate $\dot{V}(x,u,\xi)$ along the trajectories of (3.71a) and (3.71b), thus obtaining

$$\begin{aligned}\dot{W}(x,u) &= \nabla_1 W(x,u) p(x,u) + \nabla_2 W(x,u) \dot{u} \\ &\overset{(a)}{\leq} -c_3 \|x - h(u)\|^2 + \nabla_2 W(x,u) \dot{u} \\ &\overset{(b)}{\leq} -c_3 \|x - h(u)\|^2 + c_5 \|x - h(u)\| \|\dot{u}\|,\end{aligned} \tag{3.74}$$

where in (a) we use (3.73b), and in (b) we use the Cauchy-Schwartz inequality with condition (3.73c). Note that

$$G(x) = G_1(x, \mathbf{1}_{N,d}\sigma(x)) + \nabla\phi(x) \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} G_2(x, \mathbf{1}_{N,d}\sigma(x)).$$

Then, by adding and subtracting $\alpha_1 \nabla h(u) G_1(x, \mathbf{1}_{N,d}\sigma(x))$ into (3.71b), we get

$$\begin{aligned}\dot{u} = &-\alpha_1 \nabla h(u) G(x) - \alpha_1 \nabla h(u) \left(G_1(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) - G_1(x, \mathbf{1}_{N,d}\sigma(x))\right) \\ &- \alpha_1 \nabla h(u) \nabla\phi(x) \left(G_2(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x)) + \mathcal{R}_2\xi\right).\end{aligned} \tag{3.75}$$

Moreover, by using the Lipschitz continuity properties given in Assumption 3.10, we

can write

$$\|G_1(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) - G_1(x, \mathbf{1}_{N,d}\sigma(x))\| \leq L_1\|\xi\| \tag{3.76a}$$
$$\|G_2(x, \mathcal{R}_1\xi + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x))\| \leq L_2\|\xi\| \tag{3.76b}$$
$$\|\nabla\phi(x)\| \leq L_3, \tag{3.76c}$$

Further, by exploiting Assumption 3.9, it holds

$$\|\nabla h(u)\| \leq L_h. \tag{3.76d}$$

Then, we combine (3.75), the Cauchy-Schwartz inequality, and the bounds (3.76) to obtain

$$\begin{aligned}\|\dot{u}\| &\leq \alpha_1 \|\nabla h(u)G(x)\| + \alpha_1 L_h(L_1 + (1+L_2)L_3)\|\xi\| \\ &\overset{(a)}{\leq} \alpha_1 \|\nabla h(u)G(h(u))\| + \alpha_1 \|\nabla h(u)G(x) - \nabla h(u)G(h(u))\| \\ &\quad + \alpha_1 L_h(L_1 + (1+L_2)L_3)\|\xi\| \\ &\overset{(b)}{\leq} \alpha_1 \|\nabla h(u)G(h(u))\| + \alpha_1 L_h L_0 \|x - h(u)\| + \alpha_1 L_h(L_1 + (1+L_2)L_3)\|\xi\|, \end{aligned} \tag{3.77}$$

where in (a) we add and subtract within the norm $\nabla h(u)G(h(u))$ and use the triangle inequality, in (b) use the Lipschitz continuity of h and G (cf. Assumptions 3.9 and 3.10), Finally, we use (3.77) to bound (3.74). The proof follows by setting $c_4 = L_h L_0$ and $c_6 = L_h(L_1 + (1+L_2)L_3)$. ■

We note that, by choosing $\alpha_1 \leq c_3/c_4$, the conditions (3.72) in Lemma 3.10 guarentees that, for any $u \in \mathbb{R}^m$, the point $h(u)$ is globally exponentially stable for the subsystem (3.71a) when $\nabla h(u)G(h(u)) = 0$ (cf. [91, Theorem 4.10]).

Lemma 3.11. *There exists a radially unbounded function $S : \mathbb{R}^m \to \mathbb{R}$ such that, along the trajectories of* (3.71b)*, it holds*

$$\begin{aligned}\dot{S}(u) \leq &-\alpha_1 \|\nabla h(u)G(h(u))\|^2 + \alpha_1 d_1 \|\nabla h(u)G(h(u))\| \, \|x - h(u)\| \\ &+ \alpha_1 d_2 \|\nabla h(u)G(h(u))\| \, \|\xi\|, \end{aligned} \tag{3.78}$$

for some $d_1, d_2 > 0$.

Proof. Let us consider

$$S(u) := F(h(u), \sigma(h(u))), \tag{3.79}$$

where $F(\cdot,\cdot)$ has been defined in (3.57). We remark that, in light of Assumption 3.10, S

is radially unbounded (see Definition A.2 in Appendix A). We point out that

$$\nabla F(h(u), \sigma(h(u))) = \nabla h(u) G(h(u)). \tag{3.80}$$

Thus, to evaluate $\dot{S}(u)$ along the trajectories of (3.71b), we exploit (3.80) obtaining

$$\begin{aligned}
\dot{S}(u) &= (\nabla h(u) G(h(u)))^\top \dot{u}. \\
&\overset{(a)}{\leq} -\alpha_1 (\nabla h(u) G(h(u)))^\top (\nabla h(u) G(x)) + \alpha_1 d_2 \left\| \nabla h(u) G(h(u)) \right\| \left\| \xi \right\| \\
&\overset{(b)}{=} -\alpha_1 \left\| \nabla h(u) G(h(u)) \right\|^2 \\
&\quad - \alpha_1 (\nabla h(u) G(h(u)))^\top \big(\nabla h(u) \nabla F(x, \sigma(x)) - \nabla h(u) \nabla F(h(u), \sigma(h(u))) \big) \\
&\quad + \alpha_1 d_2 \left\| \nabla h(u) G(h(u)) \right\| \left\| \xi \right\| \\
&\overset{(c)}{=} -\alpha_1 \left\| \nabla h(u) G(h(u)) \right\|^2 + \alpha_1 L_h L_0 \left\| \nabla h(u) G(h(u)) \right\| \left\| x - h(u) \right\| \\
&\quad + \alpha_1 d_2 \left\| \nabla h(u) G(h(u)) \right\| \left\| \xi \right\|,
\end{aligned}$$

where in (a) we use the results (3.75) and the bounds (3.76) setting $d_2 = L_h(L_1 + (1 + L_2)L_3)$, in (b) we add and subtract the term $\nabla h(u) G(h(u), \sigma(h(u)))$ within the brackets, and in (c) we use the Lipschitz continuity of h and $\nabla F(\cdot, \sigma(\cdot))$ (cf. Assumptions 3.9 and 3.10) and the Cauchy-Schwarz inequality. The proof follows by setting $d_1 = L_h L_0$. ■

For $\xi = 0$ and $x = h(u)$ the condition (3.78) allows us to use LaSalle arguments to claim the asymptotic convergence of u to the set $\{u \in \mathbb{R}^m \mid \nabla h(u) G(h(u)) = 0\}$.

Lemma 3.12. *There exists a function $U : \mathbb{R}^{2(N-1)d} \to \mathbb{R}$ such that, along the trajectories of* (3.71c)*, it holds*

$$b_1 \left\| \xi \right\|^2 \leq U(\xi) \leq b_2 \left\| \xi \right\|^2 \tag{3.81a}$$

$$\dot{U}(\xi) \leq -\frac{\alpha_1 b_3}{\alpha_2} \left\| \xi \right\|^2 + b_4 \left\| \xi \right\| \left\| x - h(u) \right\|, \tag{3.81b}$$

for some $b_1, b_2, b_3, b_4 > 0$.

Proof. In light of Assumption 3.11, the matrix $-R^\top L R$ is Hurwitz. Thus, there exist $P_1, P_2 \in \mathbb{R}^{(N-1)d \times (N-1)d}$ such that $P_1 = P_1^\top > 0$, $P_2 = P_2^\top > 0$, and

$$-P_1 R^\top L R - (R^\top L R)^\top P_1 = -Q_1 \tag{3.82a}$$

$$-P_2 R^\top L R - (R^\top L R)^\top P_2 = -Q_2, \tag{3.82b}$$

for any $Q_1, Q_2 \in \mathbb{R}^{(N-1)d \times (N-1)d}$ such that $Q_1 = Q_1^\top > 0$ and $Q_2 = Q_2^\top > 0$. Then, let us

consider

$$U(\xi) := \xi^\top P \xi.$$

Then, the conditions (3.81a) are satisfied by denoting $b_1 > 0$ and $b_2 > 0$ the smallest and largest eigenvalue of P, respectively. In order to show (3.81b), let $\xi_1, \xi_2 \in \mathbb{R}^{(N-1)d}$ be such that $\xi = \text{col}(\xi_1, \xi_2)$. Then, by using (3.71c) and (3.82), we can write

$$\begin{aligned}\dot{U}(\xi) = &-\frac{\alpha_1}{\alpha_2}\xi_1^\top Q_1 \xi_1 - \xi_2^\top Q_2 \xi_2 + \frac{2\alpha_1}{\alpha_2}\xi_2^\top P_2 R^\top L(G_2(x, \xi_1 + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x))) \\ &- 2\xi^\top P \nabla\bar{\psi}(x)\, p(x,u). \end{aligned} \tag{3.83}$$

Moreover, by using the Lipschitz continuity of $\nabla_2 f_i$ (cf. Assumption 3.10), we can write

$$\|G_2(x, \xi_1 + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x))\| \leq L_2 \|\xi_1\|,$$

that, combined with the application of the Cauchy-Schwarz, leads to

$$\xi_2^\top P_2 R^\top L(G_2(x, \xi_1 + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x))) \leq L_2 \left\|P_2 R^\top L\right\| \|\xi_2\| \|\xi_1\|. \tag{3.84}$$

Then, given $Q_2 > 0$, we compute P_2 such that (3.82b), and define $k_1(Q_2) := L_2 \left\|P_2 R^\top L\right\|$. Now, let us denote with q_1, q_2 the smallest eigenvalues of Q_1 and Q_2, and define

$$\tilde{Q} := \begin{bmatrix} q_1 & -k_1(Q_2) \\ -k_1(Q_2) & q_2 \end{bmatrix}.$$

Then, by using (3.84), we can write

$$\begin{aligned}&-\xi_1^\top Q_1 \xi_1 - \xi_2^\top Q_2 \xi_2 + 2\xi_2^\top P_2 R^\top L(G_2(x, R\,\xi_1 + \mathbf{1}_{N,d}\sigma(x)) - G_2(x, \mathbf{1}_{N,d}\sigma(x))) \\ &\leq -\begin{bmatrix}\|\xi_1\| & \|\xi_2\|\end{bmatrix} \tilde{Q}\, \text{col}(\|\xi_1\|, \|\xi_2\|). \end{aligned} \tag{3.85}$$

Let us choose $b_3 \in (0, q_2)$ and $Q_1 > 0$ such that $q_1 > (b_3 q_2 + k_1(Q_2)^2)/(q_2 - b_3)$. The, it holds $\tilde{Q} \geq b_3 I$ which, combined with (3.85), allows to bound (3.83) as

$$\dot{U}(\xi) \leq -\frac{\alpha_1 b_3}{\alpha_2}\|\xi\|^2 + 2\xi^\top P \nabla\bar{\psi}(x)\, p(x,u). \tag{3.86}$$

Since $p(h(u), u) = 0$ (see Assumption 3.9), it holds

$$\xi^\top P \nabla\bar{\psi}(x)\, p(x,u) = \xi^\top P \nabla\bar{\psi}(x)\, (p(x,u) - p(h(u),u)).$$

Recall that, thanks to Assumption 3.9, it holds

$$\|p(x,u) - p(h(u),u)\| \leq L_p \|x - h(u)\|.$$

On the other hand, by using the Cauchy-Schwarz inequality, Assumption 3.10, $\|R\| = 1$, and $\|1_N 1_n^\top\| = \sqrt{Nn}$, we can write the bound

$$\begin{aligned}\|\nabla\bar{\psi}(x)\| &\leq \left\|\begin{bmatrix} R^\top & 0 \\ 0 & R^\top \end{bmatrix}\right\| \left\|\begin{bmatrix} \nabla\phi(x) \\ \nabla G_2(x, \mathbf{1}_{N,d}\sigma(x)) \end{bmatrix}\right\| \\ &\leq (L_2\sqrt{Nn} + L_3).\end{aligned}$$

Thus, we can bound (3.86) as

$$\dot{U}(\xi) \leq -\frac{\alpha_1 b_3}{\alpha_2} \|\xi\|^2 + b_4 \|x - h(u)\|,$$

with $b_4 := \frac{L_p\|P\|(L_2\sqrt{Nn}+L_3)}{2}$ and the proof is given. ■

We highlight that Lemma 3.12 proves that, if $x = h(u)$, then the origin is a globally exponentially stable equilibrium point for (3.71c) (cf. [91, Theorem 4.10]).

Proof of Theorem 3.3

By using the functions W, S, and U provided by Lemma 3.10, 3.11, and 3.12, respectively, we define

$$V(x,u,\xi) = U(\xi) + W(x,u) + S(u).$$

Moreover, let us introduce

$$\begin{aligned} y(u) &:= \nabla h(u) G(h(u)) \\ k_2 &:= \frac{d_1 + c_5}{2}, \\ H_1(\alpha_1) &:= \begin{bmatrix} c_3 - \alpha_1 c_4 & -\alpha_1 k_2 \\ -\alpha_1 k_2 & \alpha_1 \end{bmatrix}. \end{aligned}$$

Then, by evaluating $\dot{V}(x,u,\xi)$ along the trajectories of (3.71) and by using (3.72b), (3.78), and (3.81b), we get

$$\begin{aligned}\dot{V}(x,u,\xi) \leq &- \begin{bmatrix} \|x - h(u)\| & \|y(u)\| \end{bmatrix} H_1(\alpha_1) \begin{bmatrix} \|x - h(u)\| \\ \|y(u)\| \end{bmatrix} + \alpha_1 d_2 \|y(u)\| \|\xi\| \\ &- \frac{\alpha_1 b_3}{\alpha_2} \|\xi\|^2 + (b_4 + \alpha_1 c_5 c_6) \|\xi\| \|x - h(u)\|. \end{aligned} \tag{3.87}$$

By Sylvester Criterion, we know that the matrix $H_1(\alpha_1) = H_1(\alpha_1)^\top \in \mathbb{R}^2$ is positive definite if and only if the following conditions are satisfied

$$\begin{cases} c_3 > \alpha_1 c_4 \\ c_3\alpha_1 > \alpha_1^2(k_2^2 + c_4). \end{cases} \tag{3.88}$$

Let $\bar{\alpha}_1 := \max\left\{c_3/c_4, c_3/(k_2^2 + c_4)\right\}$. Then, with any $\alpha_1 \in (0, \bar{\alpha}_1)$, both conditions (3.88) are satisfied allowing us to claim the positive definiteness of $H_1(\alpha_1)$. Let $h_1 > 0$ be the smallest eigenvalue of the matrix $H_1(\alpha_1)$. Then, for any $\alpha_1 \in (0, \bar{\alpha}_1)$, the right-hand member of (3.87) can be bounded by

$$\begin{aligned} \dot{V}(x,u,\xi) \leq -h_1(\|x - h(u)\|^2 + \|y(u)\|^2) + \alpha_1 d_2 \|y(u)\| \|\xi\| - \frac{\bar{\alpha}_1 b_3}{\alpha_2} \|\xi\|^2 \\ + (b_4 + \bar{\alpha}_1 c_5 c_6) \|\xi\| \|x - h(u)\|. \end{aligned} \tag{3.89}$$

Let us introduce

$$\begin{aligned} e(x,u) &:= \text{col}(x - h(u), y(u)) \\ k_3 &:= \frac{d_2 + b_4 + \bar{\alpha}_1 c_5 c_6}{2} \\ H_2(\alpha_2) &:= \begin{bmatrix} h_1 & -k_3 \\ -k_3 & \frac{\bar{\alpha}_1}{\alpha_2} b_3 \end{bmatrix}. \end{aligned}$$

Then, we can bound (3.89) as

$$\dot{V}(x,u,\xi) \leq -\begin{bmatrix} \|e(x,u)\| \\ \xi \end{bmatrix}^\top H_2(\alpha_2) \begin{bmatrix} \|e(x,u)\| \\ \xi \end{bmatrix}. \tag{3.90}$$

By Sylvester Criterion, we know that for any $\alpha_2 \in (0, \bar{\alpha}_2)$, with $\bar{\alpha}_2 := \bar{\alpha}_1 b_3 h_1 / k_3^2$, the matrix $H_2(\alpha_2) = H_2(\alpha_2)^\top \in \mathbb{R}^2$ is positive definite. Let $h_2 > 0$ be the smallest eigenvalue of $H_2(\alpha_2)$. Then, the inequality (3.90) leads to

$$\dot{V}(x,u,\xi) \leq -h_2 \|\text{col}(\|e(x,u)\|, \|\xi\|)\|^2. \tag{3.91}$$

Let us study the set in which the right-hand side of (3.91) is zero. To this end, let

$$\mathcal{U} := \{u \in \mathbb{R}^m \mid \nabla h(u) G(h(u)) = 0\},$$

and

$$E := \{(x,u,\xi) \in \mathbb{R}^{n_E} \mid x = h(u), u \in \mathcal{U}, \xi = 0\}. \tag{3.92}$$

Then $\dot{V}(x,u,\xi) = 0$ for any $(x,u,\xi) \in E$. By studying system (3.71) restricted to the subset E, we get

$$\dot{x}\big|_{(x,u,\xi)\in E} = 0 \tag{3.93a}$$

$$\dot{u}\big|_{(x,u,\xi)\in E} = 0 \tag{3.93b}$$

$$\dot{\xi}\big|_{(x,u,\xi)\in E} = 0. \tag{3.93c}$$

Hence, by (3.93) we guarantee that the largest invariant set contained within E for the dynamics (3.71) coincides with E. Therefore, by using the LaSalle's invariance principle (cf. [91, Theorem 4.4]), it holds

$$\lim_{t\to\infty} \left\| \begin{bmatrix} x(t) \\ u(t) \\ \xi(t) \end{bmatrix} \right\|_E = 0. \tag{3.94}$$

We remark that Assumption 3.9 guarantees that $\ker(\nabla h(u)) = 0$ for any $u \in \mathbb{R}^m$. As a consequence, $\mathcal{U}$ can be rewritten as

$$\mathcal{U} \equiv \{u \in \mathbb{R}^m \,|\, G(h(u)) = 0\},$$

which, in turn, allows us to claim that $(x,u,\xi) \in E \implies G(x) = 0$. The proof follows by using (3.94) and noting that $G(x) \equiv \nabla F(x, \sigma(x))$.

3.4.2 Aggregative Tracking Feedback with Single Integrator Dynamics

In this section, we adapt Aggregative Tracking Feedback for the case in which system (3.54) is replaced by single integrator dynamics. Thus, we now consider N systems whose evolution is governed by

$$\dot{x}_i = u_i, \tag{3.95}$$

for all $i \in \{1,\dots,N\}$. Indeed, despite the fact that single integrator dynamics may be simpler than the more generic one (3.54), we point out that system (3.95) does not have a steady-state configuration associated to each input u_i and, thus, violates Assumption 3.9. Moreover, in this setting, we enforce the following condition about problem (3.55).

Assumption 3.12 (Convexity)**.** *The global objective function $f(x,\sigma(x))$ is μ-strongly convex.* $\triangle$

Assumption 3.12 implies the existence of a unique optimal solution $x^\star$ (cf. Proposition A.2 in Appendix A). Hence, now the distributed feedback optimization law should

steer the network to a steady-state configuration corresponding to the optimal solution $x^\star$.

In this setting, we choose each input u_i according to the algebraic law

$$u_i = -\nabla_1 f_i(x_i, w_i + \phi_i(x_i)) - \nabla\phi_i(x_i) z_i, \tag{3.96}$$

with w_i and z_i with same role and same dynamics as in Algorithm 7. Hence, the input law (3.96), plugged into the single integrator dynamics (3.95) and combined with dynamics of the auxiliary variables w_i (cf. (3.62c)) and z_i (cf. (3.62d)), leads to the whole closed loop system described by

$$\dot{x} = -\nabla_1 f(x, w + \phi(x)) - \nabla\phi(x) z - \nabla\phi(x))\nabla_2 f(x, w + \phi(x)) \tag{3.97a}$$

$$\dot{w} = -\frac{1}{\alpha} L w - \frac{1}{\alpha} L\phi(x) \tag{3.97b}$$

$$\dot{z} = -\frac{1}{\alpha} L z - \frac{1}{\alpha} L\nabla_2 f(x, w + \phi(x)), \tag{3.97c}$$

where x, w, z, $\nabla_1 f$, $\nabla_2 f$, ϕ, and L have the same meaning as in (3.97), while $\alpha > 0$ is a control parameter.

In the next, we provide a sketch of the analysis needed to prove that system (3.97) exponentially converges to a configuration in which x corresponds to the optimal solution $x^\star$ of problem (3.55).

This analysis consists of the following main steps.

(i) We introduce a suitable error dynamics with respect to the minimum of problem (3.55). The obtained dyanmics is a *singularly perturbed system*, i.e, the interconnection of a slow subsystem with a fast one.

(ii) By "freezing" the slow system state within the fast one, we find a parametrized equilibrium that depends on this frozen state. We use this equilibrium to build the so-called boundary-layer system. Then, we provide a suitable Lyapunov function (independent of the slow state), showing the global exponential stability of the origin for the boundary layer system.

(iii) We consider the fast state lying on the steady state of the boundary layer system introduced in (ii) to build the so called reduced system. We show that the origin is a globally exponentially stable equilibrium point for the obtained system.

(iv) Finally, the stability results of steps (ii) and (iii) are exploited to demonstrate the global exponential stability of the origin for the whole interconnected system by proving that the agents' states reach the optimal solution of problem (3.55) with a linear rate.

Lemma 3.13. *Consider system* (3.97) *and denote* $\tilde{x} := x - x^\star$. *Then, there exist two changes of variables*

$$\begin{bmatrix} x \\ wz \end{bmatrix} \longmapsto \begin{bmatrix} \tilde{x} \\ \tilde{w} \\ \tilde{z} \end{bmatrix}, \quad \begin{bmatrix} \tilde{x} \\ \tilde{w} \\ \tilde{z} \end{bmatrix} \longmapsto \begin{bmatrix} \tilde{x} \\ \psi \end{bmatrix},$$

such that (3.63) *is equivalent to*

$$\dot{\tilde{x}} = h(\tilde{x}, \psi) \tag{3.98a}$$

$$\alpha\dot{\psi} = g(\tilde{x}, \psi), \tag{3.98b}$$

where $\tilde{x} \in \mathbb{R}^n$, $\tilde{z}, \tilde{w} \in \mathbb{R}^{Nd}$, $\psi \in \mathbb{R}^{2(N-d)}$ *and*

$$\begin{aligned} h(\tilde{x}, \psi) := &-\nabla_1 f\left(\tilde{x} + x^\star, \begin{bmatrix} R_d & 0 \end{bmatrix} \psi + \Delta\phi(\tilde{x}) + \mathbf{1}_{N,d}\sigma(x^\star)\right) \\ &- \nabla\phi(\tilde{x} + x^\star)\nabla_2 f\left(\tilde{x} + x^\star, \begin{bmatrix} R_d & 0 \end{bmatrix} \psi + \Delta\phi(\tilde{x}) + \mathbf{1}_{N,d}\sigma(x^\star)\right) \\ &+ \nabla\phi(\tilde{x} + x^\star)\nabla_2 f\left(x^\star, \mathbf{1}_{N,d}\sigma(x^\star)\right) \\ &- \nabla\phi(\tilde{x} + x^\star)\frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N}\nabla_2 f\left(x^\star, \mathbf{1}_{N,d}\sigma(x^\star)\right) - \nabla\phi(\tilde{x} + x^\star)\begin{bmatrix} 0 & R_d \end{bmatrix}\psi, \end{aligned}$$

where the matrix R_d *has the same meaning as in* (3.64), *and*

$$\begin{aligned} g(\tilde{x}, \psi) := &\begin{bmatrix} -R_d^\top L R_d & 0 \\ 0 & -R_d^\top L R_d \end{bmatrix}\psi + \begin{bmatrix} 0 \\ R_d^\top L\nabla_2 f\left(x^\star, \mathbf{1}_{N,d}\sigma(x^\star)\right) \end{bmatrix} \\ &- \begin{bmatrix} R_d^\top L\Delta\phi(\tilde{x}) \\ R_d^\top L\nabla_2 f\left(\tilde{x} + x^\star, \begin{bmatrix} R_d & 0 \end{bmatrix}\psi + \Delta\phi(\tilde{x}) + \mathbf{1}_{N,d}\sigma(x^\star)\right) \end{bmatrix}, \end{aligned}$$

with $\Delta\phi(\tilde{x}) := \phi(\tilde{x} + x^\star) - \phi(x^\star)$. △

System (3.98) has a structure known in the literature as continuous-time singularly perturbed system (see [91, Chapter 11]), i.e., the continuous-time counterpart of the discret-time systems investigated in Appendix C. In particular, as in the discret-time case, we denote the subsystem (3.98a) as the slow system, and (3.98b) as the fast one. See Figure 3.6 for a schematic representation of (3.98).

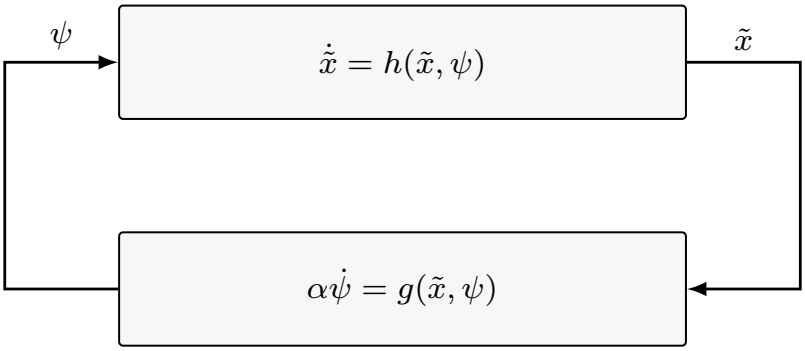


Figure 3.6: Singularly perturbed system (3.98).

The next step consists in "freezing" the slow state $\tilde{x}$ within the fast system (3.98b). After this operation, the boundary layer system associated to the interconnection (3.98) can be built and studied as follows.

Lemma 3.14. *Consider system* (3.98b) *and let* $\tilde{x} \in \mathbb{R}^n$ *be fixed. Consider*

$$\bar{\psi}(\tilde{x}) = \begin{bmatrix} -R_d^\top \Delta\phi(\tilde{x}) \\ R_d^\top \left(\nabla_2 f\left(x^\star, \mathbf{1}_{N,d}\sigma(x^\star)\right) - \nabla_2 f\left(\tilde{x} + x^\star, \mathbf{1}_{N,d}\sigma(\tilde{x} + x^\star)\right)\right) \end{bmatrix}.$$

Then, for any $\tilde{x} \in \mathbb{R}^n$, $\bar{\psi}(\tilde{x})$ *is an equilibrium for* (3.98b). *Let* $\xi := \psi - \bar{\psi}(\tilde{x})$ *and write* (3.98b) *as the boundary layer system*

$$\dot{\xi} = g(\tilde{x}, \xi + \bar{\psi}(\tilde{x})). \tag{3.99}$$

Then, there exists a function $U : \mathbb{R}^m \to \mathbb{R}$ *so that*

$$b_1 \|\xi\|^2 \leq U(\xi) \leq b_2 \|\xi\|^2 \tag{3.100a}$$

$$\frac{\partial U(\xi)}{\partial x} g(x, \xi + \bar{\psi}(x)) \leq -b_3 \|\xi\|^2 \tag{3.100b}$$

$$\left\| \frac{\partial U(\xi)}{\partial \xi} \right\| \leq b_4 \|\xi\| \tag{3.100c}$$

$$\frac{\partial U(\xi)}{\partial \tilde{x}} = 0, \tag{3.100d}$$

for all $\xi \in \mathbb{R}^m$ *and for some constants* $b_1, b_2, b_3, b_4 > 0$. △

Remark 3.4. In view of Assumption 3.11, it is possible to show that the matrix $-R_d^\top L R_d$ is Hurwitz. Then, by explicitly combining the definitions of g and $\bar{\psi}$, system (3.99) reads as a stable linear system perturbed with a vanishing perturbation. △

Once $\bar{\psi}(\tilde{x})$ has been found, we can use it to build the reduced system and study the stability of its origin.

Lemma 3.15. *Consider the reduced system defined as*

$$\dot{\tilde{x}} = h(\tilde{x}, \bar{\psi}(\tilde{x})). \tag{3.101}$$

Then, the origin is a globally exponentially stable equilibrium point for system (3.101). △

Remark 3.5. By explicitly combining the definitions of h and $\bar{\psi}$, the dynamics (3.101) reads as the gradient flow related to problem (3.55) in error coordinates with respect to $x^\star$. Hence, the proof follows from Assumption 3.12. △

Once the global exponential stability of the origin for both the boundary layer and reduced system has been proved, we can establish the convergence properties of system (3.97)

Theorem 3.4. *Consider system* (3.97). *Let Assumptions 3.10, 3.11, and 3.12 hold and pick intial conditons* $(x(0), w(0), z(0))$ *such that* $\mathbf{1}_{N,d}^\top w(0) = \mathbf{1}_{N,d}^\top z(0) = 0$. *Then, there exist* $\bar{\alpha} > 0$, $a_1 > 0$, *and* $a_2 > 0$ *such that, for all* $\alpha \in (0, \bar{\alpha})$, *it holds*

$$\|x_i - x_i^\star\| \leq a_1 \exp(-a_2 t),$$

for all $i \in \{1, \dots, N\}$, *where* $x_i^\star \in \mathbb{R}^{n_i}$ *is the* i*-th block of the optimal solution* $x^\star \in \mathbb{R}^n$ *of problem* (3.55).

Proof. By performing the change of variables introduced in Lemma 3.13, we equivalently rewrite system (3.63) as

$$\dot{\tilde{x}} = h(\tilde{x}, \psi), \tag{3.102a}$$

$$\alpha\dot{\psi} = g(\tilde{x}, \psi), \tag{3.102b}$$

with $\tilde{w}$, ψ, h, and g with same meaning as in Lemma 3.13. Lemma 3.14 and Lemma 3.15 respectively ensures the global exponential stability of the origin for the boundary layer system and for the reduced system associated to (3.102). Moreover, Assumption 3.10 ensures that functions h, g, and $\bar{\psi}$ are Lipschitz continuous. Hence, with the arguments of [91, Theorem 11.4], we claim that there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha})$ the origin is a globally exponentially stable equilibrium point for system (3.102), i.e., it holds

$$\|\text{col}(\tilde{x}, \psi)\| \leq a_3 \|\text{col}(\tilde{x}(0), \psi(0))\| \exp(-a_2 t),$$

for some $a_2, a_3 > 0$. The proof follows by the trivial fact $\|x_i - x_i^\star\| \leq \|\tilde{x}\| \leq \|\text{col}(\tilde{x}, \psi)\|$ and by setting $a_1 = a_3 \|\text{col}(\tilde{x}(0), \psi(0))\|$. ■

We point out that [91, Theorem 11.4] only provides semi-global exponential stability.

However, it can be shown that the additional condition (3.100d) can be used to prove the global exponential stability.

3.4.3 Numerical Simulations

In this section, we employ our Aggregative Tracking Feedback to address a multi-robot surveillance scenario. We consider a network of N mobile robots, whose planar position is $x_i \in \mathbb{R}^2$, that aim to surveil a collection of N intruders. In particular, each robot i of the surveillance team is associated to an intruder located at $y_i \in \mathbb{R}^2$. Given the orientation $\theta_i \in \mathbb{R}$ of the robot i, we describe its dynamics through the unicycle model

$$\dot{x}_i = \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} v_i \tag{3.103a}$$

$$\dot{\theta}_i = \omega_i, \tag{3.103b}$$

where $v_i, \omega_i \in \mathbb{R}$ are the low-level inputs denoting the linear and the angular speed, respectively. Let $u_i \in \mathbb{R}^2$ be a reference position, then [183] proposes the following low-level controller

$$v_i(x_i, \theta_i, u_i) = k_i \left\| x_i - u_i \right\| \cos(\theta_{i,\text{err}}(x_i, \theta_i)) \tag{3.104a}$$

$$\begin{aligned} \omega_i(x_i, \theta_i, u_i) = {} & \tfrac{k_i}{\|x_i - u_i\|} \cos(\theta_{i,\text{err}}(x_i, \theta_i)) \sin(\theta_{i,\text{err}}(x_i, \theta_i)) \\ & + \tfrac{k_i}{\|x_i - u_i\|} \sin(\theta_{i,\text{err}}(x_i, \theta_i)), \end{aligned} \tag{3.104b}$$

with $k_i > 0$ and $\theta_{i,\text{err}}(x_i, \theta_i) = \text{atan2}(x_{i,1}, x_{i,2}) - \theta_i$, where $x_{i,1}$ and $x_{i,2}$ are the components of x_i, i.e., we write $x_i := \text{col}(x_{i,1}, x_{i,2})$. Thus, the overall closed-loop dynamics reads as

$$\dot{x}_i = \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} v_i(x_i, \theta_i, u_i) \tag{3.105a}$$

$$\dot{\theta}_i = \omega_i(x_i, \theta_i, u_i). \tag{3.105b}$$

As shown in [183, Lemma 2.1], for any reference $u_i \in \mathbb{R}^2$, the point $\text{col}(u_i, 0)$ is an almost globally asymptotically stable equilibrium point for (3.105). Moreover, the trajectories of (3.105) exponentially converge to u_i (cf. [183, Lemma 2.1]). Thus, system (3.105) satisfies Assumption 3.9, namely it has a steady-state map $h_i(u_i) = u_i$ with exponential convergence guarantees.

As for the environment, we consider a non convex scenario in which altitude changes and $n_c \in \mathbb{N}$ crevasses are present. Let $\text{col}(\ell_1, \ell_2)$ be the planar coordinates describing a given location. Then, we model the altitude profile of the environment through a function $z_{\text{alt}} : \mathbb{R}^2 \to \mathbb{R}$ given by the sum of a sinusoidal term and a series of gaussian

functions modeling the crevasses, namely

$$z_{\text{alt}}(\ell_1, \ell_2) = -a_1 \cos(\rho \ell_1) \sin(\rho \ell_2) - \sum_{g=1}^{n_c} a_{c,g} \exp\left(-\frac{(\ell_1 - \mu_{g,1})^2 + (\ell_2 - \mu_{g,2})^2}{s_g} \right), \tag{3.106}$$

where $a_1, \rho > 0$ are respectively the amplitude and the frequency of the sinusoidal term, while $a_{c,1}, \dots, a_{c,n_c}, s_1, \dots, s_{n_c} > 0$ are the parameters characterizing the gaussian functions whose respective centers are located in $(\mu_{1,1}, \mu_{1,2}), \dots, (\mu_{n_c,1}, \mu_{n_c,2})$. It is worth noting that this environment profile as well as the nonlinear dynamics give rise to a nonconvex optimization problem.

The surveillance strategy of the team consists of a trade-off between the following competing objectives: each robot (i) tries to stay close to the intruder, (ii) tries to occupy locations with higher altitudes, and (iii) tries to stay close to the weighted center of mass. This scenario falls into the distributed aggregative feedback optimization framework. Specifically, in problem (3.55), we choose the objective function f_i of each agent $i \in \{1, \dots, N\}$ as

$$f_i(x_i, \sigma(x)) = \frac{\gamma_1}{2} \|x_i - y_i\|^2 - z_{\text{alt}}(x_{i,1}, x_{i,2}) + \frac{\gamma_2}{2} \|x_i - \sigma(x)\|^2, \tag{3.107}$$

where $\gamma_1, \gamma_2 > 0$, while the term $-z_{\text{alt}}(x_i)$ increases the cost according to the altitude of the location x_i (cf. (3.106)). Further, we choose our aggregative variable as the weighted center of mass of the defending team

$$\sigma(x) = \frac{1}{N} \sum_{i=1}^{N} \beta_i x_i, \tag{3.108}$$

for some weights $\beta_i > 0$. In particular, we set $N = 6$, $\gamma_1 = 1$, $\gamma_2 = 0.3$, $n_c = 5$, and we randomly generate the weights $\beta_1, \dots, \beta_N$ within the interval $(0, 1)$, the amplitudes $a_{c,1}, \dots, a_{c,n_c}$ within the interval $[0, 5]$, the terms $s_1, \dots, s_{n_c}$ within the interval $(5, 10)$, and the locations $\mu_1 := \text{col}(\mu_{1,1}, \mu_{1,2}), \dots, \mu_{n_c} := \text{col}(\mu_{n_c,1}, \mu_{n_c,2})$, $y_1, \dots, y_N$, and b within the interval $[0, 100]^2$. As regards the sinusoidal terms, we choose $a_1 = 10$ and $\rho = 0.02$, while, as for the algorithm parameters, we set $\alpha_1 = 0.75$, $\alpha_2 = 0.01$, while the initial conditions $x_i(0)$ and $u_i(0)$ are randomly selected. As predicted by Theorem 3.3, Fig. 3.7 shows that the optimality error $\|\nabla F(x(t), \sigma(x(t)))\|$ asymptotically converges to 0.

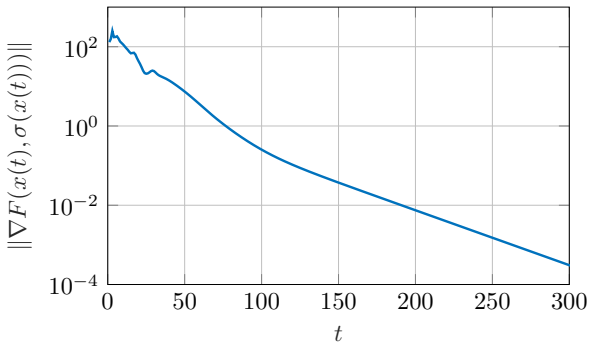
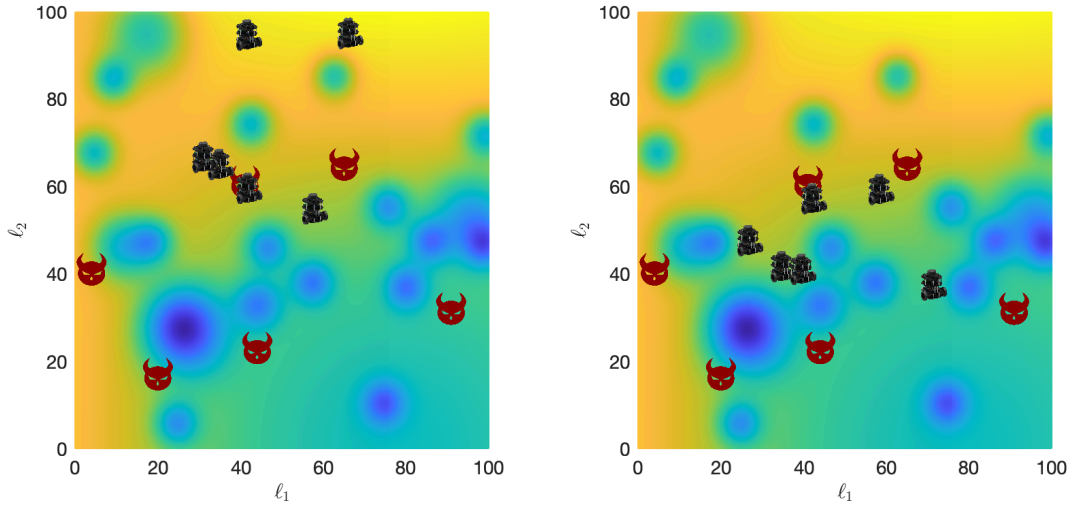


Figure 3.7: Optimality error evolution.

Considering the same simulation, Fig. 3.8 provides the initial and final configuration of the team. Each robot icon denotes an agent of the surveillance team, while each devil icon denotes an intruder. The color of the background represents the altitude: blue background denotes the lowest locations, while yellow background denotes the highest ones.



(a) Initial configuration.

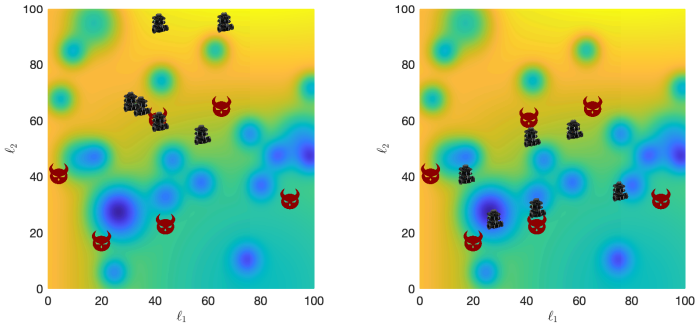
(b) Final configuration.

Figure 3.8: Multi-robot surveillance: nonconvex scenario.

Fig. 3.8 highlights the role played by the altitude in determining the final configuration achieved by the agents. Indeed, some of the robots remain far from their associated intruders because closer locations would have lower altitudes. In order to emphasize this aspect, we perform the same simulations without taking into account the altitude z_{alt} in the cost, i.e., by considering the problem in which each objective function reads as

$$f_i(x_i, \sigma(x)) = \frac{\gamma_1}{2} \|x_i - y_i\|^2 + \frac{\gamma_2}{2} \|x_i - \sigma(x)\|^2 , \tag{3.109}$$

for all $i \in \{1, \dots, N\}$. Fig. 3.9 provides the initial and final team configuration of such a simulation. In Fig. 3.9, differently from the case inspected in Fig. 3.8, the robots go closer to their associated intruders thus occupying locations with low altitudes.



(a) Initial configuration. (b) Final configuration.

Figure 3.9: Multi-robot surveillance: strongly convex scenario.

Finally, we note that in both cases the robots arrange themselves inside the polygon whose vertices coincide with the positions occupied by the intruder. In fact, the outer configurations at the same (i) distance from the invaders and (ii) altitude suffer from a higher cost due to the presence of the aggregative term $\|x_i - \sigma(x)\|^2$.

Finally, we address the same strongly convex problem with objective functions (3.109) for the case with single integrator dynamics (3.95) to test the effectiveness of (3.97). We choose the parameters of the problem as above and set $\alpha = 0.1$. As predicted by Theorem 3.3, Figure 3.7 shows an exponential decay of the optimality error $\|x(t) - x^\star\|$, where $x^\star$ is the minimizer of the problem computed by a centralized solver.

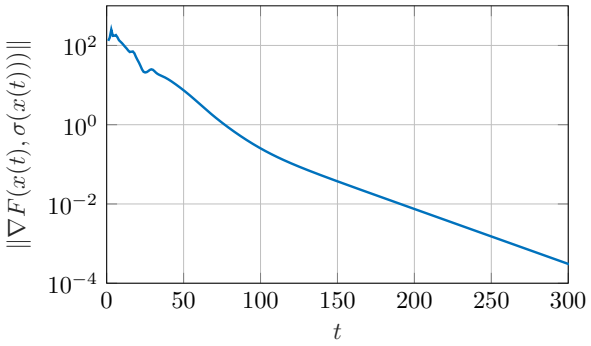


Figure 3.10: Optimality error evolution.

Chapter 4

Tracking-Based Distributed Equilibrium Seeking Algorithms for Aggregative Games

In this chapter, we focus on the development of fully-distributed algorithm for Nash equilibrium seeking in aggregative games over networks, i.e., the scenario already introduced in Section 1.4.

In particular, we first consider the case where only local constraints are present and we design an algorithm combining, for each agent, (i) the projected pseudo-gradient descent and (ii) a tracking mechanism to locally reconstruct the aggregative variable. To handle coupling constraints arising in generalized settings, we propose another distributed algorithm based on (i) a recently emerged augmented primal-dual scheme and (ii) two tracking mechanisms to reconstruct, for each agent, both the aggregative variable and the coupling constraint satisfaction. Leveraging tools from singular perturbations analysis, we prove linear convergence to the Nash equilibrium for both schemes. The results of this chapter are based on [27].

4.1 Literature Review

Recent years have seen an increasing attention to the computation of (generalized) Nash equilibria in games over networks [59, 125, 167]. Indeed, numerous applications falling within different domains such as smart grids management [8, 131], economic market analysis [143], cooperative control of robots [58], electric vehicles charging [36, 53, 66], network congestion control [7], and synchronization of coupled oscillators in power grids [208] can be modelled as networks of selfish agents – aiming at optimizing their strategy according to an associated individual cost function – that compete with each other over shared resources.

Among these examples, one can often find instances modelled as an aggregative game, where the strategies of all the agents in the network are coupled through the so-called aggregative variable (expressing, e.g., the mean strategy), upon which each agent's cost function depends; see, e.g., [12, 87, 152] for a comprehensive overview. This chapter investigates such a framework proposing novel distributed algorithms for generalized Nash equilibrium (GNE) seeking under *partial information*, i.e., assuming that each agent is only aware of its own local information (e.g., its strategy set and cost function) and can communicate only with few agents in the network. This restriction naturally calls for the design of fully-distributed mechanisms for GNE seeking.

We recall the difference with respect to the distributed aggregative optimization framework, where agents in a network collaborate to minimize the sum of individual objective functions depending both on local decision variables and an aggregative variable, see Chapter 3.

In the context of NE problems in aggregative form, first attempts to design equilibrium seeking algorithms involve semi-decentralized approaches in which a central entity gathers and shares global quantities (such as the aggregative variable and/or a dual multiplier) with all the agents [9, 10, 51, 74, 75, 90, 147, 205].

To relax the communication requirements, [95] proposes a gradient-based algorithm for non-generalized games with diminishing step-size that relies on dynamic average consensus to reconstruct the aggregative variable in each agent. Such a method has been refined in [204] to deal with privacy issues and, as a consequence, only guaranteeing approximate equilibrium computations. In [151], the distributed computation of an approximate Nash equilibrium is guaranteed through a best-response-based algorithm requiring multiple communication exchanges per iteration. In [35], instead, an asynchronous distributed algorithm based on proximal dynamics is proposed.

Looking at GNE problems where the agents' strategies are coupled also by means of constraints, in [150] the distributed computation of an approximate NE is guaranteed through an algorithm requiring, however, several communication exchanges per iteration. Exact convergence is instead guaranteed in [11], where a distributed algorithm with diminishing step-size is proposed, combining dynamic tracking mechanisms, monotone operator splitting, and the Krasnosel'skii-Mann fixed-point iteration. An exactly convergent distributed equilibrium-seeking algorithm with constant step-size is given in [69], where the authors propose a distributed method based on a forward-backward splitting of two preconditioned operators requiring a double communication exchange per iteration.

4.2 Distributed Aggregative Games over Networks

We recall as follows the formalization of the aggregative games over networks already given in Section 1.4.

We consider a population of $N \in \mathbb{N}$ agents who, given all other agents' strategies, aim at finding a local strategy solving the optimization problem:

$$\forall i \in \{1,\dots,N\} : \begin{cases} \min\limits_{x_i \in X_i} & J_i(x_i, \sigma(x)) \\ \text{s.t.} & A_i x_i + \sum\limits_{j \in \{1,\dots,N\}\setminus\{i\}} A_j x_j \leq \sum\limits_{i \in \{1,\dots,N\}} b_i, \end{cases} \tag{4.1}$$

where $X_i \subseteq \mathbb{R}^{n_i}$, $A_i \in \mathbb{R}^{m \times n_i}$, and $b_i \in \mathbb{R}^m$ model the feasible strategy set for agent i, while the cost function $J_i : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}$ depends on the i-th individual strategy $x_i \in \mathbb{R}^{n_i}$, as well as on the *aggregative variable* $\sigma(x) \in \mathbb{R}^d$, with $x \coloneqq \operatorname{col}(x_1,\dots,x_N) \in \mathbb{R}^n$, $n \coloneqq \sum_{i=1}^N n_i$. We consider $m \leq n$. As in Chapter 3, the aggregative variable is given by $\sigma(x) \frac{1}{N} \sum_{i=1}^N \phi_i(x_i)$, where each *aggregation rule* $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}^d$ models the contribution of the corresponding strategy x_i to the aggregate $\sigma(x)$. We recall the constraint functions (see (1.10)) $c_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$, $c_{-i} : \mathbb{R}^{n-n_i} \to \mathbb{R}^m$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ as follows

$$c_i(x_i) = A_i x_i - b_i, \tag{4.2a}$$

$$c_{-i}(x_{-i}) = \sum_{j \in \{1,\dots,N\}\setminus\{i\}} (A_j x_j - b_j), \tag{4.2b}$$

$$c(x) = c_i(x_i) + c_{-i}(x_{-i}) = Ax - b, \tag{4.2c}$$

where $x_{-i} \coloneqq \operatorname{col}(x_1,\dots,x_{i-1},x_{i+1},\dots,x_N) \in \mathbb{R}^{n-n_i}$, $A \coloneqq [A_1 \dots A_N] \in \mathbb{R}^{m \times n}$, and $b \coloneqq \sum_{i=1}^N b_i$. Then, the collective vector of strategies x belongs to the feasible set $\mathcal{C} \coloneqq \{\mathrm{x} \in X \mid c(x) \leq 0\} \subseteq \mathbb{R}^n$, where $X \coloneqq \prod_{i=1}^N X_i \subseteq \mathbb{R}^n$.

The goal of this chapter is to develop fully-distributed schemes to compute the GNE of (4.1), see Section 1.4 for further details about the mathematical definition of a GNE.

Next, we formalize customary assumptions that establish the regularity of some local quantities in (4.1).

Assumption 4.1 (Local feasible sets and cost functions)**.** *For all $i \in \{1,\dots,N\}$, we have that:*

(i) The feasible set X_i is nonempty, closed, and convex;

(ii) The function $J_i(\cdot, \phi_i(\cdot)/N + \sigma_{-i}(x_{-i}))$ is of class $\mathcal{C}^1$, i.e., its derivative exists and is continuous, for all $x_{-i} \in \mathbb{R}^{n-n_i}$. △

A key device in this game-theoretic framework is the so-called *pseudo-gradient*

mapping $F : \mathbb{R}^n \to \mathbb{R}^n$:

$$F(x) \coloneqq \operatorname{col}(\nabla_{x_1} J_1(x_1, \sigma(x)), \dots, \nabla_{x_N} J_N(x_N, \sigma(x))). \tag{4.3}$$

With this regard, we also make the following assumption.

Assumption 4.2 (Strong monotonicity and Lipschitz continuity)**.** *F is μ-strongly monotone, i.e., there exists $\mu > 0$ such that*

$$(F(x) - F(y))^\top (x - y) \geq \mu \|x - y\|^2,$$

for any $x, y \in \mathbb{R}^n$. Moreover, given any $x_i, x_i' \in \mathbb{R}^{n_i}$ and $y, y' \in \mathbb{R}^{n-n_i}$, for all $i \in \{1, \dots, N\}$, we assume that

$$\begin{aligned}
\|\nabla_{x_i} J_i(x_i, \phi_i(x_i)/N + y) - \nabla_{x_i'} J_i(x_i', \phi_i(x_i')/N + y')\| \\
\leq \overline{L}_1 \left\| \operatorname{col}(x_i, y) - \operatorname{col}(x_i', y') \right\|, \\
\left\| \nabla_1 J_i(x_i, y) - \nabla_1 J_i(x_i', y') \right\| \leq \overline{L}_1 \left\| \operatorname{col}(x_i, y) - \operatorname{col}(x_i', y') \right\|, \\
\left\| \nabla_2 J_i(x_i, y) - \nabla_2 J_i(x_i', y') \right\| \leq \overline{L}_2 \left\| \operatorname{col}(x_i, y) - \operatorname{col}(x_i', y') \right\|, \\
\left\| \phi_i(x_i) - \phi_i(x_i') \right\| \leq \overline{L}_3 \left\| x_i - x_i' \right\|.
\end{aligned}$$

△

While assumptions on strong monotonicity and Lipschitz continuity of the game mapping are quite standard in the literature [12, 147, 205], in the second part of Assumption 4.2 we further specialize the Lipschitz properties of the gradients of the cost functions in both the local and aggregate variables, as well as of each single aggregation rule $\phi_i(\cdot)$.

Note that we assume partial information, i.e., each agent i is only aware of its own local information x_i, J_i, ϕ_i, X_i, A_i, and b_i. Moreover, each agent can exchange information with a subset of $\{1, \dots, N\}$ only. Specifically, we consider a network of agents whose communication is performed according to a directed graph $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$, with $\mathcal{E} \subset \{1, \dots, N\}^2$. The following assumption characterizes the communication graphs considered:

Assumption 4.3 (Network)**.** *The graph $\mathcal{G}$ is strongly connected and the matrix $\mathcal{W}_\mathcal{G}$ is doubly stochastic.* △

4.2.1 Primal TRADES: Algorithm Description and Analysis

In this section we introduce and analyze Primal TRacking-based Aggregative Distributed Equilibrium Seeking (TRADES), a fully-distributed iterative NE seeking algorithm for

aggregative games given by (4.1) without coupling constraints, i.e.,

$$\forall i \in \{1, \dots, N\} : \min_{x_i \in X_i} J_i(x_i, \sigma(x)). \tag{4.4}$$

Our distributed algorithm is then able to steer the strategies of the network to a NE of the game.

The proposed scheme is iterative with k denoting the iteration index. Let $\mathrm{x}_i^k \in \mathbb{R}^{n_i}$ be the strategy chosen by each agent i at iteration $k \geq 0$. Taking its convex combination with a projected pseudo-gradient step may be an effective way to steer each agent's strategy to the best response $x_{i,\mathrm{br}}(\sigma_{-i}(\mathrm{x}_{-i}^k))$. When applied to problem (4.4), it reads as

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k + \delta \left(P_{X_i} \left[\mathrm{x}_i^k - \gamma \left(\nabla_{x_i} J_i(\mathrm{x}_i^k, \sigma(\mathrm{x}^k)) \right) \right] - \mathrm{x}_i^k \right), \tag{4.5}$$

where $\delta \in (0, 1)$ is a constant performing the combination and $\gamma > 0$ plays the role of the gradient step-size. We point out that the chain rule and the definition of $\sigma(\mathrm{x}^k)$ (cf. (1.9)) lead to $\nabla_{x_i} J_i(\mathrm{x}_i^k, \sigma(\mathrm{x}^k)) = \nabla_1 J_i(\mathrm{x}_i^k, \sigma(\mathrm{x}^k)) + \frac{\nabla \phi_i(\mathrm{x}_i^k)}{N} \nabla_2 J_i(\mathrm{x}_i^k, \sigma(\mathrm{x}^k))$. In our distributed setting, however, agent i cannot access the global aggregate variable $\sigma(\mathrm{x}^k)$. To compensate this lack of information, we rely on the locally available $\phi_i(\mathrm{x}_i^k)$ and the auxiliary variable $\mathrm{z}_i^k \in \mathbb{R}^d$. Thus, for all $i \in \{1, \dots, N\}$, we introduce the operator $\tilde{F}_i : \mathbb{R}^{n_i} \times \mathbb{R}^d \to \mathbb{R}^{n_i}$ defined as

$$\tilde{F}_i(x_i, \mathrm{s}) \coloneqq \nabla_1 J_i(x_i, s) + \frac{\nabla \phi_i(x_i)}{N} \nabla_2 J_i(x_i, s),$$

and, in accordance, we modify the update (4.5) as

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k + \delta \left(P_{X_i} \left[\mathrm{x}_i^k - \gamma \tilde{F}_i \left(\mathrm{x}_i^k, \phi_i(\mathrm{x}_i^k) + \mathrm{z}_i^k \right) \right] - \mathrm{x}_i^k \right), \tag{4.6}$$

which can be directly implemented without violating the distributed nature of the algorithm. In case

$$\mathrm{z}_i^k \to -\phi_i(\mathrm{x}_i^k) + \sigma(\mathrm{x}^k), \tag{4.7}$$

then the implementable law (4.6) coincides with the desired one given in (4.5). Note that z_i^k encodes the estimate of $\sigma(\mathrm{x}_i^k) - \phi_i(\mathrm{x}_i^k)$, i.e., the aggregate of all other agents' strategies except for the i-th one. For this reason, we update each auxiliary variable z_i^k according to the following causal version of the perturbed average consensus scheme (see (2.4), where it has been used to locally compensate the missing knowledge of the global gradient in the distributed consensus optimization setting):

$$\mathrm{z}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \mathrm{z}_j^k + \sum_{j \in \mathcal{N}_i} w_{ij} \phi_j(\mathrm{x}_j^k) - \phi_i(\mathrm{x}_i^k). \tag{4.8}$$

This is implementable in a fully-distributed fashion since it only requires communi-

cation with neighboring agents $j \in \mathcal{N}_i$. We report the whole algorithmic structure in Algorithm 8 and, from now on, we will refer to it as Primal TRacking-based Aggregative Equilibrium Seeking (TRADES)

Algorithm 8 Primal TRADES (Agent i)

Initialization: $\mathrm{x}_i^0 \in X_i, \mathrm{z}_i^0 = 0$.

for $k = 1, 2, \dots$ **do**

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k + \delta \left(P_{X_i} \left[\mathrm{x}_i^k - \gamma \tilde{F}_i \left(\mathrm{x}_i^k, \phi_i(\mathrm{x}_i^k) + \mathrm{z}_i^k \right) \right] - \mathrm{x}_i^k \right) \tag{4.9a}$$

$$\mathrm{z}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \mathrm{z}_j^k + \sum_{j \in \mathcal{N}_i} w_{ij} \phi_j(\mathrm{x}_j^k) - \phi_i(\mathrm{x}_i^k). \tag{4.9b}$$

end for

We note that Algorithm 8 requires the initialization $\mathrm{z}_i^0 = 0$ for all $i \in \{1, \dots, N\}$; we will discuss in the sequel the interpretation of this particular initialization. The local update (4.9) leads to the stacked vector form of Primal TRADES, namely

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta \left(P_X \left[\mathrm{x}^k - \gamma \tilde{F} \left(\mathrm{x}^k, \phi(\mathrm{x}^k) + \mathrm{z}^k \right) \right] - \mathrm{x}^k \right), \tag{4.10a}$$

$$\mathrm{z}^{k+1} = \mathcal{W}_d \mathrm{z}^k + (\mathcal{W}_d - I) \phi(\mathrm{x}^k), \tag{4.10b}$$

with $\mathcal{W}_d \coloneqq \mathcal{W}_{\mathcal{G}} \otimes I_d \in \mathbb{R}^{Nd}$, $\mathrm{z}^k \coloneqq \mathrm{col}(\mathrm{z}_{1,k}, \dots, \mathrm{z}_{N,k})$, $\phi(\mathrm{x}^k) \coloneqq \mathrm{col}(\phi_1(\mathrm{x}_1^k), \dots, \phi_N(\mathrm{x}_N^k))$, and $\tilde{F}(\mathrm{x}^k, \phi(\mathrm{x}^k) + \mathrm{z}^k) \coloneqq \mathrm{col}(\tilde{F}_1(\mathrm{x}_1^k, \phi_1(\mathrm{x}_1^k) + z_1^k), \dots, \tilde{F}_N(\mathrm{x}_N^k, \phi_N(\mathrm{x}_N^k) + z_N^k))$. We establish next the properties of Primal TRADES in computing the NE of problem (4.4).

Theorem 4.1. *Consider the dynamics in* (4.10). *There exist constants* $\bar{\delta}, \bar{\gamma}, a_1, a_2 > 0$ *such that, for any* $\delta \in (0, \bar{\delta})$, $\gamma \in (0, \bar{\gamma})$ *and* $(\mathrm{x}^0, \mathrm{z}^0) \in \mathbb{R}^{n+Nd}$ *such that* $\mathbf{1}_{N,d}^\top \mathrm{z}^0 = 0$, *it holds*

$$\left\| \mathrm{x}^k - x^\star \right\| \leq a_1 \exp(-a_2 k). \qquad \triangle$$

The proof of Theorem 4.1 relies on a *singular perturbation* analysis of system (4.10), and will be given in the next subsection.

Proof of Theorem 4.1

We build the framework to prove Theorem 4.1 by analyzing (4.10) under a singular perturbations lens. We therefore establish the related proof in five steps:

1. *Bringing* (4.10) *in the form of* (C.18)*:* We leverage the initialization z^0 so that

$\mathbf{1}_{N,d}^\top z^0 = 0$ to introduce coordinates $\bar{z} \in \mathbb{R}^d$ and $z_\perp \in \mathbb{R}^{(N-1)d}$ defined as:

$$\begin{bmatrix} \bar{z} \\ z_\perp \end{bmatrix} := \begin{bmatrix} \frac{\mathbf{1}_{N,d}^\top}{N} \\ R_d^\top \end{bmatrix} z \implies z = \mathbf{1}_{N,d}\bar{z} + R_d z_\perp, \tag{4.11}$$

where $R_d \in \mathbb{R}^{Nd\times(N-1)d}$ with $\|R_d\| = 1$ is such that

$$R_d R_d^\top = I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N} \text{ and } R_d^\top \mathbf{1}_{N,d} = 0. \tag{4.12}$$

Then, by using the definition of $\bar{z}$ given in (4.11), the associated dynamics reads as

$$\begin{aligned} \bar{z}^{k+1} = \frac{\mathbf{1}_{N,d}^\top}{N} z^{k+1} &\overset{(a)}{=} \frac{\mathbf{1}_{N,d}^\top}{N}\mathcal{W}_d z^k + \frac{\mathbf{1}_{N,d}^\top}{N}(\mathcal{W}_d - I)\phi(x^k) \\ &\overset{(b)}{=} \frac{\mathbf{1}_{N,d}^\top}{N} z^k \overset{(c)}{=} \frac{\mathbf{1}_{N,d}^\top}{N}\left(\mathbf{1}_{N,d}\bar{z}^k + R_d z_\perp^k\right) \overset{(d)}{=} \bar{z}^k, \end{aligned} \tag{4.13}$$

where in (a) we exploit the update (4.10), in (b) we use the facts that, in view of Assumption 4.3, (i) $\mathbf{1}_{N,d}^\top \mathcal{W}_d = \mathbf{1}_{N,d}^\top$ and (ii) $\mathbf{1}_{N,d}^\top(\mathcal{W}_d - I) = 0$, in (c) we rewrite z^k according to (4.11), and in (d) we use the fact that $\mathbf{1}_{N,d}^\top R_d = 0$. Thus, (4.13) leads to $\bar{z}^{k+1} \equiv \bar{z}^0 \equiv 0$ for all $k \geq 0$, where the last equality follows by the initialization $\mathbf{1}_{N,d}^\top z^0 = 0$ and the definition of $\bar{z}$ (cf. (4.11)). We are thus entitled to ignore the null dynamics of $\bar{z}^k$ and, according to (4.11), we equivalently rewrite (4.10) as

$$x^{k+1} = x^k + \delta\left(P_X\left[x^k - \gamma\tilde{F}(x^k, \phi(x^k) + R_d z_\perp^k)\right] - x^k\right), \tag{4.14a}$$

$$z_\perp^{k+1} = R_d^\top \mathcal{W}_d R_d z_\perp^k + R_d^\top(\mathcal{W}_d - I)\phi(x^k). \tag{4.14b}$$

For any $k \geq 0$, the interconnected system (4.14) can be seen as singularly perturbed system in the generic form of (C.18) (in Appendix C) by setting

$$\begin{aligned} w^t &:= z_\perp^k, \\ f(x^k, w^k) &:= P_X\left[x^k - \gamma\tilde{F}(x^k, \phi(x^k) + R_d w^k)\right] - x^k, \\ g(w^k, x^k) &:= R_d^\top \mathcal{W}_d R_d w^k + R_d^\top(\mathcal{W}_d - I)\phi(x^k). \end{aligned} \tag{4.15}$$

In particular, we refer to the subsystem (4.14a) as the slow system, while we refer to (4.14b) as the fast one.

2. Equilibrium function h : For any $x^k \in \mathbb{R}^n$, under the expression for $R_d R_d^\top$ in (4.12) and since $\mathcal{W}_\mathcal{G}$ is doubly stochastic (cf. Assumption 4.3) notice that for any $x^k = x \in \mathbb{R}^n$,

$$z_\perp = h(x) := -R_d^\top \phi(x) \tag{4.16}$$

constitutes an equilibrium of (4.14b). Since $R_d^\top \mathcal{W}_d R_d$ is Schur in view of Assumption 4.3,

we interpret (4.14b) as a strictly stable linear system with nonlinear input $R_d^\top(\mathcal{W}_d - I)\phi(\mathrm{x}^k)$ parametrizing the equilibrium of the subsystem. The role of γ is to slow down the variation of x^k so that the stability of $h(\mathrm{x}^k)$ for (4.14b) is preserved.

3. Boundary layer system and satisfaction of (C.21)*:* The so-called boundary layer system associated to (4.14) can be constructed by fixing $\mathrm{x}^k = x$ for all $k \geq 0$, for some arbitrary $x \in \mathbb{R}^n$ in (4.14b), and rewriting it according to the error coordinates $\tilde{\mathrm{z}}^k := \mathrm{z}_\perp^k - h(\mathrm{x}^k)$. Using (4.12), we obtain that

$$\tilde{\mathrm{z}}^{k+1} = R_d^\top \mathcal{W}_d R_d \tilde{\mathrm{z}}^k. \tag{4.17}$$

Notice that the latter is in the form of (C.20) (in Appendix C) with $\psi = \tilde{\mathrm{z}}^k$, and $g(\psi + h(\mathrm{x}), \mathrm{x}) - h(\mathrm{x}) = R_d^\top \mathcal{W}_d R_d \tilde{\mathrm{z}}^k$. The next lemma provides a Lyapunov function for (4.17).

Lemma 4.1. *Consider system* (4.17)*. Then, there exists a continuous function* $U : \mathbb{R}^{(N-1)d} \to \mathbb{R}$ *satisfying* (C.21) *(in Appendix C) with* $\tilde{\mathrm{z}}$ *in place of* ψ*.*

Proof. System (4.17) is a linear autonomous system whose state matrix $R_d^\top \mathcal{W}_d R_d \in \mathbb{R}^{(N-1)d\times(N-1)d}$ is Schur. Hence, there exists $P \in \mathbb{R}^{(N-1)d\times(N-1)d}$, $P = P^\top > 0$ for the candidate Lyapunov function $U(\tilde{\mathrm{z}}^k) = (\tilde{\mathrm{z}}^k)^\top P \tilde{\mathrm{z}}^k$, solving the Lyapunov equation

$$(R_d^\top \mathcal{W}_d R_d)^\top P R_d^\top \mathcal{W}_d R_d - P = -Q. \tag{4.18}$$

for any $Q \in \mathbb{R}^{(N-1)d\times(N-1)d}$, $Q = Q^\top > 0$. Condition (C.21a) follows then from the fact that U is quadratic with $P > 0$ so b_1 and b_2 can be chosen to be its minimum and maximum eigenvalue, respectively. The left-hand side of (C.21b) becomes $(\tilde{\mathrm{z}}^k)^\top((R_d^\top \mathcal{W}_d R_d)^\top P R_d^\top \mathcal{W}_d R_d - P)\tilde{\mathrm{z}}^k = -(\tilde{\mathrm{z}}^k)^\top Q \tilde{\mathrm{z}}^k$, where the equality is due to (4.18). Hence, (C.21b) is satisfied by taking b_3 to be the smallest eigenvalue of Q. To see (C.21c) notice that

$$\begin{aligned}
\left\|U(\tilde{\mathrm{z}}_1^k) - U(\tilde{\mathrm{z}}_2^k)\right\| &= \left\|(\tilde{\mathrm{z}}_1^k)^\top P \tilde{\mathrm{z}}_1^k - (\tilde{\mathrm{z}}_2^k)^\top P \tilde{\mathrm{z}}_2^k\right\| \\
&\leq \left\|(\tilde{\mathrm{z}}_1^k)^\top P \tilde{\mathrm{z}}_1^k - (\tilde{\mathrm{z}}_1^k)^\top P \tilde{\mathrm{z}}_2^k\right\| + \left\|(\tilde{\mathrm{z}}_2^k)^\top P \tilde{\mathrm{z}}_1^k - (\tilde{\mathrm{z}}_2^k)^\top P \tilde{\mathrm{z}}_2^k\right\| \\
&\leq \|P\| \left\|\tilde{\mathrm{z}}_1^k - \tilde{\mathrm{z}}_2^k\right\| \left\|\tilde{\mathrm{z}}_1^k\right\| + \|P\| \left\|\tilde{\mathrm{z}}_1^k - \tilde{\mathrm{z}}_2^k\right\| \left\|\tilde{\mathrm{z}}_2^k\right\|
\end{aligned} \tag{4.19}$$

where the first inequality follows from adding and subtracting $(\tilde{\mathrm{z}}_1^k)^\top P \tilde{\mathrm{z}}_2^k$ and using the triangle inequality, while the second one from the Cauchy-Schwarz inequality. The bound in (C.21c) follows then from (4.19) by taking b_4 to be the largest eigenvalue of P (recall it is symmetric). ■

4. Reduced system and satisfaction of (C.22)*:* The so-called reduced system can be obtained by plugging into (4.14a) the fast state at its steady state equilibrium, i.e., we

consider $\mathrm{z}^k = h(\mathrm{x}^k)$ for any $k \geq 0$. We thus have

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta \left(P_X \left[\mathrm{x}^k - \gamma \tilde{F}(\mathrm{x}^k, \phi(\mathrm{x}^k) + R_d h(\mathrm{x}^k)) \right] - \mathrm{x}^k \right). \tag{4.20}$$

Due to (4.12) we have that $\tilde{F}(\mathrm{x}^k, \phi(\mathrm{x}^k) + R_d h(\mathrm{x}^k)) = \tilde{F}(\mathrm{x}^k, \mathbf{1}_{N,d}\sigma(\mathrm{x}^k)) = F(\mathrm{x}^k)$, so (4.20) is equivalent to

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta \left(P_X \left[\mathrm{x}^k - \gamma F(\mathrm{x}^k) \right] - \mathrm{x}^k \right). \tag{4.21}$$

The next lemma provides a Lyapunov function for (4.20).

Lemma 4.2. *Consider system* (4.20). *Let $x^\star \in \mathbb{R}^n$ be such that $f(x^\star, h(x^\star)) = 0$ with f defined as in* (4.15). *Then, there exist a continuous function $W : \mathbb{R}^n \to \mathbb{R}$, $\bar{\gamma} > 0$ and $\bar{\delta}_2 > 0$ such that, for any $\gamma \in (0, \bar{\gamma})$ and any $\delta \in (0, \bar{\delta}_2)$, W satisfies* (C.22).

Proof. Pick the function $W : \mathbb{R}^n \to \mathbb{R}$ defined as

$$W(\mathrm{x}^k) = \frac{1}{2} \left\| \mathrm{x}^k - x^\star \right\|^2.$$

Since W is a quadratic function, conditions (C.22a) and (C.22c) (in Appendix C) are satisfied. To show (C.22b) we evaluate $\Delta W(\mathrm{x}^k) \coloneqq W(\mathrm{x}^{k+1}) - W(\mathrm{x}^k)$ along (4.21). We then have

$$\begin{aligned}
\Delta W(\mathrm{x}^k) &= \frac{1}{2} \left\| (1-\delta)\mathrm{x}^k + \delta \left(P_X \left[\mathrm{x}^k - \gamma F(\mathrm{x}^k) \right] \right) - x^\star \right\|^2 - \frac{1}{2} \left\| \mathrm{x}^k - x^\star \right\|^2 \\
&\overset{(a)}{\leq} \frac{(1-\delta)^2}{2} \left\| \mathrm{x}^k - x^\star \right\|^2 - \frac{1}{2} \left\| \mathrm{x}^k - x^\star \right\|^2 \\
&\quad + (\delta - \delta^2) \left\| \mathrm{x}^k - x^\star \right\| \left\| P_X \left[\mathrm{x}^k - \gamma F(\mathrm{x}^k) \right] - P_X \left[x^\star - \gamma F(x^\star) \right] \right\| \\
&\quad + \frac{\delta^2}{2} \left\| P_X \left[\mathrm{x}^k - \gamma F(\mathrm{x}^k) \right] - P_X \left[x^\star - \gamma F(x^\star) \right] \right\|^2 \\
&\overset{(b)}{\leq} \frac{(1-\delta)^2}{2} \left\| \mathrm{x}^k - x^\star \right\|^2 - \frac{1}{2} \left\| \mathrm{x}^k - x^\star \right\|^2 \\
&\quad + (\delta - \delta^2) \left\| \mathrm{x}^k - x^\star \right\| \left\| \mathrm{x}^k - \gamma F(\mathrm{x}^k) - x^\star + \gamma F(x^\star) \right\| \\
&\quad + \frac{\delta^2}{2} \left\| \mathrm{x}^k - \gamma F(\mathrm{x}^k) - x^\star + \gamma F(x^\star) \right\|^2,
\end{aligned} \tag{4.22}$$

where in (a) we introduce the quantity $\delta(x^\star - P_X\left[x^\star - \gamma F(x^\star)\right])$ within the first norm, as this is zero due to the definition of $x^\star$, expand the square, and use the Cauchy-Schwarz inequality. Inequality (b) follows by the fact that for any a, b, we have that $\|P_X\left[a\right] - P_X\left[b\right]\| \leq \|a - b\|$, since the projection operator is nonexpansive.

Since F is μ-strongly monotone and $\overline{L}_1$ Lipschitz continuous (cf. Assumption 4.2), set $\bar{\gamma} = 2\mu/(\overline{L}_1)^2$ and choose $\gamma \in (0, \bar{\gamma})$. Applying Lemma B.3 yields

$$\left\| \mathrm{x}^k - \gamma F(\mathrm{x}^k) - x^\star + \gamma F(x^\star) \right\| \leq (1 - \tilde{\mu}) \left\| \mathrm{x}^k - x^\star \right\|, \tag{4.23}$$

with $\tilde{\mu} = 1 - \sqrt{1 - \gamma(2\mu - \gamma(\overline{L}_1)^2)} \in (0, 1]$. Thus, by using the inequality in (4.23), we can bound (4.22) as follows

$$\begin{aligned}\Delta W(\mathrm{x}^k) &\leq \frac{(1-\delta)^2}{2}\left\|\mathrm{x}^k - x^\star\right\|^2 - \frac{1}{2}\left\|\mathrm{x}^k - x^\star\right\|^2 + (\delta - \delta^2)(1-\tilde{\mu})\left\|\mathrm{x}^k - x^\star\right\|^2 \\ &\quad + \frac{\delta^2(1-\tilde{\mu})^2}{2}\left\|\mathrm{x}^k - x^\star\right\|^2 \\ &= -\delta\tilde{\mu}\left(1 - \frac{\delta\tilde{\mu}}{2}\right)\left\|\mathrm{x}^k - x^\star\right\|^2. \end{aligned} \tag{4.24}$$

where the equality is obtained by rearranging the right-hand side of the inequality. Thus, for any $\delta \in (0, \bar{\delta}_2)$ with $\bar{\delta}_2 \coloneqq 2/\tilde{\mu}$, $\delta\tilde{\mu}(1 - \delta\tilde{\mu}/2) > 0$ in (4.24), thus establishing condition (C.22b) and concluding the proof. ■

5. Lipschitz continuity of f, g and h: As we will be invoking Theorem C.2 (in Appendix C), we need to ensure that the Lipschitz continuity assumptions required by the theorem are satisfied. In particular, we require f and g in (4.15) to be Lipschitz continuous with respect to both arguments, and h in (4.16) to be Lipschitz continuous with respect to x.

Lipschitz continuity of f follows by the fact that ∇J_i is Lipschitz continuous due to Assumption 4.2. To show Lipschitz continuity of g in (4.15) notice that for any $w, w' \in \mathbb{R}^{(N-1)d}$ and any $x, x' \in \mathbb{R}^n$,

$$\begin{aligned}\left\|R_d^\top \mathcal{W}_d R_d(w - w') + R_d^\top(\mathcal{W}_d - I)(\phi(x) - \phi(x'))\right\| &\leq \left\|R_d^\top \mathcal{W}_d R_d\right\| \|w - w'\| \\ &\quad + \overline{L}_3\left\|R_d^\top(\mathcal{W}_d - I)\right\| \|x - x'\|,\end{aligned}$$

where the inequality is due to triangle inequality and the fact that by Assumption 4.2, ϕ is Lipschitz continuous with Lipschitz constant $\overline{L}_3$. To show Lipschitz continuity of h, notice that for any $x, x' \in \mathbb{R}^n$,

$$\left\|h(x) - h(x')\right\| \leq \overline{L}_3\|R_d\|\left\|x - x'\right\| = \overline{L}_3\left\|x - x'\right\|,$$

where the inequality follows from (4.16) and Lipschitz continuity of ϕ, while the equality from the fact that $\|R_d\| = 1$.

By combining Lemma 4.1 and 4.2 with the Lipschitz conditions expressed above, Theorem C.2 can therefore be applied. Thus, there exists $\bar{\delta} \in (0, \bar{\delta}_2)$ so that $(x^\star, h(x^\star))$ is an exponentially stable equilibrium for (4.14). □

4.2.2 Primal-Dual TRADES: Algorithm Description and Analysis

In this section we introduce the Primal-Dual TRADES algorithm, i.e., a distributed iterative methodology to find a GNE in aggregative games with local and linear coupling

constraints as formalized in (4.1).

In addition to the assumptions made in Section 4.2, we need some further conditions for our mathematical developments.

Assumption 4.4 (Feasibility). *The set $\mathcal{C} \neq \emptyset$ and, for all $i \in \{1, \dots, N\}$, for any $x_{-i} \in \mathbb{R}^{n-n_i}$, (i) $\mathcal{C}_i(c_{-i}(x_{-i})) \neq \emptyset$ and (ii) $J_i(x_{i,br}(x_{-i}), \phi_i(x_{i,br}(x_{-i}))/N + \sum_{j\neq i} \phi_j(x_j)/N) > -\infty$.* △

Consider the following variational inequality, defined by the mapping F in (4.3) and the domain $\mathcal{C}$:

$$F(x^\star)^\top (x - x^\star) \geq 0, \text{ for all } x \in \mathcal{C}. \tag{4.25}$$

It is known that every point $x^\star \in \mathcal{C}$ for which (4.25) holds is a GNE of the game (4.1) and, specifically, a *variational* GNE (v-GNE) (cf. [59, Th. 2.1]). The converse, however, does not hold in general due to the presence of the coupling constraints. Since F is strongly monotone (cf. Assumption 4.2) and $\mathcal{C}$ nonempty, closed and convex (cf. Assumption 4.4), a v-GNE is guaranteed to exist and it is also unique by [167, Th. 3].

We will devise an iterative algorithm that will asymptotically return the (unique) v-GNE of (4.1). Inspired by [159], where an augmented primal-dual scheme was used for continuous-time, centralized optimization, we require the following additional condition on the matrix A of the coupling constraints (cf. (4.1)):

Assumption 4.5 (Full-row rank). *Matrix A satisfies $\operatorname{rank}(A) = m$, and there exist $\kappa_1, \kappa_2 > 0$ such that $\kappa_1 I_m \preccurlyeq AA^\top \preccurlyeq \kappa_2 I_m$.* △

Following [159], for all $i \in \{1, \dots, N\}$ we consider the augmented Lagrangian function $L_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined as

$$L_i(\mathrm{x}, \lambda) \coloneqq J_i(\mathrm{x}_i, \sigma(\mathrm{x})) + \underbrace{\sum_{\ell=1}^{m} H_\ell([A\mathrm{x} - b]_\ell, [\lambda]_\ell)}_{=: H(A\mathrm{x}-b, \lambda)}, \tag{4.26}$$

where

$$H_\ell([A\mathrm{x} - b]_\ell, [\lambda]_\ell) \coloneqq \begin{cases} [A\mathrm{x} - b][\lambda]_\ell + \frac{\rho}{2}([A\mathrm{x} - b]_\ell)^2 & \text{if } \rho([A\mathrm{x} - b]_\ell) + [\lambda]_\ell \geq 0 \\ -\frac{1}{2\rho}[\lambda]_\ell^2 & \text{if } \rho([A\mathrm{x} - b]_\ell) + [\lambda]_\ell < 0, \end{cases}$$

with $\lambda \in \mathbb{R}^m$ being the multiplier associated to the coupling constraints, and $\rho > 0$ a constant. We therefore address the v-GNE seeking problem by obtaining a saddle point of (4.26) through the discrete-time dynamics:

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k + \delta\Big(P_{X_i}\Big[\mathrm{x}_i^k - \gamma \nabla_{\mathrm{x}_i} J_i(\mathrm{x}_i^k, \sigma(\mathrm{x}^k)) - \gamma \nabla_{\mathrm{x}_i} H(A\mathrm{x}^k - b, \lambda^k)\Big] - \mathrm{x}_i^k\Big) \tag{4.27a}$$

$$\lambda^{k+1} = \lambda^k + \delta\gamma\nabla_\lambda H(A\mathrm{x}^k - b, \lambda^k), \tag{4.27b}$$

where x_i^k, δ, and γ have the same meaning as in (4.5), $\lambda^k \in \mathbb{R}^m$ is the multiplier at $k \geq 0$, and the explicit form of the gradients $\nabla_{\mathrm{x}_i} H(A\mathrm{x}^k - b, \lambda^k)$ and $\nabla_\lambda H(A\mathrm{x}^k - b, \lambda^k)$ reads as

$$\begin{aligned}
\nabla_{\mathrm{x}_i} H(A\mathrm{x}^k - b, \lambda^k) &= \sum_{\ell=1}^m \nabla_{\mathrm{x}_i} H_\ell([A\mathrm{x}^k - b]_\ell, [\lambda^k]_\ell) \\
&= \sum_{\ell=1}^m \max\left\{\rho([A\mathrm{x}^k - b]_\ell) + [\lambda^k]_\ell, 0\right\}[A_i]_\ell^\top, \qquad (4.28a) \\
\nabla_\lambda H(A\mathrm{x}^k - b, \lambda^k) &= \sum_{\ell=1}^m \nabla_\lambda H_\ell([A\mathrm{x}^k - b]_\ell, [\lambda^k]_\ell) \\
&= \sum_{\ell=1}^m \frac{1}{\rho} e_\ell(\max\left\{\rho([A\mathrm{x}^k - b]_\ell) + [\lambda^k]_\ell, 0\right\} - [\lambda^k]_\ell), \qquad (4.28b)
\end{aligned}$$

where $e_\ell \in \mathbb{R}^m$ is the ℓ-th vector of the canonical basis of $\mathbb{R}^m$, $\ell \in \{1, \dots, m\}$. The stacked-column form of (4.27) is

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta\Big(P_X\left[\mathrm{x}^k - \gamma F(\mathrm{x}) - \gamma\nabla_{\mathrm{x}} H(A\mathrm{x}^k - b, \lambda^k)\right] - \mathrm{x}^k\Big), \tag{4.29a}$$

$$\lambda^{k+1} = \lambda^k + \delta\gamma\nabla_\lambda H(A\mathrm{x}^k - b, \lambda^k), \tag{4.29b}$$

where $\nabla_x H(A\mathrm{x}^k - b, \lambda^k) \coloneqq \mathrm{col}(\nabla_{x_1} H(A\mathrm{x}^k - b, \lambda^k), \dots, \nabla_{x_N} H(A\mathrm{x}^k - b, \lambda^k)$.

However, since agent i does not have access neither to $\sigma(\mathrm{x}^k)$ nor to $A\mathrm{x}^k - b$, the scheme in (4.27) cannot be directly implemented. Moreover, dynamics (4.27) requires a central unit that can compute the global quantity $A\mathrm{x}^k - b$ and communicate the multiplier λ^k to all the agents. For this reason, in Algorithm 9 we introduce for all $i \in \{1, \dots, N\}$ (i) two additional variables $\mathrm{z}_i \in \mathbb{R}^d$ and $\mathrm{y}_i \in \mathbb{R}^m$ to compensate the local unavailability of $\sigma(\mathrm{x}^k)$ and $A\mathrm{x}^k - b$, respectively, (ii) a copy $\lambda_i \in \mathbb{R}^m$ of the multiplier λ, and (iii) an additional average consensus step to enforce agreement among the multipliers λ_i, (cf. (4.31b)-(4.31d)). We choose causal perturbed consensus dynamics to update z_i and y_i. For all $i \in \{1, \dots, N\}$, we then introduce operators $G_{\mathrm{x},i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^{n_i}$ and $G_{\lambda,i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ as

$$\begin{aligned}
G_{\mathrm{x},i}(s_1, s_2) &\coloneqq \sum_{\ell=1}^m \max\{\rho([s_1]_\ell) + [s_2]_\ell, 0\}[A_i]_\ell^\top, \\
G_{\lambda,i}(s_1, s_2) &\coloneqq \frac{1}{\rho}\sum_{\ell=1}^m \left(\max\{\rho([s_1]_\ell) + [s_2]_\ell, 0\} - [s_2]_\ell\right) e_\ell.
\end{aligned} \tag{4.30}$$

In Algorithm 9, these operators encode the component of the gradients in (4.28) available to agent i at iteration k, plus the auxiliary variable y_i^k that is used to track $A\mathrm{x}^k - b$ (see

Algorithm 9 Primal-Dual TRADES (Agent i)

Initialization: $\mathrm{x}_i^0 \in X_i, \lambda_i^k \in \mathbb{R}_+^m, \mathrm{z}_i^0 = 0, \mathrm{y}_i^0 = 0$.
for $k = 0, 1, \dots$ **do**

$$\mathrm{x}_i^{k+1} = \mathrm{x}_i^k + \delta\left(P_{X_i}\left[\mathrm{x}_i^k - \gamma\tilde{F}_i(\mathrm{x}_i^k, \phi_i(\mathrm{x}_i^k) + \mathrm{z}_i^k) - \gamma G_{\mathrm{x},i}(N(A_i\mathrm{x}_i^k - b_i) + \mathrm{y}_i^k, \lambda_i^k)\right] - \mathrm{x}_i^k\right) \tag{4.31a}$$

$$\lambda_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\lambda_j^k + \delta\gamma G_{\lambda,i}(N(A_i\mathrm{x}_i^k - b_i) + \mathrm{y}_i^k, \lambda_i^k) \tag{4.31b}$$

$$\mathrm{z}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{z}_j^k + \sum_{j\in\mathcal{N}_i} w_{ij}\phi_j(\mathrm{x}_j^k) - \phi_i(\mathrm{x}_i^k) \tag{4.31c}$$

$$\mathrm{y}_i^{k+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\mathrm{y}_j^k + \sum_{j\in\mathcal{N}_i} w_{ij}N(A_j\mathrm{x}_j^k - b_j) - N(A_i\mathrm{x}_i^k - b_i), \tag{4.31d}$$

end for

(4.31a) and (4.31b) in Algorithm 9). The main steps of the proposed method are hence summarized in Algorithm 9 from the perspective of agent i, which is then referred as Primal-Dual TRADES. Note that all the quantities involved in the agent's calculations are purely local, thus making Algorithm 9 fully distributed.

Differently from customary primal-dual schemes, (4.31b) does not need the projection over the positive orthant $\mathbb{R}_+^m$ due to the chosen augmented Lagrangian functions L_i (see (4.26)). We only need to initialize $\lambda_i^0 \geq 0$ for all $i \in \{1, \dots, N\}$, and choose parameters δ, γ, and ρ appropriately so that we avoid situations where $\lambda_i^k \geq 0$ implies $\lambda_i^{k+1} < 0$. To see this notice first that if $\lambda_i^k = 0$, then it is easy to check $G_{\lambda,i}(N(A_i\mathrm{x}_i^k - b_i) + \mathrm{y}_i^k, \lambda_i^k) \geq 0$ and, thus, $\lambda_i^{k+1} \geq 0$. The critical scenario for agent i occurs when all the multipliers of its neighbors are zero, namely $\lambda_j^k = 0$ for any $j \in \mathcal{N}_i$, and when $\max\{\rho([N(A_i\mathrm{x}_i^k - b_i + \mathrm{y}_i^k]_\ell) + [\lambda_i^k]_\ell, 0\} = 0$ for at least one $\ell \in \{1, \dots, m\}$. Indeed, specializing (4.31b) for this case leads to the following update of that ℓ-th component of λ_i^k

$$[\lambda_i^{k+1}]_\ell = \left(w_{ii} - \frac{\delta\gamma}{\rho}\right)[\lambda_i^k]_\ell. \tag{4.32}$$

From (4.32), we conclude that $[\lambda_i^{k+1}]_\ell$ remains non-negative if $[\lambda_i^k]_\ell$ is non-negative, thus alleviating the need for a projection, as long as δ, γ, and ρ satisfy $w_{ii} > \delta\gamma/\rho$.

As in the case without coupling constraints, the purpose of the initialization step will become clear in the next subsection. The steps of Algorithm 9 in (4.31) can be compactly written as:

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta f_X(\mathrm{x}^k, \lambda^k, \mathrm{z}^k, \mathrm{y}^k), \tag{4.33a}$$

$$\lambda^{k+1} = \mathcal{W}_m \lambda^k + \delta\gamma G_\lambda(N(\bar{A}\mathrm{x}^k - \bar{b}) + \mathrm{y}^k, \lambda^k), \tag{4.33b}$$
$$\mathrm{z}^{k+1} = \mathcal{W}_d \mathrm{z}^k + (\mathcal{W}_d - I)\phi(\mathrm{x}^k), \tag{4.33c}$$
$$\mathrm{y}^{k+1} = \mathcal{W}_m \mathrm{y}^k + (\mathcal{W}_m - I)N(\bar{A}\mathrm{x}^k - \bar{b}). \tag{4.33d}$$

where $f_X : \mathbb{R}^n \times \mathbb{R}^{Nm} \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nm} \to \mathbb{R}^n$ is defined as

$$f_X(\mathrm{x}, \lambda, \mathrm{z}, \mathrm{y}) \coloneqq P_X\left[x - \gamma\tilde{F}(\mathrm{x}, \phi(\mathrm{x}) + \mathrm{z}) - \gamma G_\mathrm{x}(N(\bar{A}\mathrm{x} - \bar{b}) + \mathrm{y}, \lambda)\right] - \mathrm{x},$$

and, similarly to (4.10), $\lambda : \mathrm{col}(\lambda_1, \dots, \lambda_N)$, $\mathcal{W}_d \coloneqq \mathcal{W}_\mathcal{G} \otimes I_d$, $\mathcal{W}_m \coloneqq \mathcal{W}_\mathcal{G} \otimes I_m$, $G_\mathrm{x}(N(\bar{A}\mathrm{x}^k - \bar{b}) + \mathrm{y}^k, \lambda^k) \coloneqq \mathrm{col}(G_{\mathrm{x},1}(N(A_1\mathrm{x}_1^k - b_1) + \mathrm{y}_1^k, \lambda_1^k), \dots, G_{\mathrm{x},N}(N(A_N\mathrm{x}_N^k - b_N) + \mathrm{y}_N^k, \lambda_N^k))$, and $G_\lambda(N(\bar{A}\mathrm{x}^k - \bar{b}) + \mathrm{y}^k, \lambda^k) \coloneqq \mathrm{col}(G_{\lambda,1}(N(A_1\mathrm{x}_1^k - b_1) + \mathrm{y}_1^k, \lambda_1^k), \dots, G_{\lambda,N}(N(A_N\mathrm{x}_N^k - b_N) + \mathrm{y}_N^k, \lambda_N^k))$.

The next theorem establishes the convergence properties of Primal-Dual TRADES in computing the v-GNE of (4.1).

Theorem 4.2. *Consider 4.33 and Assumptions 4.4, 4.5. Let* $(\mathrm{x}^0, \lambda^0, \mathrm{z}^0, \mathrm{y}^0) \in X \times \mathbb{R}_+^{Nm} \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nm}$ *satisfy* $\mathbf{1}_{N,d}^\top \mathrm{z}^0 = 0$ *and* $\mathbf{1}_{N,m}^\top \mathrm{y}^0 = 0$*. Then, there exist constants* $\bar{\delta}, \bar{\gamma}, a_1, a_2 > 0$ *such that, for any* $\delta \in (0, \bar{\delta})$*,* $\gamma \in (0, \bar{\gamma})$*, with* $w_{ii} > \frac{\delta\gamma}{\rho}$ *for all* $i \in \{1, \dots, N\}$*, it holds*

$$\left\|\mathrm{x}^k - x^\star\right\| \leq a_1 \exp(-a_2 k). \qquad \triangle$$

Note that the additional condition $w_{ii} > \delta\gamma/\rho$ needs to be satisfied by δ and γ, given ρ, to ensure the dual variables remain non-negative, as discussed below (4.32). As in the case of NE seeking without coupling constraints, the proof of Theorem 4.2 relies on a *singular perturbations* analysis of system (4.33). We provide this in the next subsection.

Proof of Theorem 4.2

As with the proof of Theorem 4.1, we show that the setting of Theorem 4.2 fits the framework of Theorem C.2 (in Appendix C), and organize its proof in five steps.

1. Bringing (4.33) *in the form of* (C.18)*:* We introduce the change of coordinates

$$\begin{bmatrix}\bar{\mathrm{z}}^k \\ \mathrm{z}_\perp^k\end{bmatrix} = \begin{bmatrix}\frac{\mathbf{1}_{N,d}^\top}{N} \\ R_d^\top\end{bmatrix}\mathrm{z}^k, \quad \begin{bmatrix}\bar{\mathrm{y}}^k \\ \mathrm{y}_\perp^k\end{bmatrix} = \begin{bmatrix}\frac{\mathbf{1}_{N,m}^\top}{N} \\ R_m^\top\end{bmatrix}\mathrm{y}^k, \begin{bmatrix}\bar{\lambda}^k \\ \lambda_\perp^k\end{bmatrix} = \begin{bmatrix}\frac{\mathbf{1}_{N,m}^\top}{N} \\ R_m^\top\end{bmatrix}\lambda^k, \tag{4.34}$$

where $R_d \in \mathbb{R}^{Nd\times(N-1)d}$, $R_m \in \mathbb{R}^{Nm\times(N-1)m}$, $\|R_d\| = 1$, $\|R_m\| = 1$, and

$$R_d R_d^\top = I - \frac{\mathbf{1}_{N,d}\mathbf{1}_{N,d}^\top}{N},\ R_m R_m^\top = I - \frac{\mathbf{1}_{N,m}\mathbf{1}_{N,m}^\top}{N}. \tag{4.35}$$

As in the proof of Theorem 4.2, we use the initialization $\mathbf{1}_{N,d}^\top \mathrm{z}^0 = 0$ and $\mathbf{1}_{N,m}^\top \mathrm{y}^0 = 0$ to ensure that $\bar{\mathrm{z}}^k = 0$ and $\bar{\mathrm{y}}^k = 0$ for all $k \geq 0$. In view of (4.34), we can therefore

rewrite (4.33) by ignoring the dynamics of $\bar{\mathrm{z}}^k$ and $\bar{\mathrm{y}}^k$, thus obtaining the system

$$\chi^{k+1} = \chi^k + \delta f(\chi^k, \mathrm{w}^k), \tag{4.36a}$$
$$\mathrm{w}^{k+1} = S\mathrm{w}^k + K(\delta,\gamma)u(\chi^k). \tag{4.36b}$$

in which

$$\chi^k := \begin{bmatrix} \mathrm{x}^k \\ \bar{\lambda}^k \end{bmatrix}, \quad \mathrm{w}^k := \begin{bmatrix} \lambda_\perp^k \\ \mathrm{z}_\perp^k \\ \mathrm{y}_\perp^k \end{bmatrix}, \tag{4.37a}$$

$$f(\chi^k, \mathrm{w}^k) := \begin{bmatrix} f_X(\mathrm{x}^k, \mathbf{1}_{N,m}\bar{\lambda}^k + R_m\lambda_\perp^k, R_d\mathrm{z}_\perp^k, R_m\mathrm{y}_\perp^k) \\ \gamma\frac{\mathbf{1}_{N,m}^\top}{N} G_\lambda(N(\bar{A}\mathrm{x}^k - \bar{b}) + R_m\mathrm{y}_\perp^k, \mathbf{1}_{N,m}\bar{\lambda}^k + R_m\lambda_\perp^k) \end{bmatrix}, \tag{4.37b}$$

$$S := \begin{bmatrix} R_m^\top \mathcal{W}_m R_m & 0 & 0 \\ 0 & R_d^\top \mathcal{W}_d R_d & 0 \\ 0 & 0 & R_m^\top \mathcal{W}_m R_m \end{bmatrix}, \tag{4.37c}$$

$$K(\delta,\gamma) := \begin{bmatrix} \delta\gamma R_m^\top & 0 & 0 \\ 0 & R_d^\top(\mathcal{W}_d - I) & 0 \\ 0 & 0 & R_m^\top(\mathcal{W}_m - I) \end{bmatrix}, \tag{4.37d}$$

$$u(\chi^k) := \begin{bmatrix} G_\lambda(N(\bar{A}\mathrm{x}^k - \bar{b}) + R_m\mathrm{y}_\perp^k, \mathbf{1}_{N,m}\bar{\lambda}^k + R_m\lambda_\perp^k) \\ \phi(\mathrm{x}^k) \\ N(\bar{A}\mathrm{x}^k - \bar{b}) \end{bmatrix}. \tag{4.37e}$$

where We view (4.36) as a singularly perturbed system, namely the interconnection between the slow dynamics (4.36a) and the fast one (4.36b). Indeed, system (4.36) can be obtained from (C.18) by considering χ^k as the state of (C.18a) and setting

$$g(\chi^k, \mathrm{w}^k, \delta) := S\mathrm{w}^k + K(\delta,\gamma)u(\chi^k). \tag{4.38}$$

2. Equilibrium function h: Under the double stochasticity condition of $\mathcal{W}_\mathcal{G}$ due to Assumption 4.3 and using (4.35), for any $\chi^k = \chi$,

$$h(\chi) := \begin{bmatrix} 0 \\ -R_d^\top \phi\left(\begin{bmatrix} I_n & 0 \end{bmatrix}\chi\right) \\ -R_m^\top N\left(\bar{A}\begin{bmatrix} I_n & 0 \end{bmatrix}\chi - \bar{b}\right) \end{bmatrix} \tag{4.39}$$

constitutes an equilibrium of (4.36b) (parametrized by x).

3. Boundary layer system and satisfaction of (C.21)*:* The so-called boundary layer system associated to (4.36) can be constructed by fixing $\chi^k = \chi = \mathrm{col}(\mathrm{x}, \bar{\lambda})$ for some arbitrary $(\mathrm{x}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$, and rewriting it according to the error coordinates $\tilde{\mathrm{w}} :=$

$\text{col}(\tilde{\lambda}_\perp, \tilde{z}_\perp, \tilde{y}_\perp) \coloneqq w - h(\chi)$. Using (4.35), we then obtain that

$$\tilde{\mathrm{w}}^{k+1} = S\tilde{\mathrm{w}}^k + \delta\gamma\tilde{u}(\chi, \tilde{\mathrm{w}}^k), \tag{4.40}$$

where

$$\tilde{u}(\chi, \tilde{\mathrm{w}}^k) \coloneqq \begin{bmatrix} R_m^\top G_\lambda \left(\mathbf{1}_{N,m}(A\mathrm{x} - b) + R_m\tilde{\mathrm{y}}_\perp^k, \mathbf{1}_{N,m}\bar{\lambda} + R_m\tilde{\lambda}_\perp^k\right) \\ 0 \\ 0 \end{bmatrix}.$$

The next lemma provides a Lyapunov function for (4.40).

Lemma 4.3. *Consider system* (4.40)*. Then, there exists a continuous function* $U : \mathbb{R}^{(N-1)(2m+d)} \to \mathbb{R}$ *and* $\bar{\delta}_1 > 0$ *such that for any* $\delta \in (0, \bar{\delta}_1)$ *and any* $\gamma > 0$*,* U *satisfies* (C.21) *with* $\tilde{\mathrm{w}}$ *in place of* ψ*.*

Proof. Since $R_m^\top \mathbf{1}_{N,m} = 0$, we can write

$$\begin{aligned} &R_m^\top G_\lambda \left(\mathbf{1}_{N,m}(A\mathrm{x} - b) + R_m\tilde{\mathrm{y}}_\perp^k, \mathbf{1}_{N,m}\bar{\lambda} + R_m\tilde{\lambda}_\perp^k\right) \\ &= R_m^\top \Big(G_\lambda \left(\mathbf{1}_{N,m}(A\mathrm{x} - b) + R_m\tilde{\mathrm{y}}_\perp^k, \mathbf{1}_{N,m}\bar{\lambda} + R_m\tilde{\lambda}_\perp^k\right) - \mathbf{1}_{N,m}\nabla_\lambda H(A\mathrm{x} - b, \bar{\lambda}) \Big) \\ &= R_m^\top \Big(G_\lambda \left(\mathbf{1}_{N,m}(A\mathrm{x} - b) + R_m\tilde{\mathrm{y}}_\perp^k, \mathbf{1}_{N,m}\bar{\lambda} + R_m\tilde{\lambda}_\perp^k\right) - G_\lambda(\mathbf{1}_{N,m}(A\mathrm{x} - b), \mathbf{1}_{N,m}\bar{\lambda}) \Big), \end{aligned} \tag{4.41}$$

where in the last equality we used $\mathbf{1}_{N,m}\nabla_\lambda H(A\mathrm{x} - b, \bar{\lambda}) = G_\lambda(\mathbf{1}_{N,m}(A\mathrm{x} - b), \mathbf{1}_{N,m}\bar{\lambda})$. Following [159, Lemma 3], notice that, for any $r_1, r_2 \in \mathbb{R}$, there exists $\epsilon(r_1, r_2) \in [0, 1]$ so that[1]

$$\max\{r_1, 0\} - \max\{r_2, 0\} = \epsilon(r_1, r_2)(r_1 - r_2). \tag{4.42}$$

Let us introduce

$$\begin{aligned} q_i^k &\coloneqq \sum_{\ell=1}^m [R_m\tilde{\mathrm{y}}_\perp^k]_{\ell+(i-1)m} e_\ell \\ p_i^k &\coloneqq \sum_{\ell=1}^m [R_m\tilde{\lambda}_\perp^k]_{\ell+(i-1)m} e_\ell, \end{aligned} \tag{4.43}$$

and use them to define

$$\begin{aligned} r_{1,i}^k &\coloneqq \rho(A\mathrm{x} - b + q_i^k) + \bar{\lambda} + p_i^k \\ r_{2,i} &\coloneqq \rho(A\mathrm{x} - b) + \bar{\lambda}. \end{aligned} \tag{4.44}$$

[1] If $r_1 \neq r_2$, pick $\epsilon = \frac{\max\{r_1,0\} - \max\{r_2,0\}}{r_1 - r_2}$, otherwise set $\epsilon = 0$.

By the definition of $\tilde{u}(\chi, \tilde{\mathrm{w}}^k)$ we have that its norm $\left\|\tilde{u}(\chi, \tilde{\mathrm{w}}^k)\right\|$ is equal to the norm of the quantity in (4.41). As such, for any $\chi \in \mathbb{R}^{n+m}$ and $\tilde{\mathrm{w}}^k \in \mathbb{R}^{(N-1)(2m+d)}$, we use the definition of G_λ in (4.30), $r^k_{1,i}$ and $r_{2,i}$ in (4.44), and apply (4.42) for each component of $\tilde{u}(\chi, \tilde{\mathrm{w}}^k)$ obtaining

$$\begin{aligned}
\left\|\tilde{u}(\chi, \tilde{\mathrm{w}}^k)\right\| &\leq \left\| R_m^\top \frac{1}{\rho} \textsc{col}\left(\sum_{\ell=1}^{m} \epsilon([r^k_{1,i}]_\ell, [r_{2,i}]_\ell)\big([r^k_{1,i} - \bar{\lambda} - p^k_i]_\ell - [r_{2,i} - \bar{\lambda}]_\ell\big) e_\ell \right)_{i=1}^{N} \right\| \\
&\overset{(a)}{\leq} \left\| R_m^\top \frac{1}{\rho} \textsc{col}\left(\sum_{\ell=1}^{m} \big([r^k_{1,i} - \bar{\lambda} - p^k_i]_\ell - [r_{2,i} - \bar{\lambda}]_\ell\big) e_\ell \right)_{i=1}^{N} \right\| \\
&\overset{(b)}{=} \left\| R_m^\top \frac{1}{\rho} \textsc{col}\left(\sum_{\ell=1}^{m} \rho [q^k_i]_\ell e_\ell \right)_{i=1}^{N} \right\| \\
&\overset{(c)}{=} \left\| R_m^\top R_m \tilde{\mathrm{y}}^k_\perp \right\| \overset{(d)}{\leq} \left\| \tilde{\mathrm{w}}^k \right\|,
\end{aligned} \tag{4.45}$$

where in (a) we use the fact that $\epsilon([r^k_{1,i}]_\ell, [r_{2,i}]_\ell) \in [0,1]$ for all $\ell \in \{1, \dots, m\}$ and $i \in \{1, \dots, N\}$, (b) uses the definitions in (4.44) to simplify the terms, (c) follows from (4.43), and (d) uses $R_m^\top R_m = I$ and $\left\|\tilde{\mathrm{y}}^k_\perp\right\| \leq \left\|\tilde{\mathrm{w}}^k\right\|$ that holds since $\tilde{\mathrm{y}}^k_\perp$ is a component of $\tilde{\mathrm{w}}^k$.

Pick now $U : \mathbb{R}^{(N-1)(2m+d)} \to \mathbb{R}$ defined as

$$U(\tilde{\mathrm{w}}) = (\tilde{\mathrm{w}})^\top M \tilde{\mathrm{w}},$$

where $M \in \mathbb{R}^{(N-1)(2m+d)\times(N-1)(2m+d)}$ with $M = M^\top > 0$, such that

$$S^\top M S - M = -I. \tag{4.46}$$

We remark that such a matrix M always exists because, in light of Assumption 4.3, both $R_d^\top \mathcal{W}_d R_d$ and $R_m^\top \mathcal{W}_m R_m$ are Schur matrices and, thus, S is Schur as well. Under this choice of U, conditions (C.21a) and (C.21c) are satisfied. To show (C.21b) we evaluate $\Delta U(\tilde{\mathrm{w}}^k) \coloneqq U(\tilde{\mathrm{w}}^{k+1}) - U(\tilde{\mathrm{w}}^k)$ along the trajectories of (4.40), obtaining

$$\begin{aligned}
\Delta U(\tilde{\mathrm{w}}^k) &= (S\tilde{\mathrm{w}}^k + \delta\gamma\tilde{u}(\chi, \tilde{\mathrm{w}}^k))^\top M (S\tilde{\mathrm{w}}^k + \delta\gamma\tilde{u}(\chi, \tilde{\mathrm{w}}^k)) - (\tilde{\mathrm{w}}^k)^\top M \tilde{\mathrm{w}}^k \\
&= -\left\|\tilde{\mathrm{w}}^k\right\|^2 + 2\delta\gamma(\tilde{\mathrm{w}}^k)^\top S^\top M \tilde{u}(\chi, \tilde{\mathrm{w}}^k) + \delta^2\gamma^2 \tilde{u}(\chi, \tilde{\mathrm{w}}^k)^\top M \tilde{u}(\chi, \tilde{\mathrm{w}}^k) \\
&\leq -(1 - \delta\gamma\mu_1 - \delta^2\gamma^2\mu_2)\left\|\tilde{\mathrm{w}}^k\right\|^2,
\end{aligned} \tag{4.47}$$

where the second equality is due to (4.46), and the inequality is due to (4.45) and the Cauchy-Schwarz inequality , with the constants $\mu_1 \coloneqq 2\|S\|\,\|M\|$ and $\mu_2 \coloneqq \|M\|$. Thus, there always exists $\bar{\delta}_1 > 0$ small enough so that $(1 - \delta\gamma\mu_1 - \delta^2\gamma^2\mu_2) > 0$ for any $\delta \in (0, \bar{\delta}_1)$ and $\gamma > 0$, concluding the proof. ■

4. Reduced system and satisfaction of (C.22)*:* The so-called reduced system can be obtained by considering the fast dynamics in (4.36a) at steady state, i.e., $\mathrm{w}^k = h(\chi^k)$ for any $k \geq 0$. We thus have

$$\chi^{k+1} = \chi^k + \delta f(\chi^k, h(\chi^k)). \tag{4.48}$$

Let us expand (4.48). Using (4.35), we obtain

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta\Big(P_X\left[\mathrm{x}^k - \gamma \tilde{F}\left(\mathrm{x}^k, \mathbf{1}_{N,d}\sigma(\mathrm{x}^k)\right) - \gamma G_{\mathrm{x}}\left(\mathbf{1}_{N,m}(A\mathrm{x}^k - b), \mathbf{1}_{N,m}\bar{\lambda}^k\right)\right] - \mathrm{x}^k\Big), \tag{4.49a}$$

$$\bar{\lambda}^{k+1} = \bar{\lambda}^k + \delta\gamma \frac{\mathbf{1}_{N,m}^\top}{N} G_\lambda\left(\mathbf{1}_{N,m}(A\mathrm{x}^k - b), \mathbf{1}_{N,m}\bar{\lambda}^k\right). \tag{4.49b}$$

Notice that

$$\tilde{F}\left(\mathrm{x}, \mathbf{1}_{N,d}\sigma(\mathrm{x})\right) = F(\mathrm{x}),$$
$$G_{\mathrm{x}}\left(\mathbf{1}_{N,m}(A\mathrm{x}^k - b), \mathbf{1}_{N,m}\bar{\lambda}^k\right) = \nabla_{\mathrm{x}} H(A\mathrm{x}^k - b, \bar{\lambda}^k),$$

and also

$$\frac{\mathbf{1}_{N,m}^\top}{N} G_\lambda\left(\mathbf{1}_{N,m}(A\mathrm{x}^k - b), \mathbf{1}_{N,m}\bar{\lambda}^k\right) = \nabla_\lambda H(A\mathrm{x}^k - b, \bar{\lambda}^k).$$

Therefore, (4.48) is identical to the original update (4.29). Given the unique v-GNE $x^\star$ of (4.1) (see Assumptions 4.4, 4.5) and the associated multiplier $\lambda^\star \in \mathbb{R}^m$, the next lemma provides a Lyapunov function for (4.48), hence for (4.29).

Lemma 4.4. *Consider system* (4.48) *and Assumptions 4.4, 4.5. Then, there exist a continuous function* $W : \mathbb{R}^{n+m} \to \mathbb{R}$*,* $\bar{\delta} > 0$*, and* $\bar{\gamma} > 0$ *such that for any* $\delta \in (0, \bar{\delta})$ *and* $\gamma \in (0, \bar{\gamma})$*,* W *satisfies* (C.22) *with* χ *in place of* x*.*

Proof. The proof is inspired by [159, Theorem 2, Lemma 3, Lemma 4], adapted to our framework. Let $\mathcal{F} : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ and $\mathcal{H} : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ be defined as

$$\mathcal{F}(\chi^k) \coloneqq \begin{bmatrix} F\left(\begin{bmatrix} I & 0 \end{bmatrix} \chi^k\right) \\ 0 \end{bmatrix}, \tag{4.50a}$$

$$\mathcal{H}(\chi^k) \coloneqq \begin{bmatrix} \nabla_{\mathrm{x}} H\left(A\begin{bmatrix} I & 0 \end{bmatrix}\chi^k - b, \begin{bmatrix} 0 & I \end{bmatrix}\chi^k\right) \\ -\nabla_\lambda H\left(A\begin{bmatrix} I & 0 \end{bmatrix}\chi^k - b, \begin{bmatrix} 0 & I \end{bmatrix}\chi^k\right) \end{bmatrix}. \tag{4.50b}$$

Applying (4.42) to each of the components of $\mathcal{H}(\chi^k) - \mathcal{H}(\chi^\star)$, for any $\chi^k \in \mathbb{R}^{n+m}$ we obtain

$$\mathcal{H}(\chi^k) - \mathcal{H}(\chi^\star) = \begin{bmatrix} \rho A^\top E(\chi^k, \chi^\star) A & A^\top E(\chi^k, \chi^\star) \\ -E(\chi^k, \chi^\star) A & -\frac{1}{\rho}(E(\chi^k, \chi^\star) - I) \end{bmatrix} (\chi^k - \chi^\star), \tag{4.51}$$

where $E(\chi^k, \chi^\star) \coloneqq \mathrm{diag}(\epsilon_1(\chi^k, \chi^\star), \dots, \epsilon_m(\chi^k, \chi^\star))$ and $\epsilon_\ell(\chi^k, \chi^\star) \in [0, 1]$ so that

$$\begin{aligned}&\max\{\rho([A\mathrm{x}^k - b]_\ell) + [\bar{\lambda}^k]_\ell, 0\} - \max\{\rho([Ax^\star - b]_\ell) + [\lambda^\star]_\ell, 0\}\\ &= \epsilon_\ell(\chi^k, \chi^\star)(\rho([A\mathrm{x}^k - b]_\ell - [Ax^\star - b]_\ell) + [\bar{\lambda}^k]_\ell - [\lambda^\star]_\ell),\end{aligned}$$

for all $\ell \in \{1, \dots, m\}$ and $\chi^k \coloneqq \mathrm{col}(\mathrm{x}^k, \bar{\lambda}^k) \in \mathbb{R}^{n+m}$. Moreover, for any $\mathrm{x}^k \in \mathbb{R}^n$, we have

$$\begin{aligned}F(\mathrm{x}^k) - F(x^\star) &= \int_0^1 \nabla F((1-\nu)x^\star + \nu\mathrm{x}^k)(\mathrm{x}^k - x^\star)d\nu\\ &\overset{(a)}{=} \left[\int_0^1 \nabla F((1-\nu)x^\star + \nu\mathrm{x}^k)d\nu\right](\mathrm{x}^k - x^\star)\\ &\overset{(b)}{=} B(\mathrm{x}^k, x^\star)(\mathrm{x}^k - x^\star).\end{aligned} \tag{4.52}$$

where in (a) we have extracted the term $(\mathrm{x}^k - x^\star)$ from the integral and in (b) we have introduced $B(\mathrm{x}^k, x^\star) \coloneqq \int_0^1 \nabla F((1-\nu)x^\star + \nu\mathrm{x}^k)d\nu$. Since F is μ-strongly monotone and $\overline{L}_1$-Lipschitz continuous (cf. Assumption 4.2), we can uniformly bound the integrand term of (4.52) as

$$\mu I \preccurlyeq \nabla F((1-\nu)x^\star + \nu\mathrm{x}^k) \preccurlyeq \overline{L}_1 I,$$

which leads to

$$\mu I \preccurlyeq \int_0^1 \mu I d\nu \preccurlyeq B(\mathrm{x}^k, x^\star) \preccurlyeq \int_0^1 \overline{L}_1 I d\nu \preccurlyeq \overline{L}_1 I. \tag{4.53}$$

Combining the definitions (4.50) with (4.51) and (4.52), we can write

$$-\mathcal{F}(\chi^k) + \mathcal{F}(\chi^\star) - (\mathcal{H}(\chi^k) - \mathcal{H}(\chi^\star)) = D(\chi^k, \chi^\star)(\chi^k - \chi^\star), \tag{4.54}$$

where $D(\chi^k, \chi^\star) \in \mathbb{R}^{(n+m)\times(n+m)}$ is given by

$$D(\chi^k, \chi^\star) \coloneqq \begin{bmatrix} -B(\chi^k, \chi^\star) - \rho A^\top E(\chi^k, \chi^\star)A & -A^\top E(\chi^k, \chi^\star) \\ E(\chi^k, \chi^\star)A & \frac{1}{\rho}(E(\chi^k, \chi^\star) - I)\end{bmatrix}.$$

Consider now $M \in \mathbb{R}^{(n+m)\times(n+m)}$ defined as

$$M \coloneqq \begin{bmatrix} cI & A^\top \\ A & cI\end{bmatrix}, \tag{4.55}$$

and notice that choosing c such that $c^2 > \kappa_2$ (cf. Assumption 4.5) ensures that $M > 0$ (see also [159, Theorem 1]). Now, let

$$\mathcal{P}_{\mathcal{X}}[\chi] \coloneqq P_{X\times\mathbb{R}^m}[\chi]. \tag{4.56}$$

We can employ matrix M to show that $\left\|\mathcal{P}_{\mathcal{X}}\left[\chi^k-\gamma\mathcal{F}(\chi^k)-\gamma\mathcal{H}(\chi^k)\right]-\chi^\star\right\|_M$ enjoys certain contraction properties under the M-weighted norm. Note that $\mathcal{P}_{\mathcal{X}}\left[\chi^k-\gamma\mathcal{F}(\chi^k)-\gamma\mathcal{H}(\chi^k)\right]$ combines both the projected descent step for x^{k+1} and the ascent step for λ^{k+1} in (4.33); this also justifies the opposite sign in the two block rows of $\mathcal{H}(\chi^k)$ (cf. (4.50b)) and hence also of $D(\chi^k,\chi^\star)$.

We then have that

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{X}}\left[\chi^k-\gamma\mathcal{F}(\chi^k)-\gamma\mathcal{H}(\chi^k)\right]-\chi^\star\right\|_M^2\\
&\overset{(a)}{\leq}\left\|\chi^k-\chi^\star-\gamma(\mathcal{F}(\chi^k)-\mathcal{F}(\chi^\star))-\gamma(\mathcal{H}(\chi^k)-\mathcal{H}(\chi^\star))\right\|_M^2\\
&\overset{(b)}{\leq}\left\|\chi^k-\chi^\star+\gamma D(\chi^k,\chi^\star)(\chi^k-\chi^\star)\right\|_M^2\\
&\overset{(c)}{=}\left\|\chi^k-\chi^\star\right\|_M^2+\gamma^2\left\|D(\chi^k,\chi^\star)(\chi^k-\chi^\star)\right\|_M^2\\
&\quad+\gamma(\chi^k-\chi^\star)^\top(D(\chi^k,\chi^\star)^\top M+MD(\chi^k,\chi^\star))(\chi^k-\chi^\star),
\end{aligned}\tag{4.57}$$

where in (a) we use the relation $\chi^\star=\mathcal{P}_{\mathcal{X}}\left[\chi^\star-\gamma\mathcal{F}(\chi^\star)-\gamma\mathcal{H}(\chi^\star)\right]$, and the non-expansiveness property of the projection since X is closed and convex (cf. Assumption 4.1), in (b) we use (4.54), and in (c) we expand $\|\cdot\|_M^2$. In light of (4.53), selecting

$$c\coloneqq 20\overline{L}_1\left(\max\left\{\frac{\rho\kappa_2}{\mu},\frac{\overline{L}_1}{\mu}\right\}\right)^2\left(\max\left\{\frac{1}{\overline{L}_1\rho},\frac{\overline{L}_1}{\mu}\right\}\right)^2\frac{\kappa_2}{\kappa_1}$$

and $\tau\coloneqq\frac{\kappa_1}{2c}$, we can apply [159, Lemma 4] to $D(\chi^k,\chi^\star)$, obtaining

$$D(\chi^k,\chi^\star)^\top M+MD(\chi^k,\chi^\star)\leq-\tau M.\tag{4.58}$$

We then have that for any $\chi^k\in\mathbb{R}^{n+m}$,

$$\left\|D(\chi^k,\chi^\star)(\chi^k-\chi^\star)\right\|_M^2\leq\mu_1\left\|\chi^k-\chi^\star\right\|_M^2,\tag{4.59}$$

where $\mu_1\coloneqq\left(\max\left\{\overline{L}_1+\rho\left\|A\right\|^2,\frac{1}{\rho}\right\}\right)^2$ and the inequality follows by inspection of $D(\chi^k,\chi^\star)(\chi^k-\chi^\star)$ and using $\left\|E(\chi^k,\chi^\star)\right\|\leq 1$. Thus, we bound the right-hand side of (4.57) as

$$\left\|\mathcal{P}_{\mathcal{X}}\left[\chi^k-\gamma\mathcal{F}(\chi^k)-\gamma\mathcal{H}(\chi^k)\right]-\chi^\star\right\|_M^2\leq(1-\gamma\tau+\gamma^2\mu_1)\left\|\chi^k-\chi^\star\right\|_M^2.\tag{4.60}$$

Setting $\bar{\gamma}=\frac{\tau}{\mu_1}$, for any $\gamma\in(0,\bar{\gamma})$, we have that $0<1-\gamma\tau+\gamma^2\mu_1<1$. Therefore,

$$\left\|\mathcal{P}_{\mathcal{X}}\left[\chi^k-\gamma\mathcal{F}(\chi^k)-\gamma\mathcal{H}(\chi^k)\right]-\chi^\star\right\|_M\leq(1-\tilde{\mu})\left\|\chi^k-\chi^\star\right\|_M,\tag{4.61}$$

where $\tilde{\mu} \coloneqq 1 - \sqrt{1 - \gamma\tau + \gamma^2\mu_1} \in (0, 1)$.

Consider now $W : \mathbb{R}^{n+m} \to \mathbb{R}$ defined as

$$W(\chi) = (\chi - \chi^\star)^\top M(\chi - \chi^\star), \tag{4.62}$$

where M is as in (4.55). Since $M > 0$, W satisfies conditions (C.22a) and (C.22c). To show (C.22b) we evaluate $\Delta W(\chi^k) \coloneqq W(\chi^{k+1}) - W(\chi^k)$ along the trajectories of (4.48), obtaining

$$\begin{aligned}
\Delta W(\chi^k) &= \left\|\chi^k + \delta f(\chi^k, h(\chi^k)) - \chi^\star\right\|_M^2 - \left\|\chi^k - \chi^\star\right\|_M^2 \\
&\overset{(a)}{\leq} \left\|\chi^k + \delta\left(\mathcal{P}_{\mathcal{X}}\left[\chi^k - \gamma\mathcal{F}(\chi^k) - \gamma\mathcal{H}(\chi^k)\right] - \chi^k - \chi^\star\right)\right\|_M^2 - \left\|\chi^k - \chi^\star\right\|_M^2 \\
&\overset{(b)}{\leq} (1-\delta)^2\left\|\chi^k - \chi^\star\right\|_M^2 - \left\|\chi^k - \chi^\star\right\|_M^2 \\
&\quad + 2(\delta - \delta^2)\left\|\chi^k - \chi^\star\right\|_M \left\|\mathcal{P}_{\mathcal{X}}\left[\chi^k - \gamma\mathcal{F}(\chi^k) - \gamma\mathcal{H}(\chi^k)\right] - \chi^\star\right\|_M \\
&\quad + \delta^2\left\|\mathcal{P}_{\mathcal{X}}\left[\chi^k - \gamma\mathcal{F}(\chi^k) - \gamma\mathcal{H}(\chi^k)\right] - \chi^\star\right\|_M^2 \\
&\overset{(c)}{\leq} (1-\delta)^2\left\|\chi^k - \chi^\star\right\|_M^2 - \left\|\chi^k - \chi^\star\right\|_M^2 + 2(\delta - \delta^2)(1-\tilde{\mu})\left\|\chi^k - \chi^\star\right\|_M^2 \\
&\quad + \delta^2(1-\tilde{\mu})^2\left\|\chi^k - \chi^\star\right\|_M^2 \\
&\overset{(d)}{\leq} -\delta\tilde{\mu}(2 - \delta\tilde{\mu})\left\|\chi^k - \chi^\star\right\|_M^2,
\end{aligned} \tag{4.63}$$

where (a) uses the definitions of f and $\mathcal{P}_{\mathcal{X}}$ (cf. (4.37b), (4.56)) to explicitly write the update, in (b) we expand the squared norm, (c) follows by (4.61), while in (d) we rearrange the terms. Setting $\bar{\delta} \coloneqq 2/\tilde{\mu}$, (4.63) ensures that for any $\delta \in (0, \bar{\delta})$, W satisfies (C.22b), and the proof follows. ■

5. Lipschitz continuity of f, g and h: As we will be invoking Theorem C.2, we need to ensure that the required Lipschitz continuity assumptions are satisfied. In particular, we need to show that f, g in (4.37b) and (4.38), respectively, and h in (4.39) are Lipschitz with respect to their arguments. This is guaranteed by the Lipschitz continuity of the aggregation rules and the gradients of the cost functions (cf. Assumption 4.2), the nonexpansiveness of the projection operator (since X is closed and convex, see Assumption 4.1), and the Lipschitz continuity of G_x and G_λ (that appear in f and g), which is ensured as shown in (4.42) within the proof of Lemma4.3.

By combining Lemmas 4.3 and 4.4 with the Lipschitz continuity properties expressed above, Theorem C.2 can be applied. Then, there exists $\bar{\delta} \in (0, \min(\bar{\delta}_1, \bar{\delta}_2))$ so that, for any $\delta \in (0, \bar{\delta})$, $\text{col}(x^\star, \lambda^\star, h(x^\star, \lambda^\star))$ is an exponentially stable equilibrium point for (4.36). □

4.2.3 Numerical Simulations

We demonstrate the efficacy of Primal TRADES and Primal-Dual TRADES, and compare them with the most closely related distributed equilibrium seeking algorithms from the literature. First, we consider the case with local constraints only, and then we focus also on problems with coupling constraints.

Example without coupling constraints

In this subsection, we consider an instance of problem (4.4) and perform a numerical simulations in which we compare Primal TRADES with Algorithm 2 proposed in [151].

We consider the multi-agent demand response problem considered in [151]. Consider N loads whose electricity consumption $\mathrm{x}_i \coloneqq \mathrm{col}(x_{i,1}, \dots, x_{i,T}) \in \mathbb{R}^T$ with $T \in \mathbb{N}$ has to be chosen to solve

$$\forall i \in \{1, \dots, N\} : \min_{x_i \in X_i} \ \rho_i \left\| x_i - \hat{u}_i \right\|^2 + (\lambda \sigma(x) + p_0)^\top x_i,$$

where $\hat{u}_i \in \mathbb{R}^T$ denotes some nominal energy profile, $\rho_i > 0$ is a constant weighting parameter, the term $\lambda\sigma(x) + p_0$ with $\lambda \in \mathbb{R}$, $p_0 \in \mathbb{R}^T$ models the unit price which is taken to be an affine increasing function of the aggregate (average) energy demand $\sigma(x) = (1/N)\sum_{i=1}^N x_i$. As for the local feasible set $X_i \subseteq \mathbb{R}^T$, for all $i \in \{1, \dots, N\}$, we pick

$$X_i \coloneqq \left\{ x_i \in \mathbb{R}^T \mid s_{i,\tau+1}(x_i) \in \mathcal{S}_i \text{ and } x_{i,\tau} \in \mathcal{U}_i \ \forall \tau \in \{1, \dots, T\}, \sum_{\tau=1}^T x_{i,\tau} = \sum_{\tau=1}^T \hat{u}_{i,\tau} \right\},$$

where $\mathcal{U}_i \subseteq \mathbb{R}$, $\mathcal{S}_i \subseteq \mathbb{R}$, and $s_{i,\tau}(\mathrm{x}_i)$ is the state of the i-th load at time τ that, given the parameters $a_i, b_i \in \mathbb{R}$, is computed according to the linear dynamics

$$s_{i,\tau} = a_i^{\tau-1} s_{i,1} + \sum_{t=1}^{\tau-1} a^{k-1} b_i x_{i,\tau-t},$$

where $s_{i,1} \in \mathcal{S}_i$ is the initial condition of the state of the i-th load. To instantiate the problem, we set $T = 24$ and randomly generate values for $\hat{u}_i$, ρ_i, λ, p_0, a_i, b_i, $s_{i,1}$ and initial strategies $\mathrm{x}_{i,1}$ from uniform distributions. As for the sets $\mathcal{U}_i$ and $\mathcal{S}_i$, we pick the intervals $[0, 1]$ and $[0, 10]$, respectively. We consider a network with $N = 10$ players communicating according to an undirected, connected Erdős-Rényi graph with parameter 0.3.

This setting satisfies our Assumptions. We compare our scheme, namely, Primal TRADES with Algorithm 2 in [151]. We tune the latter with $v_1 = v_2 = 50$ communication rounds per iterate and update the auxiliary variable z^k according to $\mathrm{z}^{k+1} = (1 - \lambda)\mathrm{z}^k +$

$\lambda \mathcal{A}_{v_1,v_2}$ with $\lambda = 0.01$ (the quantity $\mathcal{A}_{v_1,v_2}$ is a proxy for the unavailable aggregative variable $\sigma(x)$, see [151] for more details.) As for the parameters of our scheme, we set $\delta = \gamma = 0.1$. Figure 4.1 shows the evolution of the normalized distance $\left\|\mathrm{x}^k - x^\star\right\| / \left\|x^\star\right\|$ from the NE $x^\star$ as the communication rounds (corresponding to iterations) progress. Our algorithm exhibits faster convergence and achieves higher accuracy in the calculation of the equilibrium $x^\star$. This was anticipated as the method in [151] is not guaranteed to converge to the exact NE.

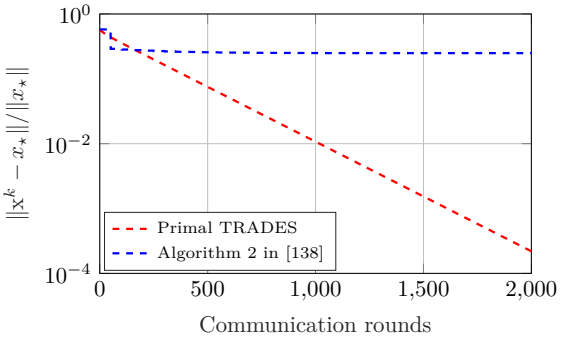


Figure 4.1: Comparison in terms of the normalized distance of the iterates from the NE between Primal TRADES (Algorithm 8) and the algorithm by [151], on a case study introduced in [151].

Example with coupling constraints

We address two Nash-Cournot games formulated as in (4.1) to compare our Primal-Dual TRADES algorithm with the distributed methods proposed in [11] and [69]. For a fair comparison we test the scheme by [11] with a constant step-size even if convergence was theoretically proven only with a diminishing one; note that slower convergence is expected by using a diminishing step-size.

We first compare the algorithms in [11] and [69] with Algorithm 9 on the case study already presented in Section 1.4.2, i.e., the one from [11] that we recall as follwos. Consider N firms that compete over n_m markets. In particular, for each market $\tau \in \mathcal{M} := \{1, \dots, n_m\}$, firm i is characterized by a production $g_{i,k} \geq 0$ and sales $s_{i,\tau} \geq 0$. For each $i \in \{1, \dots, N\}$ and $\tau \in \mathcal{M}$, the cost of production amounts to

$$f_{i,\tau}(g_{i,\tau}) = q_{i,\tau} g_{i,\tau}^2 + c_{i,\tau} g_{i,\tau}.$$

The revenue of firm i at market τ is modelled as $(d_\tau - \bar{s}_\tau) s_{i,\tau}$, where $d_\tau > 0$ is the total demand for location τ, and $\bar{s}_\tau := \sum_{i \in \{1,\dots,N\}} s_{i,\tau}$ represents the aggregate sales at location τ. For all firms $i \in \{1, \dots, N\}$ and markets $\tau \in \mathcal{M}$, we assume a production limitation $u_{i,\tau}$. Moreover, in each market τ, the total production $\sum_{i \in \{1,\dots,N\}} g_{i,\tau}$ must cover the demand d_τ without exceeding a maximum capacity r_τ. We can thus cast this setting as an instance of the GNEP in (4.1) with each strategy vector given by

$\mathrm{x}_i \coloneqq \mathrm{col}(g_{i,1}, \ldots, g_{i,n_m}, s_{i,1}, \ldots, s_{i,n_m}) \in \mathbb{R}^{2n_m}$, and cost function

$$J_i(x_i, \sigma(x)) = x_i^\top Q_i x_i + \ell_i^\top x_i + (\Delta\sigma(x))^\top x_i,$$

where we introduce the symbols $Q_i \coloneqq \mathrm{diag}(q_{i,1}, \ldots, q_{i,n_m}, 0, \ldots, 0) \in \mathbb{R}^{2n_m \times 2n_m}$, $\ell_i \coloneqq \mathrm{col}(c_{i,1}, \ldots, c_{i,n_m}, -d_1, \ldots, -d_{n_m}) \in \mathbb{R}^{2n_m}$, $\Delta = \mathrm{blkdiag}(0_{n_m}, NI_{n_m})$, and set the aggregation rule as $\phi_i(x_i) = x_i$ for all $i \in \{1, \ldots, N\}$. As for the constraints, for all $i \in \{1, \ldots, N\}$, we have the local constraint set

$$X_i \coloneqq \left\{ x_i \in \mathbb{R}^{2n_m} \mid \begin{bmatrix} -1_{2n_m}^\top \\ 1_{2n_m}^\top \end{bmatrix} x_i \leq 0, 0 \leq g_{i,\tau} \leq u_{i,\tau}, 0 \leq s_{i,\tau}, \tau = 1, \ldots, n_m \right\},$$

while the coupling constraints are defined by

$$A_i \coloneqq \begin{bmatrix} I_{n_m} & 0_{n_m} \\ -I_{n_m} & 0_{n_m} \end{bmatrix}, \quad b_i \coloneqq \frac{1}{N} \begin{bmatrix} r_1 & \ldots & r_{n_m} & -d_1 & \ldots & -d_{n_m} \end{bmatrix}^\top.$$

Following [11], we choose $N = 20$, $n_m = 10$, an undirected and connected graph with doubly stochastic weighted adjacency matrix chosen according to the Metropolis rule, and we generate values for the parameters of the problem from uniform distributions. Note that this game satisfies our Assumptions. In particular, for all $i \in \{1, \ldots, N\}$ and $k \in \mathcal{M}$, we pick $q_{i,k} \in [2, 3]$, $c_{i,k} \in [2, 12]$, $u_{i,k} \in [50, 100]$, $d_k \in [90, 100]$, and $r_k \in [d_k, 2d_k]$. We tune the algorithm as suggested in [11], i.e., with $\delta = \min(1, 1/L)$, $\tau = 1.05/(2\delta)$, $\gamma = 1$, $\alpha_i \leq 0.95/(\|A_i + \tau\|)$, and $\overline{L}_i \leq 1/(\|A\| + \tau)$ for all $i \in \{1, \ldots, N\}$, where $L > 0$ denotes the Lipschitz constant of the pseudo-gradient of the problem. To instantiate the algorithm in [69], we choose $c = 4$, $k = 1/200$, $\tau = 1/800$, $\alpha = 1/120$, and $v = 1/120$, while we implement our scheme with $\delta = 0.25$, $\gamma = 0.01$, and $\rho = 0.1$. Figure 4.2 compares the performance of these algorithms with our proposed Algorithm 9 in terms of the normalized distance $\left\|\mathrm{x}^k - x^\star\right\| / \|x^\star\|$ from the GNE $x^\star$. We observe from Figure 4.2 that Algorithm 9 outperforms the others in terms of accuracy and convergence speed.

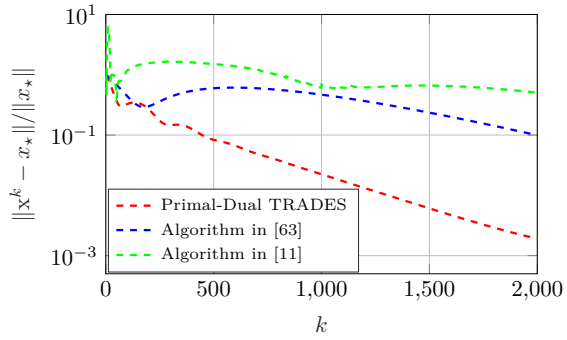


Figure 4.2: Comparison in terms of the normalized distance of the iterates from the GNE between Primal-Dual TRADES (Algorithm 9), and the algorithms by [11] and [69], on a case study introduced in [11].

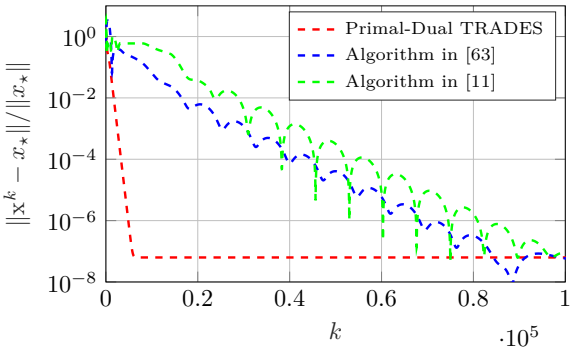


Figure 4.3: Comparison in terms of the normalized distance of the iterates from the GNE between Primal-Dual TRADES (Algorithm 9), and the algorithms by [11] and [69], on a case study introduced in [69].

We now focus on the case study considered in [69]. Specifically, we consider a Nash-Cournot game over a network for a single market with production constraints and globally coupling capacity constraints, which can be formulated as an instance of (4.1). In particular, we consider $N = 20$, $X_i = [0, 10]$ for all $i \in \{1, \dots, N\}$, $A = [1 \ \dots \ 1]$, $b = 20$, and the cost function

$$J_i(x_i, \sigma(x)) = (1 + 2(i-1))x_i - x_i \left(60 - \sigma(x) - \frac{1}{2}x_i\right).$$

As in [69], we consider a graph with ring topology. To achieve a fair comparison with [69], we follow the authors' tuning and choose $c = 4$, $k = 1/200$, $\tau = 1/800$, $\alpha = 1/120$, and $v = 1/120$, while we tune the scheme in [11] as above. As for the parameters of our algorithm, we empirically tune them as $\delta = \gamma = \rho = 0.1$. In Figure 4.3, we compare the performance of the algorithms in [11] and [69] with Algorithm 9 in terms of the normalized distance $\left\|\mathbf{x}^k - x^\star\right\| / \left\|x^\star\right\|$ from the GNE $x^\star$. Also in this case the proposed scheme exhibits faster convergence.

Conclusions

In this thesis, we addressed several challenges arising in peer-to-peer networks of agents with the aim of optimizing their local decision varlables with respect to a global cost (distributed optimization) or a private one (games). This thesis proposed a system theoretical point of view to design and analyze distributed algorithms to address this kind of problems. In detail, we first considered the consensus optimization framework and, in particular, we focused on the existing Gradient Tracking method. By resorting to an original singular perturbations perspective, we provided theoretical guarantees about its convergence in the case of nonconvex objective functions. Then, we extended the standard scheme to improve its performance and to deal with more complex frameworks, like, e.g., the case in which the gradients of the objective functions are not available or the one in which the objective functions vary over time. Subsequently, we tackled the recently emerged distributed aggregative optimization framework. In this field, we designed and analyzed novel algorithms arising in the online setup, the "personalized" optimization setting, and the feedback optimization framework. Finally, we considered the so-called aggregative games over networks. In this regard, we designed two fully distributed schemes to compute the Nash equilibrium of the game. In detail, the first algorithm has been designed to deal only with local constraints, while the second one has been tailored to deal also with linear coupling constraints among the decision variables of the agents of the game. In both cases, system theoretical arguments are provided to guarantee the effectiveness of the schemes.

Future research directions involve the extensions of the developed schemes and methodologies to more general stochastic frameworks. In particular, it would be interesting to generalize the investigated frameworks by allowing for uncertainties of the problem parameters that require learning-oriented techniques for which we only provided a first attempt. Such a contribution would definitely find applications for big data and deep learning purposes. Further, it may be also interesting to give more insights about the investigated feedback optimization setup, namely implementing the emerged methods on physical devices (like, e.g., mobile robots or drones), i.e., in closed-loop with their own dynamics. Moreover, it could be interesting to investigate the extension of our algorithms to a more general non monotone game setting. Finally,

the developed methodologies may also provide interesting contributions to the so-called constraint-coupled setup, i.e., another distributed optimization framework which has not been investigated in this thesis.

Appendix A

Optimization Basics

In this appendix, we provide some basic concepts and results about optimization.

Definition A.1 (Local and Global Minimum [13])**.** *Consider the function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$. Then, the point $x^\star \in \mathbb{R}^n$ is said to be a (strict) local minimum of f over X if there exists $\epsilon > 0$ such that it holds*

$$f(x^\star) \overset{(<)}{\leq} f(x), \tag{A.1}$$

for any $x \in X$ satisfying $\|x - x^\star\| \leq \epsilon$. Moreover, if the inequality (A.1) *holds for any $\mathrm{x} \in X$, then $x^\star$ is said to be the global minimum of f over X.* △

Definition A.2 (Radially unboundedness [14])**.** *Consider a closed function $f : \mathbb{R}^n \to \mathbb{R}$. Then, f is said to be radially unbounded if for any sequence $\{\mathrm{x}^k\}_{k\in\mathbb{N}}$ such that $\left\|\mathrm{x}^k\right\| \to \infty$ we have*

$$\lim_{k\to\infty} f(\mathrm{x}^k) = \infty.$$ △

Proposition A.1 ([14, Proposition 3.2.1])**.** *Consider the function $f : \mathbb{R}^n \to \mathbb{R}$. If f is radially unbounded, then the set of local minima of f over $\mathbb{R}^n$ is nonempty and compact.* △

Definition A.3 (Convexity [14])**.** *Consider the function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$. Then, f is said to be (strictly) convex if it holds*

$$f(x) \overset{(>)}{\geq} f(y) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|x - y\|^2,$$

for any $x, y \in X$. Moreover, given $\mu > 0$, f is said to be μ-strongly convex if it holds

$$f(x) \geq f(y) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|x - y\|^2,$$

for any $x, y \in X$. △

Proposition A.2 ([14, Proposition 3.1.1]). *If $X \subseteq \mathbb{R}^n$ is a convex subset and f is a convex function, then a local minimum of f over X is also a global minimum. If in addition f is strictly convex, then there exists the global minmum $x^\star \in \mathbb{R}^n$ of f over X.* △

Convergence rates

Let $\{\mathrm{x}^k\}_{k\in\mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^n$. Assume the sequence converges to some $\bar{x} \in \mathbb{R}^n$. We say that the sequence converges *linearly* to $\bar{x}$ if there exists a number $\eta \in (0,1)$ such that

$$\lim_{k\to\infty} \frac{\|\mathrm{x}^{k+1} - \bar{x}\|}{\|\mathrm{x}^k - \bar{x}\|} = \eta.$$

If $\eta = 0$ we say that the sequence converges *superlinearly*. It is common to denote the rate of convergence using the big-O notation. For instance, a sequence that goes to zero as $O(1/k)$ converges sublinearly, while a sequence that goes to zero as $O(\lambda^k)$, with $\lambda \in (0,1)$, converges linearly.

Appendix B

Auxiliary Results

In this appendix, we provide a series of auxiliary results that turn out to be useful in deriving some of the intermediate results of this thesis.

Lemma B.1 ([85, Theorem 6.3.12])**.** *Let M_0, $E \in \mathbb{R}^{n\times n}$ and let λ be a simple eigenvalue of M_0. Let v and w be, respectively, the right and left eigenvectors of M_0 corresponding to the eigenvalue λ. Then, for each $\epsilon > 0$, there exists a $\bar{\delta} > 0$ such that, for all $\delta \in \mathbb{R}$ with $|\delta| < \bar{\delta}$, there is a unique eigenvalue $\lambda(\delta)$ of $M_0 + \delta E$ such that*

$$\left|\lambda(\delta) - \lambda - \delta\frac{w^H E v}{w^H v}\right| \leq |\delta|\,\epsilon,$$

in which w^H denotes the Hermitian of w. Moreover $\lambda(\delta)$ is continuous at $\delta = 0$ and

$$\lim_{\delta\to 0}\lambda(\delta) = \lambda.$$

Moreover $\lambda(\delta)$ is differentiable at $\delta = 0$ and it holds

$$\left.\frac{d\lambda(\delta)}{d\delta}\right|_{\delta=0} = \frac{w^H E v}{w^H v}. \qquad \triangle$$

Lemma B.2. *Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex and with $\overline{L}$-Lipschitz continuous gradient. Let $x^\star \in \mathbb{R}^n$ its (unique) minimizer. Moreover, let $D \in \mathbb{R}^{n\times n}$ be positive definite diagonal matrix such that $D_{ii} \in [\epsilon, M]$ for all $i = 1, \ldots, n$ with $M \geq \epsilon > 0$ and $M < \infty$. Let $\overline{L}_M := M\overline{L}$ and $\mu_\epsilon = \epsilon\mu$. Let $\mathrm{x}^{k+1} = \mathrm{x}^k - \gamma D\nabla f(\mathrm{x}^k)$, with $\gamma \in (0, \frac{2}{\overline{L}_M}]$. Then $\|\mathrm{x}^{k+1} - x^\star\| \leq \max\{(1-\gamma\mu_\epsilon), (1-\gamma\overline{L}_M)\}\|\mathrm{x}^k - x^\star\|$.*

Proof. Let $h(x)$ be a function such that $\nabla h(x) = D\nabla f(x)$ for all x. It can be easily shown that h has $\overline{L}_M$-Lipschitz continuous gradients, in fact

$$\begin{aligned}\|\nabla h(x) - \nabla h(y)\| &= \|D\nabla f(x) - D\nabla f(y)\| \\ &\leq \|D\|\|\nabla f(x) - \nabla f(y)\| \leq \|D\|L\|x-y\| \leq ML\|x-y\|.\end{aligned}$$

Moreover h is μ_ϵ-strongly convex, since $\nabla^2 h(x) = D\nabla^2 f(x) \succeq D\sigma I \geq \epsilon\sigma I$. Define $g(x) = h(x) - \frac{\mu_\epsilon}{2}\|x\|^2$. Notice that, by definition, g is convex and with $(L-\mu_\epsilon)$-Lipschitz continuous gradient. Thus, by definition we have

$$\langle \nabla g(x) - \nabla g(y), x - y\rangle \geq \frac{1}{L-\mu_\epsilon}\|\nabla g(x) - \nabla g(y)\|^2. \tag{B.1}$$

Now, by using the definition of g one has

$$\langle \nabla h(x) - \mu_\epsilon x - \nabla h(y) + \mu_\epsilon y, x - y\rangle = \langle \nabla h(x) - \nabla h(y), x - y\rangle - \mu_\epsilon\|x - y\|^2. \tag{B.2}$$

Moreover

$$\begin{aligned}\|\nabla g(x) - \nabla g(y)\|^2 &= \|\nabla h(x) - \mu_\epsilon x - \nabla h(y) + \mu_\epsilon y\|^2 \\ &= \|\nabla h(x) - \nabla h(y)\|^2 + \mu_\epsilon^2\|x - y\|^2 - 2\mu_\epsilon\langle \nabla h(x) - \nabla h(y), x - y\rangle.\end{aligned} \tag{B.3}$$

By combining (B.1), (B.2), and (B.3) we get

$$\langle \nabla h(x) - \nabla h(y), x - y\rangle \geq \frac{\mu_\epsilon \overline{L}_M}{\mu_\epsilon + \overline{L}_M}\|x - y\|^2 + \frac{1}{\mu_\epsilon + \overline{L}_M}\|\nabla h(x) - \nabla h(y)\|^2. \tag{B.4}$$

Now, by using the update rule, one has

$$\begin{aligned}\|\mathrm{x}^{k+1} - x^\star\|^2 &= \|\mathrm{x}^k - \gamma D\nabla f(\mathrm{x}^k) - x^\star\|^2 \\ &= \|\mathrm{x}^k - x^\star\|^2 - 2\gamma\langle D\nabla f(\mathrm{x}^k), \mathrm{x}^k - x^\star\rangle + \gamma^2\|D\nabla f(\mathrm{x}^k)\|^2 \\ &= \|\mathrm{x}^k - x^\star\|^2 - 2\gamma\langle D\nabla f(\mathrm{x}^k) - D\nabla f(x^\star), \mathrm{x}^k - x^\star\rangle \\ &\quad + \gamma^2\|D\nabla f(\mathrm{x}^k) - D\nabla f(x^\star)\|^2.\end{aligned}$$

By using the result (B.4) with $\nabla h(x) = D\nabla f(x)$, we have

$$\begin{aligned}\|\mathrm{x}^{k+1} - x^\star\|^2 &\leq \|\mathrm{x}^k - x^\star\|^2 + \gamma^2\|D\nabla f(\mathrm{x}^k) - D\nabla f(x^\star)\|^2 \\ &\quad - 2\gamma\frac{\mu_\epsilon \overline{L}_M}{\mu_\epsilon + \overline{L}_M}\|\mathrm{x}^k - x^\star\|^2 - \frac{2\gamma}{\mu_\epsilon + \overline{L}_M}\|D\nabla f(\mathrm{x}^k) - D\nabla f(x^\star)\|^2 \\ &= \left(1 - 2\gamma\frac{\mu_\epsilon \overline{L}_M}{\mu_\epsilon + \overline{L}_M}\right)\|\mathrm{x}^k - x^\star\|^2 \\ &\quad + \gamma\left(\gamma - \frac{2}{\mu_\epsilon + \overline{L}_M}\right)\|D\nabla f(\mathrm{x}^k) - D\nabla f(x^\star)\|^2 \\ &\leq \left(1 - 2\gamma\frac{\mu_\epsilon \overline{L}_M}{\mu_\epsilon + \overline{L}_M}\right)\|\mathrm{x}^k - x^\star\|^2 + \gamma\left(\gamma\overline{L}_M^2 - \frac{2\mu_\epsilon^2}{\mu_\epsilon + \overline{L}_M}\right)\|\mathrm{x}^k - x^\star\|^2 \\ &\leq \max\{(1 - \gamma\mu_\epsilon)^2, (1 - \gamma\overline{L}_M)^2\}\|\mathrm{x}^k - x^\star\|^2.\end{aligned}$$

The proof follows by taking the square root of both sides. ■

Lemma B.3 (Contraction of strongly monotone operator)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be μ-strongly monotone and $\overline{L}$-Lipschitz continuous. If $\gamma \in (0, 2\mu/\overline{L}^2)$, then for any $x, x' \in \mathbb{R}^n$ it holds that*

$$\left\|x - \gamma F(x) - x' + \gamma F(x')\right\| \leq (1 - \tilde{\mu}) \left\|x - x'\right\|,$$

where $\tilde{\mu} \coloneqq 1 - \sqrt{1 - \gamma(2\mu - \gamma\overline{L}^2)} \in (0, 1]$. □

Proof. We have that

$$\begin{aligned}\left\|x - \gamma F(x) - x' + \gamma F(x')\right\|^2 &= \left\|x - x'\right\|^2 + \gamma^2 \left\|F(x) - F(x')\right\|^2 \\ &\quad - 2\gamma(x - x')^\top (F(x) - F(x')) \\ &\overset{(a)}{\leq} \left\|x - x'\right\|^2 - \gamma(2\mu - \gamma\overline{L}^2) \left\|x - x'\right\|^2, \qquad \text{(B.5)}\end{aligned}$$

where in (a) we use the strong monotonicity and the Lipschitz continuity of F. By construction, $\tilde{\mu} \in (0, 1]$ is equivalent to $\gamma(2\mu - \gamma\overline{L}^2) > 0$ and $\gamma(2\mu - \gamma L^2) \leq 1$. The former holds since $\gamma \in (0, 2\mu/\overline{L}^2)$. To see the latter, notice that, by definition of μ-strong monotonicity and L-Lipschitz continuity, we have

$$\mu \left\|x - x'\right\|^2 \leq (F(x) - F(x'))^\top (x - x') \leq \left\|F(x) - F(x')\right\| \left\|x - x'\right\| \leq \overline{L} \left\|x - x'\right\|^2,$$

for any x, x', hence $\mu \leq \overline{L}$. Thus, for any γ, it holds that $1 - 2\mu\gamma + \gamma^2 L^2 \geq 1 - 2\gamma L + \gamma^2 L^2 = (1 - \gamma L)^2 \geq 0$. △ ■

Appendix C

Discrete-Time Singularly Perturbed Systems

In this appendix, we consider the class of systems known in literature as singularly perturbed systems, i.e., the interconnection between two schemes referred to as slow and fast subsystem, respectively. A key feature of this class of systems is that the fast scheme has an equilibrium parametrized in the slow state. In the following, we provide results extending the ones existing in literature (see, e.g., [Proposition 9.1][19] for discrete-time or [91] for continuous-time). Although we explicitly apply the theorems of this appendix both in Chapter 2 and 4, they represent results that can be useful per se in the analysis of generic schemes given by the interconnection of two subsystems.

In detail, the next theorem provides a LaSalle's invariance principle for discrete-time singularly perturbed systems. First, we provide the definition of invariant set which turns out to be instrumental for such a theorem.

Definition C.1 (Invariant Set [124])**.** *Consider $\mathcal{M} \subseteq \mathbb{R}^n$ and $\mathrm{x}^{k+1} = T(\mathrm{x}^k)$, with $T : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathrm{x} \in \mathbb{R}^n$. Let $T(\mathcal{M}) := \{y \in \mathbb{R}^n \mid y = T(x) \text{ for some } x \in \mathcal{M}\}$. $\mathcal{M}$ is invariant if $T(\mathcal{M}) \equiv \mathcal{M}$.* △

Theorem C.1. *Consider the system*

$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, \zeta^k) \tag{C.1a}$$

$$\zeta^{k+1} = g(\zeta^k, \bar{\mathrm{x}}^k, \gamma), \tag{C.1b}$$

with $\bar{\mathrm{x}}^k \in \mathbb{R}^n$, $\zeta^k \in \mathbb{R}^m$, $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$, and $\gamma > 0$. Assume that ϕ and g are Lipschitz continuous in $\bar{\mathrm{x}}$ and ζ with parameters $\overline{L}_1 > 0$ and $\overline{L}_g(\gamma) > 0$, respectively, where $\overline{L}_g$ is continuous. Assume that there exists $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ such that $g(h(\bar{\mathrm{x}}, \gamma), \bar{\mathrm{x}}, \gamma) = h(\bar{\mathrm{x}}, \gamma)$ for any $\bar{\mathrm{x}} \in \mathbb{R}^n$ and that h is Lipschitz continuous in $\bar{\mathrm{x}}$ with

parameter $\overline{L}_h(\gamma) > 0$, where $\overline{L}_h$ is continuous. Let

$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h(\bar{\mathrm{x}}^k, \gamma)) \tag{C.2}$$

be the reduced system and

$$\psi^{k+1} = g(\psi^k + h(\bar{\mathrm{x}}, \gamma), \bar{\mathrm{x}}, \gamma) - h(\bar{\mathrm{x}}, \gamma) \tag{C.3}$$

be the boundary layer system with $\psi^k \in \mathbb{R}^m$. Assume that there exists $\bar{\gamma}_1 > 0$ such that, for any $\gamma \in (0, \bar{\gamma}_1)$, there exists a Lyapunov function $W : \mathbb{R}^m \to \mathbb{R}$ such that

$$b_1 \|\psi\|^2 \leq W(\psi) \leq b_2 \|\psi\|^2 \tag{C.4a}$$

$$W(g(\psi + h(\bar{\mathrm{x}}, \gamma), \bar{\mathrm{x}}, \gamma) - h(\bar{\mathrm{x}}, \gamma)) - W(\psi) \leq -b_3 \|\psi\|^2 \tag{C.4b}$$

$$|W(\psi_1) - W(\psi_2)| \leq b_4 \|\psi_1 - \psi_2\| (\|\psi_1\| + \|\psi_2\|), \tag{C.4c}$$

for any $\psi, \psi_1, \psi_2 \in \mathbb{R}^m$, $\bar{\mathrm{x}} \in \mathbb{R}^n$, and some $b_1, b_2, b_3, b_4 > 0$. Further, assume there exists $\bar{\gamma}_2 > 0$ and a radially unbounded function $U : \mathbb{R}^n \to \mathbb{R}$ such that

$$U(\bar{\mathrm{x}} + \gamma\phi(\bar{\mathrm{x}}, h(\bar{\mathrm{x}}, \gamma)) - U(\bar{\mathrm{x}}) \leq -\gamma c_1 \|\phi(\bar{\mathrm{x}}, h(\bar{\mathrm{x}}, \gamma))\|^2 \tag{C.5a}$$

$$U(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2) - U(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_3) \leq c_2 \|\phi(\bar{\mathrm{x}}_1, h(\bar{\mathrm{x}}_1, \gamma))\| \|\bar{\mathrm{x}}_2 - \bar{\mathrm{x}}_3\| + c_3 \left(\|\bar{\mathrm{x}}_2\|^2 + \|\bar{\mathrm{x}}_3\|^2 \right), \tag{C.5b}$$

for any $\gamma \in (0, \bar{\gamma}_2)$, $\bar{\mathrm{x}}, \bar{\mathrm{x}}_1, \bar{\mathrm{x}}_2, \bar{\mathrm{x}}_3 \in \mathbb{R}^n$, and some $c_1, c_2, c_3 > 0$. Then, there exists $\bar{\gamma} \in (0, \min\{\bar{\gamma}_1, \bar{\gamma}_2\})$ such that, for all $\gamma \in (0, \bar{\gamma})$, any trajectory of system (C.1) *satisfies*

$$\lim_{t \to \infty} \inf_{\xi \in \mathcal{M}} \left\| \begin{bmatrix} \bar{\mathrm{x}}^k \\ \zeta^k \end{bmatrix} - \begin{bmatrix} \xi \\ h(\xi, \gamma) \end{bmatrix} \right\| = 0,$$

where $\mathcal{M} \subseteq \ker\{\phi(\cdot, h(\cdot, \gamma))\} \subseteq \mathbb{R}^n$ denotes the largest invariant set for (C.2) *contained within* $\ker\{\phi(\cdot, h(\cdot, \gamma))\}$.

Proof. We start by defining $h_\gamma(\bar{\mathrm{x}}) := h(\bar{\mathrm{x}}, \gamma)$ and

$$\overline{L}_2 := \sup_{\gamma \in [0, \bar{\gamma}_3]} \{\overline{L}_g(\gamma)\}, \quad \overline{L}_3 := \sup_{\gamma \in [0, \bar{\gamma}_3]} \{\overline{L}_h(\gamma)\},$$

where $\bar{\gamma}_3 := \max\{\bar{\gamma}_1, \bar{\gamma}_2\}$ and both $\overline{L}_2$ and $\overline{L}_3$ are finite in light of the continuity of g and h, respectively. Thus, the global Lipschitz properties of g and h with parameters $\overline{L}_g(\gamma)$ and $\overline{L}_h(\gamma)$ lead to the Lipschitz property of g and h_γ with parameters $\overline{L}_2$ and $\overline{L}_3$ in the interval $[0, \bar{\gamma}_3]$. With this result at hand, define $\psi^k := \zeta^k - h_\gamma(\bar{\mathrm{x}}^k)$, and rewrite (C.1) as

$$\bar{\mathrm{x}}^{k+1} = \bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k)) \tag{C.6a}$$

$$\psi^{k+1} = g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^{k+1}). \tag{C.6b}$$

Pick the function U satisfying (C.5). Thus, by evaluating $\Delta U(\bar{\mathrm{x}}^k) := U(\bar{\mathrm{x}}^{k+1}) - U(\bar{\mathrm{x}}^k)$ along (C.6a), we get

$$\begin{aligned}
\Delta U(\bar{\mathrm{x}}^k) &= U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))) - U(\bar{\mathrm{x}}^k) \\
&\overset{(a)}{=} U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))) - U(\bar{\mathrm{x}}^k) \\
&\quad + U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))) - U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))) \\
&\overset{(b)}{\leq} -\gamma c_1 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\quad + U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))) - U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))) \\
&\overset{(c)}{\leq} -c_1 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\quad + \gamma c_2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k)) - \phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \\
&\quad + \gamma^2 c_3 \left(\left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 + \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2\right),
\end{aligned} \tag{C.7}$$

where in (a) we add and subtract $U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k)))$, in (b) we use (C.5a) to bound $U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))) - U(\bar{\mathrm{x}}^k)$, and in (c) we use (C.5b) to bound $U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))) - U(\bar{\mathrm{x}}^k + \gamma\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k)))$. Now, we add and subtract the term $\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))$ into $\left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))\right\|^2$ and thus, the right-hand side of (C.7) becomes

$$\begin{aligned}
\Delta U(\bar{\mathrm{x}}^k) &\leq -\gamma c_1 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 + \gamma c_2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k)) - \phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \\
&\quad + \gamma^2 c_3 \left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k)) - \phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k)) + \phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\quad + \gamma^2 c_3 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\overset{(a)}{\leq} -\gamma c_1 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 + \gamma c_2 \overline{L}_1 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \left\|\psi^k\right\| + \gamma^2 c_3 2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\quad + \gamma^2 c_3 \overline{L}_1^2 \left\|\psi^k\right\|^2 + \gamma^2 c_3 2 \overline{L}_1 \left\|\psi^k\right\| \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|,
\end{aligned} \tag{C.8}$$

where (a) exploits the square norm and the fact that ϕ is Lipschitz. Now, pick W satisfying (C.4). By evaluating $\Delta W(\psi^k) := W(\psi^{k+1}) - W(\psi^k)$, we get

$$\begin{aligned}
\Delta W(\psi^k) &= W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^{k+1})) - W(\psi^k) \\
&\overset{(a)}{=} W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^k)) - W(\psi^k) \\
&\quad + W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^{k+1})) \\
&\quad - W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^k))
\end{aligned}$$

$$\overset{(b)}{\leq} -b_3 \left\|\psi^k\right\|^2 + \tilde{W}(\psi^k, \bar{\mathrm{x}}^k), \tag{C.9}$$

where in (a) we add and subtract the term $W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^k))$ and in (b) we bound the term $W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k) - h_\gamma(\bar{\mathrm{x}}^k)) - W(\psi^k)$ by applying the result (C.4b) (which holds for any $\gamma \in (0, \bar{\gamma}_1)$) and introduce

$$\tilde{W}(\psi^k, \bar{\mathrm{x}}^k) := W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^{k+1})) - W(g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^k)).$$

By using (C.4c), we bound the above term as

$$\begin{aligned}
\tilde{W}(\psi^k, \bar{\mathrm{x}}^k) &\leq b_4 \left\|h_\gamma(\bar{\mathrm{x}}^{k+1}) - h_\gamma(\bar{\mathrm{x}}^k)\right\| \left\|g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^{k+1})\right\| \\
&\quad + b_4 \left\|h_\gamma(\bar{\mathrm{x}}^{k+1}) - h_\gamma(\bar{\mathrm{x}}^k)\right\| \left\|g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^k)\right\| \\
&\overset{(a)}{\leq} b_4 \left\|h_\gamma(\bar{\mathrm{x}}^{k+1}) - h_\gamma(\bar{\mathrm{x}}^k)\right\|^2 \\
&\quad + b_4 2 \left\|h_\gamma(\bar{\mathrm{x}}^{k+1}) - h_\gamma(\bar{\mathrm{x}}^k)\right\| \left\|g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h_\gamma(\bar{\mathrm{x}}^k)\right\| \\
&\overset{(b)}{\leq} b_4 \left\|h_\gamma(\bar{\mathrm{x}}^{k+1}) - h_\gamma(\bar{\mathrm{x}}^k)\right\|^2 + b_4 2 \left\|h_\gamma(\bar{\mathrm{x}}^{k+1}) - h_\gamma(\bar{\mathrm{x}}^k)\right\| \left\|\Delta g(\psi^k, \bar{\mathrm{x}}^k, \gamma)\right\| \\
&\overset{(c)}{\leq} \gamma^2 b_4 \overline{L}_3^2 \left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 + \gamma b_4 2 \overline{L}_3 \overline{L}_2 \left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))\right\| \left\|\psi^k\right\|,
\end{aligned} \tag{C.10}$$

where in (a) we add and subtract within the second norm $h_\gamma(\bar{\mathrm{x}})$ and use the triangle inequality, in (b) we add within the last norm $g(h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - h(\bar{\mathrm{x}}^k) = 0$ and introduce $\Delta g(\psi^k, \bar{\mathrm{x}}^k, \gamma) := g(\psi^k + h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma) - g(h_\gamma(\bar{\mathrm{x}}^k), \bar{\mathrm{x}}^k, \gamma)$, and in (c) we exploit the Lipschitz continuity of h_γ and g. Now, add and subtract $\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))$ in $\left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))\right\|^2$ and $\left\|\phi(\bar{\mathrm{x}}^k, \psi^k + h_\gamma(\bar{\mathrm{x}}^k))\right\|$, use the triangle inequality and the Lipschitz property of ϕ to bound (C.10) as

$$\begin{aligned}
\tilde{W}(\psi^k, \bar{\mathrm{x}}^k) &= \gamma^2 b_4 \overline{L}_3^2 \overline{L}_1^2 \left\|\psi^k\right\|^2 + \gamma^2 b_4 \overline{L}_3^2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\quad + \gamma^2 b_4 2 \overline{L}_3^2 \overline{L}_1 \left\|\psi^k\right\| \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| + \gamma b_4 2 \overline{L}_3 \overline{L}_2 \overline{L}_1 \left\|\psi^k\right\|^2 \\
&\quad + \gamma b_4 2 \overline{L}_3 \overline{L}_2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \left\|\psi^k\right\|.
\end{aligned} \tag{C.11}$$

Thus, we can use (C.11) to bound (C.9) as

$$\begin{aligned}
\Delta W(\psi^k) &\leq -b_3 \left\|\psi^k\right\|^2 + \gamma^2 b_4 \overline{L}_3^2 \overline{L}_1^2 \left\|\psi^k\right\|^2 + \gamma^2 b_4 \overline{L}_3^2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\|^2 \\
&\quad + \gamma^2 b_4 2 \overline{L}_3^2 \overline{L}_1 \left\|\psi^k\right\| \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| + \gamma b_4 2 \overline{L}_3 \overline{L}_2 \overline{L}_1 \left\|\psi^k\right\|^2 \\
&\quad + \gamma b_4 2 \overline{L}_3 \overline{L}_2 \left\|\phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k))\right\| \left\|\psi^k\right\|.
\end{aligned} \tag{C.12}$$

Now, define $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as

$$V(\bar{\mathrm{x}}^k, \psi^k) = U(\bar{\mathrm{x}}^k) + W(\psi^k).$$

Thus, by evaluating $\Delta V(\bar{\mathrm{x}}^k, \psi^k) := V(\bar{\mathrm{x}}^{k+1}, \psi^{k+1}) - V(\bar{\mathrm{x}}^k, \psi^k) = \Delta U(\bar{\mathrm{x}}^k) + \Delta W(\psi^k)$ along the trajectories of (C.6), we can use the results (C.8) and (C.12) to write

$$\Delta V(\bar{\mathrm{x}}^k, \psi^k) \leq - \begin{bmatrix} \left\| \phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k)) \right\| \\ \left\| \psi^k \right\| \end{bmatrix}^\top H \begin{bmatrix} \left\| \phi(\bar{\mathrm{x}}^k, h_\gamma(\bar{\mathrm{x}}^k)) \right\| \\ \left\| \psi^k \right\| \end{bmatrix}, \tag{C.13}$$

where $H \in \mathbb{R}^{2\times 2}$ denotes the symmetric matrix

$$H := \begin{bmatrix} \gamma c_1 - \gamma^2 k_1 & -\gamma k_2 - \gamma^2 k_3 \\ -\gamma k_2 - \gamma^2 k_3 & b_3 - \gamma k_4 - \gamma^2 k_5 \end{bmatrix},$$

in which the notation has been shortened through the constants

$$\begin{aligned} &k_1 := b_4 \overline{L}_3^2 + c_3 2, \;\; k_2 := \tfrac{c_2 \overline{L}_1 + b_4 2 \overline{L}_2 \overline{L}_3}{2}, \;\; k_3 := \tfrac{c_3 2 \overline{L}_1 + b_4 2 \overline{L}_1 \overline{L}_3^2}{2} \\ &k_4 := b_4 2 \overline{L}_1 \overline{L}_2 \overline{L}_3, \;\; k_5 := c_3 \overline{L}_1^2 + b_4 \overline{L}_1^2 \overline{L}_3^2. \end{aligned}$$

Being $H = H^\top$, by Sylvester Criterion, $H > 0$ if and only if

$$\begin{cases} \gamma c_1 > p_1(\gamma) \\ \gamma c_1 b_3 > p_2(\gamma), \end{cases} \tag{C.14}$$

where we have introduced the polynomials

$$\begin{aligned} &p_1(\gamma) := \gamma^2 k_1 \\ &p_2(\gamma) := \gamma^2 c_1 (k_4 + \gamma k_5) + \gamma^2 k_1 (b_3 - \gamma k_4 - \gamma^2 k_5) + (\gamma k_2 + \gamma^2 k_3)^2. \end{aligned} \tag{C.15}$$

We notice that $\lim_{\gamma \to 0} p_1(\gamma)/\gamma = \lim_{\gamma \to 0} p_2(\gamma)/\gamma = 0$. Thus, there exists $\bar{\gamma} \in (0, \min\{\bar{\gamma}_1, \bar{\gamma}_2\})$ such that, for any $\gamma \in (0, \bar{\gamma})$, the conditions in (C.14) hold leading to the positiveness of H. Hence (C.13) ensures that $\Delta V(\bar{\mathrm{x}}^k, \psi^k) \leq 0$ for any $\bar{\mathrm{x}}^k \in \mathbb{R}^n$ and any $\psi^k \in \mathbb{R}^m$. In particular, the right-hand side of (C.13) is null when $\bar{\mathrm{x}}^k \in E'$, where $E' \subseteq \mathbb{R}^{n+m}$ reads as

$$E' := \{(\bar{\mathrm{x}}, \psi) \in \mathbb{R}^{n+m} \mid \bar{\mathrm{x}} \in \ker\{\phi(\cdot, h_\gamma(\cdot))\}, \psi = 0\}. \tag{C.16}$$

Thus, we apply the LaSalle's invariance principle (cf. [70, Theorem 3.7]) to conclude that, for any $\gamma \in (0, \bar{\gamma})$, any trajectory of system (C.6) approaches

$$\lim_{t \to \infty} \inf_{\xi' \in \mathcal{M}'} \left\| \begin{bmatrix} \bar{\mathrm{x}}^k \\ \psi^k \end{bmatrix} - \xi' \right\| = 0, \tag{C.17}$$

where $\mathcal{M}' \subseteq E'$ denotes the largest invariant set for system (C.6) contained within the subspace E defined in (C.16). The proof follows by noticing that (i) $\mathcal{M}' \equiv E'$, and that (ii), turning out to the coordinates $(\bar{\mathrm{x}}, \zeta)$, the result (C.17) implies that, for any $\gamma \in (0, \bar{\gamma})$, any trajectory of (C.1) converges to $\mathcal{M} := \{(\bar{\mathrm{x}}, \zeta) \in \mathbb{R}^{n+m} \mid \bar{\mathrm{x}} \in \ker\{\phi(\cdot, h_\gamma(\cdot))\}, \zeta = h_\gamma(\bar{\mathrm{x}})\}$. ■

We now provide an extension of Theorem C.1. In particular, given the existence of a globally exponentially stable equilibrium point for the reduced system, we establish the conditions to guarantee the global exponential stability of an equilibrium point for the whole interconnected system.

Theorem C.2 (Global exponential stability for singularly perturbed systems)**.** *Consider the system*

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta f(\mathrm{x}^k, \mathrm{w}^k) \tag{C.18a}$$

$$\mathrm{w}^{k+1} = g(\mathrm{w}^k, \mathrm{x}^k, \delta), \tag{C.18b}$$

with $\mathrm{x}^k \in \mathcal{D} \subseteq \mathbb{R}^n$, $\mathrm{w}^k \in \mathbb{R}^m$, $f : \mathcal{D} \times \mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$, $\delta > 0$. *Assume that* f *and* g *are Lipschitz continuous with respect to both arguments with Lipschitz constants* $\overline{L}_f > 0$ *and* $\overline{L}_g > 0$, *respectively. Assume that there exists* $x^\star \in \mathbb{R}^n$ *and* $h : \mathbb{R}^n \to \mathbb{R}^m$ *such that for any* $\mathrm{x} \in \mathbb{R}^n$

$$0 = \delta f(x^\star, h(x^\star)),$$
$$h(\mathrm{x}) = g(h(\mathrm{x}), \mathrm{x}, \delta),$$

with h *being Lipschitz continuous with Lipschitz constant* $\overline{L}_h > 0$. *Let*

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta f(\mathrm{x}^k, h(\mathrm{x}^k)) \tag{C.19}$$

be the reduced system *and*

$$\psi^{k+1} = g(\psi^k + h(\mathrm{x}), \mathrm{x}, \delta) - h(\mathrm{x}) \tag{C.20}$$

be the boundary layer system *with* $\psi^k \in \mathbb{R}^m$.

Assume that there exists a continuous function $U : \mathbb{R}^m \to \mathbb{R}$ *and* $\bar{\delta}_1 > 0$ *such that, for any* $\delta \in (0, \bar{\delta}_1)$ *(cf.* (C.18)*), there exist* $b_1, b_2, b_3, b_4 > 0$ *such that for any* $\psi, \psi_1, \psi_2 \in \mathbb{R}^m$, $x \in \mathbb{R}^n$,

$$b_1 \|\psi\|^2 \leq U(\psi) \leq b_2 \|\psi\|^2 \tag{C.21a}$$

$$U(g(\psi + h(x), x, \delta) - h(x)) - U(\psi) \leq -b_3 \|\psi\|^2 \tag{C.21b}$$

$$|U(\psi_1) - U(\psi_2)| \leq b_4 \|\psi_1 - \psi_2\| \|\psi_1\| + b_4 \|\psi_1 - \psi_2\| \|\psi_2\|. \tag{C.21c}$$

Further, assume there exists a continuous function $W : \mathcal{D} \to \mathbb{R}$ and $\bar{\delta}_2 > 0$ such that, for any $\delta \in (0, \bar{\delta}_2)$, there exist $c_1, c_2, c_3, c_4 > 0$ such that for any $x, x_1, x_2, x_3 \in \mathcal{D}$

$$c_1 \|x - x^\star\|^2 \leq W(x) \leq c_2 \|x - x^\star\|^2 \tag{C.22a}$$

$$W(x + \delta f(x, h(x))) - W(x) \leq -\delta c_3 \|x - x^\star\|^2 \tag{C.22b}$$

$$|W(x_1) - W(x_2)| \leq c_4 \|x_1 - x_2\| \|x_1 - x^\star\| + c_4 \|x_1 - x_2\| \|x_2 - x^\star\|. \tag{C.22c}$$

Then, there exist $\bar{\delta} \in (0, \min\{\bar{\delta}_1, \bar{\delta}_2\})$, $a_1 > 0$, and $a_2 > 0$ such that, for all $\delta \in (0, \bar{\delta})$, it holds

$$\left\| \begin{bmatrix} \mathrm{x}^k - x^\star \\ \mathrm{w}^k - h(\mathrm{x}^k) \end{bmatrix} \right\| \leq a_1 \left\| \begin{bmatrix} \mathrm{x}^0 - x^\star \\ \mathrm{w}^0 - h(\mathrm{x}^0) \end{bmatrix} \right\| e^{-a_2 t},$$

for any $(\mathrm{x}^0, \mathrm{w}^0) \in \mathcal{D} \times \mathbb{R}^m$.

Proof.

Define $\tilde{\mathrm{w}}^k \coloneqq \mathrm{w}^k - h(\mathrm{x}^k)$ and, in accordance, rewrite system (C.18) as

$$\mathrm{x}^{k+1} = \mathrm{x}^k + \delta f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k)) \tag{C.23a}$$

$$\tilde{\mathrm{w}}^{k+1} = g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) + \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k), \tag{C.23b}$$

where $\Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k) \coloneqq -h(\mathrm{x}^{k+1}) + h(\mathrm{x}^k)$. Pick W as in (C.22). By evaluating $\Delta W(\mathrm{x}^k) \coloneqq W(\mathrm{x}^{k+1}) - W(\mathrm{x}^k)$ along the trajectories of (C.23a), we obtain

$$\begin{aligned}
\Delta W(\mathrm{x}^k) &= W(\mathrm{x}^k + \delta f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k))) - W(\mathrm{x}^k) \\
&\overset{(a)}{=} W(\mathrm{x}^k + \delta f(\mathrm{x}^k, h(\mathrm{x}^k))) - W(\mathrm{x}^k) + W(\mathrm{x}^k + \delta f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k))) \\
&\quad - W(\mathrm{x}^k + \delta f(\mathrm{x}^k, h(\mathrm{x}^k))) \\
&\overset{(b)}{\leq} -\delta c_3 \left\| \mathrm{x}^k - x^\star \right\|^2 + W(\mathrm{x}^k + \delta f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k))) - W(\mathrm{x}^k + \delta f(\mathrm{x}^k, h(\mathrm{x}^k))) \\
&\overset{(c)}{\leq} -\delta c_3 \left\| \mathrm{x}^k - x^\star \right\|^2 + 2\delta c_4 \overline{L}_f \left\| \tilde{\mathrm{w}}^k \right\| \left\| \mathrm{x}^k - x^\star \right\| + \delta^2 c_4 \overline{L}_f \left\| \tilde{\mathrm{w}}^k \right\| \left\| f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k)) \right\| \\
&\quad + \delta^2 c_4 \overline{L}_f \left\| \tilde{\mathrm{w}}^k \right\| \left\| f(\mathrm{x}^k, h(\mathrm{x}^k)) \right\|,
\end{aligned} \tag{C.24}$$

where in (a) we add and subtract the term $W(\mathrm{x}^k + \delta f(\mathrm{x}^k, h(\mathrm{x}^k)))$, in (b) we exploit (C.22b) to bound the difference of the first two terms, in (c) we use (C.22c), the Lipschitz continuity of f, and the triangle inequality. By recalling that $f(x^\star, h(x^\star)) = 0$ we can thus write

$$\|f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k))\| = \left\| f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k)) - f(x^\star, h(x^\star)) \right\|$$

$$\begin{aligned}
&\overset{(a)}{\leq} \overline{L}_f \left\| \mathrm{x}^k - x^\star \right\| + \overline{L}_f \left\| \tilde{\mathrm{w}}^k + h(\mathrm{x}^k) - h(x^\star) \right\|, \\
&\overset{(b)}{\leq} \overline{L}_f (1 + \overline{L}_h) \left\| \mathrm{x}^k - x^\star \right\| + \overline{L}_f \left\| \tilde{\mathrm{w}}^k \right\|, \qquad \text{(C.25)}
\end{aligned}$$

where in (a) we use the Lipschitz continuity of f and h, and in (b) we use the Lipschitz continuity of h together with the triangle inequality. With similar arguments, we have

$$\left\| f(\mathrm{x}^k, h(\mathrm{x}^k)) \right\| \leq \overline{L}_f (1 + \overline{L}_h) \left\| \mathrm{x}^k - x^\star \right\|. \qquad \text{(C.26)}$$

Using inequalities (C.25) and (C.26) we then bound (C.24) as

$$\begin{aligned}
\Delta W(\mathrm{x}^k) &\leq -\delta c_3 \left\| \mathrm{x}^k - x^\star \right\|^2 + 2\delta c_4 \overline{L}_f \left\| \tilde{\mathrm{w}}^k \right\| \left\| \mathrm{x}^k - x^\star \right\| + \delta^2 c_4 \overline{L}_f^2 \left\| \tilde{\mathrm{w}}^k \right\|^2 \\
&\quad + 2\delta^2 c_4 \overline{L}_f^2 (1 + \overline{L}_h) \left\| \tilde{\mathrm{w}}^k \right\| \left\| \mathrm{x}^k - x^\star \right\| \\
&\leq -c_3 \left\| \mathrm{x}^k - x^\star \right\|^2 + \delta^2 k_3 \left\| \tilde{\mathrm{w}}^k \right\|^2 + (\delta k_1 + \delta^2 k_2) \left\| \tilde{\mathrm{w}}^k \right\| \left\| \mathrm{x}^k - x^\star \right\|, \qquad \text{(C.27)}
\end{aligned}$$

where we introduce the constants

$$k_1 \coloneqq 2c_4 \overline{L}_f, \quad k_2 \coloneqq 2c_4 \overline{L}_f^2 (1 + \overline{L}_h), \quad k_3 \coloneqq c_4 \overline{L}_f^2.$$

We now pick U as in (C.21). By evaluating $\Delta U(\tilde{\mathrm{w}}^k) \coloneqq U(\tilde{\mathrm{w}}^{k+1}) - U(\tilde{\mathrm{w}}^k)$ along the trajectories of (C.23b), we obtain

$$\begin{aligned}
\Delta U(\tilde{\mathrm{w}}) &= U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) + \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k)) - U(\tilde{\mathrm{w}}^k) \\
&\overset{(a)}{\leq} U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k)) - U(\tilde{\mathrm{w}}^k) \\
&\quad - U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k)) \\
&\quad + U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) + \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k)) \\
&\overset{(b)}{\leq} -b_3 \left\| \tilde{\mathrm{w}}^k \right\|^2 - U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k)) \\
&\quad + U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) + \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k)) \\
&\overset{(c)}{\leq} -b_3 \left\| \tilde{\mathrm{w}}^k \right\|^2 + b_4 \left\| \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k) \right\| \left\| g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) + \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k) \right\| \\
&\quad + b_4 \left\| \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k) \right\| \left\| g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) \right\| \\
&\overset{(d)}{\leq} -b_3 \left\| \tilde{\mathrm{w}}^k \right\|^2 + b_4 \left\| \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k) \right\|^2 \\
&\quad + 2b_4 \left\| \Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k) \right\| \left\| g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k) \right\|, \qquad \text{(C.28)}
\end{aligned}$$

where in (a) we add and subtract $U(g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k))$, in (b) we exploit (C.21b) to bound the first two terms, in (c) we use (C.21c) to bound the the difference of the

last two terms, and in (d) we use the triangle inequality. By exploiting the definition of $\Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k)$ and the Lipschitz continuity of h, we have that

$$\begin{aligned}\left\|\Delta h(\mathrm{x}^{k+1}, \mathrm{x}^k)\right\| &\leq \overline{L}_h \left\|\mathrm{x}^{k+1} - \mathrm{x}^k\right\| \\ &\overset{(a)}{\leq} \delta \overline{L}_h \left\|f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k))\right\| \\ &\overset{(b)}{\leq} \delta \overline{L}_h \left\|f(\mathrm{x}^k, \tilde{\mathrm{w}}^k + h(\mathrm{x}^k)) - f(x^\star, h(x^\star))\right\| \\ &\overset{(c)}{\leq} \delta \overline{L}_h \overline{L}_f (1 + \overline{L}_h) \left\|\mathrm{x}^k - x^\star\right\| + \delta \overline{L}_h \overline{L}_f \left\|\tilde{\mathrm{w}}^k\right\|, \end{aligned} \tag{C.29}$$

where in (a) we use the update (C.23a), in (b) we add the term $f(x^\star, h(x^\star))$ since this is zero, and in (c) we use the triangle inequality and the Lipschitz continuity of f and h. Moreover, since $g(h(\mathrm{x}^k), \mathrm{x}^k, \delta) = h(\mathrm{x}^k)$, we obtain

$$\left\|g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - h(\mathrm{x}^k)\right\| = \left\|g(\tilde{\mathrm{w}}^k + h(\mathrm{x}^k), \mathrm{x}^k, \delta) - g(h(\mathrm{x}^k), \mathrm{x}^k, \delta)\right\| \leq \overline{L}_g \left\|\tilde{\mathrm{w}}^k\right\|, \tag{C.30}$$

where the inequality is due to the Lipschitz continuity of g. Using inequalities (C.29) and (C.30), we then bound (C.28) as

$$\begin{aligned}\Delta U(\tilde{\mathrm{w}}) \leq & -b_3 \left\|\tilde{\mathrm{w}}^k\right\|^2 + 2\delta b_4 \overline{L}_h \overline{L}_g \overline{L}_f (1 + \overline{L}_h) \left\|\mathrm{x}^k - x^\star\right\| \left\|\tilde{\mathrm{w}}^k\right\| \\ & + 2\delta b_4 \overline{L}_h \overline{L}_g \overline{L}_f \left\|\tilde{\mathrm{w}}^k\right\|^2 + \delta^2 b_4 \overline{L}_h^2 \overline{L}_f^2 (1 + \overline{L}_h)^2 \left\|\mathrm{x}^k - x^\star\right\|^2 \\ & + 2\delta^2 b_4 \overline{L}_h^2 \overline{L}_f^2 (1 + \overline{L}_h) \left\|\mathrm{x}^k - x^\star\right\| \left\|\tilde{\mathrm{w}}^k\right\| + \delta^2 b_4 \overline{L}_h^2 \overline{L}_f^2 \left\|\tilde{\mathrm{w}}^k\right\|^2 \\ \leq & (-b_3 + \delta k_6 + \delta^2 k_7) \left\|\tilde{\mathrm{w}}^k\right\|^2 + \delta^2 k_8 \left\|\mathrm{x}^k - x^\star\right\|^2 \\ & + (\delta k_4 + \delta^2 k_5) \left\|\mathrm{x}^k - x^\star\right\| \left\|\tilde{\mathrm{w}}^k\right\|, \end{aligned} \tag{C.31}$$

where we introduce the constants

$$\begin{aligned} k_4 &\coloneqq 2b_4 \overline{L}_h \overline{L}_g \overline{L}_f (1 + \overline{L}_h), & k_5 &\coloneqq 2b_4 \overline{L}_h^2 \overline{L}_f^2 (1 + \overline{L}_h), \\ k_6 &\coloneqq 2b_4 \overline{L}_h \overline{L}_g \overline{L}_f, & k_7 &\coloneqq b_4 \overline{L}_h^2 \overline{L}_f^2, \\ k_8 &\coloneqq b_4 \overline{L}_h^2 \overline{L}_f^2 (1 + \overline{L}_h)^2. \end{aligned}$$

We pick the following Lyapunov candidate $V : \mathcal{D} \times \mathbb{R}^m \to \mathbb{R}$:

$$V(\mathrm{x}^k, \tilde{\mathrm{w}}^k) = W(\mathrm{x}^k) + U(\tilde{\mathrm{w}}^k).$$

By evaluating $\Delta V(\mathrm{x}^k, \tilde{\mathrm{w}}^k) \coloneqq V(\mathrm{x}^{k+1}, \tilde{\mathrm{w}}^{k+1}) - V(\mathrm{x}^k, \tilde{\mathrm{w}}^k) = \Delta W(\mathrm{x}^k) + \Delta U(\tilde{\mathrm{w}}^k)$ along the

trajectories of (C.23), we can use the results (C.27) and (C.31) to write

$$\Delta V(\mathrm{x}^k, \tilde{\mathrm{w}}^k) \leq -\begin{bmatrix} \left\|\mathrm{x}^k - x^\star\right\| \\ \left\|\tilde{\mathrm{w}}^k\right\| \end{bmatrix}^\top Q(\delta) \begin{bmatrix} \left\|\mathrm{x}^k - x^\star\right\| \\ \left\|\tilde{\mathrm{w}}^k\right\| \end{bmatrix}, \tag{C.32}$$

where we define the matrix $Q(\delta) = Q(\delta)^\top \in \mathbb{R}^2$ as

$$Q(\delta) \coloneqq \begin{bmatrix} \delta c_3 - \delta^2 k_8 & q_{21}(\delta) \\ q_{21}(\delta) & b_3 - \delta k_6 - \delta^2(k_3 + k_7) \end{bmatrix},$$

with $q_{21}(\delta) \coloneqq -\frac{1}{2}(\delta(k_1 + k_4) + \delta^2(k_2 + k_5))$. By relying on the Sylvester criterion [91], we know that $Q \succ 0$ if and only if

$$\delta c_3 b_3 > p(\delta) \tag{C.33}$$

where the polynomial $p(\delta)$ is defined as

$$p(\delta) \coloneqq q_{21}(\delta)^2 + \delta^2 c_3 k_6 + \delta^3 c_3(k_3 + k_7) + \delta^2 b_3 k_8 - \delta^3 k_6 k_8 - \delta^4 k_8(k_3 + k_7). \tag{C.34}$$

We note that p is a continuous function of δ and $\lim_{\delta \to 0} p(\delta)/\delta = 0$. Hence, there exists some $\bar{\delta} \in (0, \min\{\bar{\delta}_1, \bar{\delta}_2\})$ – recall that $\bar{\delta}_1$ and $\bar{\delta}_2$ exist as U and W are taken to satisfy (C.21) and (C.22) – so that (C.33) is satisfied for any $\delta \in (0, \bar{\delta})$. Under such a choice of δ, and denoting by $q > 0$ the smallest eigenvalue of $Q(\delta)$, we can bound (C.32) as

$$\Delta V(\mathrm{x}^k, \tilde{\mathrm{w}}^k) \leq -q \left\| \begin{bmatrix} \left\|\mathrm{x}^k - x^\star\right\| \\ \left\|\tilde{\mathrm{w}}^k\right\| \end{bmatrix} \right\|^2,$$

which allows us to conclude, in view of [40, Theorem 13.2], that $(x^\star, 0)$ is an exponentially stable equilibrium point for system (C.23). The theorem's conclusion follows then by considering the definition of exponentially stable equilibrium point and by reverting to the original coordinates $(\mathrm{x}^k, \mathrm{w}^k)$. ■

Ringraziamenti Personali

In questi anni, da studente di Automatica, ho appreso il ruolo chiave degli input per governare l'evoluzione di un sistema. Per questo motivo, il ringraziamento più grande e più sentito va senz'altro a Giuseppe e a tutti i preziosi consigli che mi ha dato in questi anni: dall'imbeccata sul passaggio più ostico di una dimostrazione fino alla dritta sul più piccolo dettaglio di questo mondo così strambo e coinvolgente. É principalmente grazie a questi input se, nonostante tutti i disturbi e le nonidealità che influenzano il sottoscritto, ho raggiunto i risultati raccontati in questo lavoro.

Naturalmente, dopo aver stressato per 200 pagine il potenziale di una rete cooperante e interconnessa, viene da sé ringraziare tutte le persone con cui ho interagito in questi anni. Con questa rete d'interazioni, tormentarsi per dare un contributo a queste sfide tanto appassionanti quanto impegnative, è stato molto più facile e soddisfacente. Con la competizione si va all'equilibrio di Nash, ma con la cooperazione si va all'ottimo! Grazie al CASY!

Riorganizzare il lavoro di questi tre anni è stato un po' come sfogliare un album di vecchie foto e rivivere le emozioni già provate al momento dello scatto: dall'articolo concepito in pieno lockdown, fino a quello nato durante il periodo di visita all'Università di Oxford (grazie Kostas! Grazie Fili's!). È quindi evidente come questo lavoro sia anche intrinsecamente legato alle relazioni e alle sensazioni estranee all'ambito lavorativo. Per questo motivo, l'ultimo ma sincero ringraziamento va a tutte le persone importanti che mi hanno tenuto per mano durante i passi di questo percorso.

Bibliography

[1] T Adachi, N Hayashi, and S Takai, *Distributed gradient descent method with edge-based event-driven communication for non-convex optimization*, IET Control Theory & Applications **15** (2021), no. 12, 1588–1598.

[2] M. Akbari, B. Gharesifard, and T. Linder, *Distributed online convex optimization on time-varying directed graphs*, IEEE Transactions on Control of Network Systems **4** (2015), no. 3, 417–428.

[3] _____, *Individual regret bounds for the distributed online alternating direction method of multipliers*, IEEE Transactions on Automatic Control **64** (2019), no. 4, 1746–1752.

[4] M. B Alatise and G. P Hancke, *A review on challenges of autonomous mobile robot and sensor fusion methods*, IEEE Access **8** (2020), 39830–39846.

[5] K. B Ariyur and M. Krstic, *Real-time optimization by extremum-seeking control*, John Wiley & Sons, 2003.

[6] M. Babazadeh and A. Nobakhti, *Sparsity promotion in state feedback controller design*, IEEE Transactions on Automatic Control **62** (2016), no. 8, 4066–4072.

[7] J. Barrera and A. Garcia, *Dynamic incentives for congestion control*, IEEE Transactions on Automatic Control **60** (2014), no. 2, 299–310.

[8] G. Belgioioso, W. Ananduta, S. Grammatico, and C. Ocampo-Martinez, *Energy management and peer-to-peer trading in future smart grids: A distributed game-theoretic approach*, 2020 European Control Conference (ECC), 2020, pp. 1324–1329.

[9] G. Belgioioso and S. Grammatico, *Semi-decentralized Nash equilibrium seeking in aggregative games with separable coupling constraints and non-differentiable cost functions*, IEEE control systems letters **1** (2017), no. 2, 400–405.

[10] _____, *Semi-decentralized generalized Nash equilibrium seeking in monotone aggregative games*, IEEE Transactions on Automatic Control (2021).

[11] G. Belgioioso, A. Nedić, and S. Grammatico, *Distributed generalized Nash equilibrium seeking in aggregative games on time-varying networks*, IEEE Transactions on Automatic Control **66** (2020), no. 5, 2061–2075.

[12] G. Belgioioso, P. Yi, S. Grammatico, and L. Pavel, *Distributed generalized Nash equilibrium seeking: An operator-theoretic perspective*, IEEE Control Systems Magazine **42** (2022), no. 4, 87–102.

[13] D. P Bertsekas, *Constrained optimization and lagrange multiplier methods*, Academic press, 2014.

[14] D. P Bertsekas and A. Scientific, *Convex optimization algorithms*, Athena Scientific Belmont, 2015.

[15] O. Bilenne, P. Mertikopoulos, and E. V. Belmega, *Fast optimization with zeroth-order feedback in distributed, multi-user mimo systems*, IEEE Transactions on Signal Processing **68** (2020), 6085–6100.

[16] M. Bin, I. Notarnicola, L. Marconi, and G. Notarstefano, *A system theoretical perspective to gradient-tracking algorithms for distributed quadratic optimization*, IEEE Conference on Decision and Control (CDC), 2019, pp. 2994–2999.

[17] A. Bloch, *Hamiltonian and gradient flows, algorithms and control*, Vol. 3, American Mathematical Soc., 1994.

[18] A. Boulemtafes, A. Derhab, and Y. Challal, *A review of privacy-preserving techniques for deep learning*, Neurocomputing **384** (2020), 21–45.

[19] R Bouyekhf and A El Moudni, *On analysis of discrete singularly perturbed non-linear systems: application to the study of stability properties*, Journal of the Franklin Institute **334** (1997), no. 2, 199–212.

[20] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, Vol. 15, Siam, 1994.

[21] F. Bullo, *Lectures on network systems*, Vol. 1, Kindle Direct Publishing, 2020.

[22] M. Bürger, G. Notarstefano, and F. Allgöwer, *A polyhedral approximation framework for convex and robust distributed optimization*, IEEE Transactions on Automatic Control **59** (2014), no. 2, 384–395.

[23] A. Camisa, F. Farina, I. Notarnicola, and G. Notarstefano, *Distributed constraint-coupled optimization via primal decomposition over random time-varying graphs*, Automatica **131** (2021), 109739.

[24] A. Camisa, I. Notarnicola, and G. Notarstefano, *Distributed stochastic dual subgradient for constraint-coupled optimization*, IEEE Control Systems Letters **6** (2021), 644–649.

[25] G. Carnevale, M. Bin, I. Notarnicola, L. Marconi, and G. Notarstefano, *Enhanced gradient tracking algorithms for distributed quadratic optimization via sparse gain design*, IFAC-PapersOnLine **53** (2020), no. 2, 2696–2701.

[26] G. Carnevale, A. Camisa, and G. Notarstefano, *Distributed online aggregative optimization for dynamic multi-robot coordination*, IEEE Transactions on Automatic Control (2022).

[27] G. Carnevale, F. Fabiani, F. Fele, K. Margellos, and G. Notarstefano, *Tracking-based distributed equilibrium seeking for aggregative games*, arXiv preprint arXiv:2210.14547 (2022).

[28] G. Carnevale, F. Farina, I. Notarnicola, and G. Notarstefano, *Gtadam: Gradient tracking with adaptive momentum for distributed online optimization*, IEEE Transactions on Control of Network Systems (2022).

[29] G. Carnevale, N. Mimmo, and G. Notarstefano, *Aggregative feedback optimization for distributed cooperative robotics*, IFAC-PapersOnLine **55** (2022), no. 13, 7–12.

[30] ———, *Nonconvex distributed feedback optimization for aggregative cooperative robotics*, arXiv preprint arXiv:2302.01892 (2023).

[31] G. Carnevale, I. Notarnicola, L. Marconi, and G. Notarstefano, *Triggered gradient tracking for asynchronous distributed optimization*, Automatica **147** (2023), 110726.

[32] G. Carnevale and G. Notarstefano, *A learning-based distributed algorithm for personalized aggregative optimization*, 2022 IEEE 61st Conference on Decision and Control (CDC), 2022, pp. 1576–1581.

[33] ———, *Nonconvex distributed optimization via LaSalle and singular perturbations*, IEEE Control Systems Letters **7** (2022), 301–306.

[34] R. L. Cavalcante and S. Stanczak, *A distributed subgradient method for dynamic convex optimization problems under noisy information exchange*, IEEE Journal of Selected Topics in Signal Processing **7** (2013), no. 2, 243–256.

[35] C. Cenedese, G. Belgioioso, Y. Kawano, S. Grammatico, and M. Cao, *Asynchronous and time-varying proximal type dynamics in multiagent network games*, IEEE Transactions on Automatic Control **66** (2020), no. 6, 2861–2867.

[36] C. Cenedese, F. Fabiani, M. Cucuzzella, J. M. Scherpen, M. Cao, and S. Grammatico, *Charging plug-in electric vehicles as a mixed-integer aggregative game*, 2019 IEEE 58th Conference on Decision and Control (CDC), 2019, pp. 4904–4909.

[37] C.-Y. Chang, M. Colombino, J. Cortés, and E. Dall'Anese, *Saddle-flow dynamics for distributed feedback-based optimization*, IEEE Control Systems Letters **3** (2019), no. 4, 948–953.

[38] T.-H. Chang, A. Nedić, and A. Scaglione, *Distributed constrained optimization by consensus-based primal-dual perturbation method*, IEEE Transactions on Automatic Control **59** (2014), no. 6, 1524–1538.

[39] P. Chatupromwong and A. Yokoyama, *Optimization of charging sequence of plug-in electric vehicles in smart grid considering user's satisfaction*, 2012 IEEE International Conference on Power System Technology (POWERCON), 2012, pp. 1–6.

[40] V. Chellaboina and W. M Haddad, *Nonlinear dynamical systems and control: A Lyapunov-based approach*, Princeton University Press, 2008.

[41] X. Chen, C. Gao, M. Zhang, and Y. Qin, *Randomized gradient-free distributed algorithms through sequential gaussian smoothing*, 2017 36th Chinese Control Conference (CCC), 2017, pp. 8407–8412.

[42] Z. Chen and S. Liang, *Distributed aggregative optimization with quantized communication*, Kybernetika **58** (2022), no. 1, 123–144.

[43] A. R Conn, K. Scheinberg, and L. N Vicente, *Introduction to derivative-free optimization*, SIAM, 2009.

[44] J. Cortés and M. Egerstedt, *Coordinated control of multi-robot systems: A survey*, SICE Journal of Control, Measurement, and System Integration **10** (2017), no. 6, 495–503.

[45] L. Cothren, G. Bianchin, and E. Dall'Anese, *Data-enabled gradient flow as feedback controller: Regulation of linear dynamical systems to minimizers of unknown functions*, Learning for Dynamics and Control Conference, 2022, pp. 234–247.

[46] ———, *Online optimization of dynamical systems with deep learning perception*, IEEE Open Journal of Control Systems **1** (2022), 306–321.

[47] E. Dall'Anese and A. Simonetto, *Optimal power flow pursuit*, IEEE Transactions on Smart Grid **9** (2016), no. 2, 942–952.

[48] E. Dall'Anese, A. Simonetto, S. Becker, and L. Madden, *Optimization and learning with information streams: Time-varying algorithms and applications*, IEEE Signal Processing Magazine **37** (2020), no. 3, 71–83.

[49] E. Dall'Anese, H. Zhu, and G. B Giannakis, *Distributed optimal power flow for smart microgrids*, IEEE Transactions on Smart Grid **4** (2013), no. 3, 1464–1475.

[50] A. Daneshmand, G. Scutari, and V. Kungurtsev, *Second-order guarantees of distributed gradient algorithms*, SIAM Journal on Optimization **30** (2020), no. 4, 3029–3068.

[51] C. De Persis and S. Grammatico, *Continuous-time integral dynamics for a class of aggregative games with coupling constraints*, IEEE Transactions on Automatic Control **65** (2019), no. 5, 2171–2176.

[52] Z. Deng, X. Wang, and Y. Hong, *Distributed optimisation design with triggers for disturbed continuous-time multi-agent systems*, IET Control Theory & Applications **11** (2017), no. 2, 282–290.

[53] L. Deori, K. Margellos, and M. Prandini, *Price of anarchy in electric vehicle charging control games: When Nash equilibria achieve social welfare*, Automatica **96** (2018), 150–158.

[54] P. Di Lorenzo and G. Scutari, *NEXT: In-network nonconvex optimization*, IEEE Transactions on Signal and Information Processing over Networks **2** (2016), no. 2, 120–136.

[55] J. Diakonikolas and L. Orecchia, *The approximate duality gap technique: A unified theory of first-order methods*, SIAM Journal on Optimization **29** (2019), no. 1, 660–689.

[56] J. Ding, D. Yuan, G. Jiang, and Y. Zhou, *Distributed quantized gradient-free algorithm for multi-agent convex optimization*, 29th Chinese Control and Decision conference (CCDC), 2017, pp. 6431–6435.

[57] S. Dougherty and M. Guay, *An extremum-seeking controller for distributed optimization over sensor networks*, IEEE Transactions on Automatic Control **62** (2017), no. 2, 928–933.

[58] F. Fabiani, D. Fenucci, and A. Caiti, *A distributed passivity approach to auv teams control in cooperating potential games*, Ocean Engineering **157** (2018), 152–163.

[59] F. Facchinei and C. Kanzow, *Generalized Nash equilibrium problems*, Annals of Operations Research **175** (2010), no. 1, 177–211.

[60] A. Falsone, K. Margellos, S. Garatti, and M. Prandini, *Dual decomposition for multi-agent distributed optimization with coupling constraints*, Automatica **84** (2017), 149–158.

[61] A. Falsone, I. Notarnicola, G. Notarstefano, and M. Prandini, *Tracking-ADMM for distributed constraint-coupled optimization*, Automatica **117** (2020), 108962.

[62] M. Fardad and M. R Jovanović, *On the design of optimal structured and sparse feedback gains via sequential convex programming*, IEEE American Control Conference (ACC), 2014, pp. 2426–2431.

[63] F. Farina, A. Camisa, A. Testa, I. Notarnicola, and G. Notarstefano, *Disropt: a python framework for distributed optimization*, IFAC-PapersOnLine **53** (2020), no. 2, 2666–2671.

[64] F. Farina and G. Notarstefano, *Randomized block proximal methods for distributed stochastic big-data optimization*, IEEE Transactions on Automatic Control **66** (2021), no. 9, 4000–4014.

[65] M. Fazlyab, S. Paternain, V. M Preciado, and A. Ribeiro, *Prediction-correction interior-point method for time-varying convex optimization*, IEEE Trans. on Automatic Control **63** (2017), no. 7, 1973–1986.

[66] F. Fele and K. Margellos, *Probably approximately correct Nash equilibrium learning*, IEEE Transactions on Automatic Control **66** (2020), no. 9, 4238–4245.

[67] F. Ferrante, F. Dabbene, and C. Ravazzi, *On the design of structured stabilizers for LTI systems*, IEEE Control Systems Letters **4** (2019), no. 2, 289–294.

[68] G. Folland, *Higher-order derivatives and taylor's formula in several variables*, Preprint (2005), 1–4.

[69] D. Gadjov and L. Pavel, *Single-timescale distributed gne seeking for aggregative games over networks via forward–backward operator splitting*, IEEE Transactions on Automatic Control **66** (2020), no. 7, 3259–3266.

[70] T. Ge, W. Lin, and J. Feng, *Invariance principles allowing of non-Lyapunov functions for estimating attractor of discrete dynamical systems*, IEEE Transactions on Automatic Control **57** (2011), no. 2, 500–505.

[71] A. Gharaibeh, M. A Salahuddin, S. J. Hussini, A. Khreishah, I. Khalil, M. Guizani, and A. Al-Fuqaha, *Smart cities: A survey on data management, security, and enabling technologies*, IEEE Communications Surveys & Tutorials **19** (2017), no. 4, 2456–2501.

[72] B. Gharesifard and J. Cortés, *Distributed continuous-time convex optimization on weight-balanced digraphs*, IEEE Transactions on Automatic Control **59** (2013), no. 3, 781–786.

[73] P. Giselsson and A. Rantzer, *Large-scale and distributed optimization*, Vol. 2227, Springer, 2018.

[74] S. Grammatico, *Dynamic control of agents playing aggregative games with coupling constraints*, IEEE Transactions on Automatic Control **62** (2017), no. 9, 4537–4548.

[75] S. Grammatico, F. Parise, M. Colombino, and J. Lygeros, *Decentralized convergence to Nash equilibria in constrained deterministic mean field control*, IEEE Transactions on Automatic Control **61** (2015), no. 11, 3315–3329.

[76] M. Guay, *Distributed newton seeking*, Computers & Chemical Engineering **146** (2021), 107206.

[77] M. Guay, I. Vandermeulen, S. Dougherty, and P J. McLellan, *Distributed extremum-seeking control over networks of dynamically coupled unstable dynamic agents*, Automatica **93** (2018), 498–509.

[78] V. Häberle, A. Hauswirth, L. Ortmann, S. Bolognani, and F. Dörfler, *Non-convex feedback optimization with input and output constraints*, IEEE Control Systems Letters **5** (2020), no. 1, 343–348.

[79] T. Hatanaka, N. Chopra, T. Ishizaki, and N. Li, *Passivity-based distributed optimization with communication delays using pi consensus algorithm*, IEEE Transactions on Automatic Control **63** (2018), no. 12, 4421–4428.

[80] A. Hauswirth, S. Bolognani, G. Hug, and F. Dörfler, *Optimization algorithms as robust feedback controllers*, arXiv preprint arXiv:2103.11329 (2021).

[81] A. Hauswirth, F. Dörfler, and A. Teel, *Anti-windup approximations of oblique projected dynamics for feedback-based optimization*, arXiv preprint arXiv:2003.00478 (2020).

[82] F. Hayashi, *Econometrics*, Princeton University Press, 2011.

[83] Z. He, S. Bolognani, J. He, F. Dörfler, and X. Guan, *Model-free nonlinear feedback optimization*, arXiv preprint arXiv:2201.02395 (2022).

[84] J. M Holte, *Discrete gronwall lemma and applications*, MAA-NCS meeting at the University of North Dakota, 2009, pp. 1–7.

[85] R. A Horn and C. R Johnson, *Matrix analysis*, Cambridge university press, 2012.

[86] S. Hosseini, A. Chapman, and M. Mesbahi, *Online distributed convex optimization on dynamic networks*, IEEE Transactions on Automatic Control **61** (2016), no. 11, 3545–3550.

[87] M. K. Jensen, *Aggregative games and best-reply potentials*, Economic theory **43** (2010), no. 1, 45–66.

[88] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, *Subgradient methods and consensus algorithms for solving convex optimization problems*, 2008 47th ieee conference on decision and control, 2008, pp. 4185–4190.

[89] Y. Kajiyama, N. Hayashi, and S. Takai, *Distributed subgradient method with edge-based event-triggered communication*, IEEE Transactions on Automatic Control **63** (2018), no. 7, 2248–2255.

[90] H. Kebriaei, S J. Sadati-Savadkoohi, M. Shokri, and S. Grammatico, *Multipopulation aggregative games: Equilibrium seeking via mean-field control and consensus*, IEEE Transactions on Automatic Control **66** (2021), no. 12, 6011–6016.

[91] H. K Khalil, *Nonlinear systems*, Upper Saddle River (2002).

[92] S. S Kia, J. Cortés, and S. Martínez, *Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication*, Automatica **55** (2015), 254–264.

[93] S. S Kia, B. Van Scoy, J. Cortes, R. A Freeman, K. M Lynch, and S. Martinez, *Tutorial on dynamic average consensus: The problem, its applications, and the algorithms*, IEEE Control Systems Magazine **39** (2019), no. 3, 40–72.

[94] D. P Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[95] J. Koshal, A. Nedić, and U. V Shanbhag, *Distributed algorithms for aggregative games on graphs*, Operations Research **64** (2016), no. 3, 680–704.

[96] D. Krishnamoorthy and S. Skogestad, *Real-time optimization as a feedback control problem-a review*, Computers & Chemical Engineering (2022), 107723.

[97] M. Krstić and H.-H. Wang, *Stability of extremum seeking feedback for general nonlinear dynamic systems*, Automatica **36** (2000), no. 4, 595–601.

[98] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, *Human-aware robot navigation: A survey*, Robotics and Autonomous Systems **61** (2013), no. 12, 1726–1743.

[99] K. Kvaternik and L. Pavel, *An analytic framework for decentralized extremum seeking control*, 2012 American Control Conference (ACC), 2012, pp. 3371–3376.

[100] A. Lamperski and L. Lessard, *Optimal decentralized state-feedback control with sparsity and delays*, Automatica **58** (2015), 143–151.

[101] E. A Lee, *The past, present and future of cyber-physical systems: A focus on models*, Sensors **15** (2015), no. 3, 4837–4869.

[102] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM Journal on Optimization **26** (2016), no. 1, 57–95.

[103] M. Li, G. Chesi, and Y. Hong, *Input-feedforward-passivity-based distributed optimization over jointly connected balanced digraphs*, IEEE Transactions on Automatic Control **66** (2021), no. 9, 4117–4131.

[104] X. Li, L. Xie, and Y. Hong, *Distributed aggregative optimization over multi-agent networks*, IEEE Transactions on Automatic Control **67** (2021), no. 6, 3165–3171.

[105] X. Li, L. Xie, and N. Li, *A survey of decentralized online learning*, arXiv preprint arXiv:2205.00473 (2022).

[106] X. Li, X. Yi, and L. Xie, *Distributed online convex optimization with an aggregative variable*, IEEE Transactions on Control of Network Systems **9** (2021), no. 1, 438–449.

[107] Y. Li, G. Qu, and N. Li, *Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit*, IEEE Transactions on Automatic Control **66** (2020), no. 10, 4761–4768.

[108] Z. Li, Z. Ding, J. Sun, and Z. Li, *Distributed adaptive convex optimization on directed graphs via continuous-time algorithms*, IEEE Transactions on Automatic Control **63** (2018), no. 5, 1434–1441.

[109] Z. Li, K. You, and S. Song, *Cooperative source seeking via networked multi-vehicle systems*, Automatica **115** (2020), 108853.

[110] S. Liang, L. Y. Wang, and G. Yin, *Distributed smooth convex optimization with coupled constraints*, IEEE Transactions on Automatic Control **65** (2020), no. 1, 347–353.

[111] F. Lin, M. Fardad, and M. R Jovanović, *Augmented Lagrangian approach to design of structured optimal state feedback gains*, IEEE Transactions on Automatic Control **56** (2011), no. 12, 2923–2929.

[112] ———, *Design of optimal sparse feedback gains via the alternating direction method of multipliers*, IEEE Transactions on Automatic Control **58** (2013), no. 9, 2426–2431.

[113] P. Lin, W. Ren, and J. A Farrell, *Distributed continuous-time optimization: nonuniform gradient gains, finite-time convergence, and convex constraint set*, IEEE Transactions on Automatic Control **62** (2017), no. 5, 2239–2253.

[114] J. Liu and W. Chen, *Sample-based zero-gradient-sum distributed consensus optimization of multi-agent systems*, Proc. of the 11th world congress on intelligent control and automation, 2014, pp. 215–219.

[115] Q. Liu and J. Wang, *A second-order multi-agent network for bound-constrained distributed optimization*, IEEE Transactions on Automatic Control **60** (2015), no. 12, 3310–3315.

[116] S. Liu, L. Xie, and D. E Quevedo, *Event-triggered quantized communication-based distributed convex optimization*, IEEE Transactions on Control of Network Systems **5** (2016), no. 1, 167–178.

[117] L. Ljung et al., *Theory for the user*, System Identification (1987).

[118] S. H Low, F. Paganini, and J. C Doyle, *Internet congestion control*, IEEE control systems magazine **22** (2002), no. 1, 28–43.

[119] J. Lu and C. Y. Tang, *Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case*, IEEE Transactions on Automatic Control **57** (2012), no. 9, 2348–2354.

[120] X. Luo, Y. Zhang, and M. M Zavlanos, *Socially-aware robot planning via bandit human feedback*, 2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS), 2020, pp. 216–225.

[121] P. K. R. Maddikunta, Q.-V. Pham, B Prabadevi, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, and M. Liyanage, *Industry 5.0: A survey on enabling technologies and potential applications*, Journal of Industrial Information Integration **26** (2022), 100257.

[122] D. Mateos-Núnez and J. Cortés, *Distributed online convex optimization over jointly connected digraphs*, IEEE Transactions on Network Science and Engineering **1** (2014), no. 1, 23–37.

[123] _____, *Distributed saddle-point subgradient algorithms with Laplacian averaging*, IEEE Transactions on Automatic Control **62** (2017), no. 6, 2720–2735.

[124] W. Mei and F. Bullo, *Lasalle invariance principle for discrete-time dynamical systems: A concise and self-contained tutorial*, arXiv preprint arXiv:1710.03710 (2017).

[125] I. Menache and A. Ozdaglar, *Network games: Theory, models, and dynamics*, Synthesis Lectures on Communication Networks **4** (2011), no. 1, 1–159.

[126] M. Menner, L. Neuner, L. Lünenburger, and M. N Zeilinger, *Using human ratings for feedback control: A supervised learning approach with application to rehabilitation robotics*, IEEE Transactions on Robotics **36** (2020), no. 3, 789–801.

[127] A. Menon and J. S. Baras, *Collaborative extremum seeking for welfare optimization*, 53rd IEEE Conference on Decision and Control, 2014, pp. 346–351.

[128] S. Menta, A. Hauswirth, S. Bolognani, G. Hug, and F. Dörfler, *Stability of dynamic feedback optimization with applications to power systems*, 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2018, pp. 136–143.

[129] S. Michalowsky, B. Gharesifard, and C. Ebenbauer, *Distributed extremum seeking over directed graphs*, 2017 IEEE 56th Annual Conference on Decision and Control (CDC), 2017, pp. 2095–2101.

[130] N. Mimmo, G. Carnevale, A. Testa, and G. Notarstefano, *Extremum seeking tracking for derivative-free distributed optimization*, arXiv preprint arXiv:2110.04234 (2021).

[131] A.-H. Mohsenian-Rad, V. W. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, *Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid*, IEEE Transactions on Smart Grid **1** (2010), no. 3, 320–331.

[132] A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro, *Online optimization in dynamic environments: Improved regret rates for strongly convex problems*, IEEE Conference on Decision and Control (CDC), 2016, pp. 7195–7201.

[133] H. Moradian and S. S Kia, *A distributed continuous-time modified newton–raphson algorithm*, Automatica **136** (2022), 109886.

[134] P. Nazari, D. A. Tarzanagh, and G. Michailidis, *Dadam: A consensus-based distributed adaptive gradient method for online optimization*, arXiv preprint arXiv:1901.09109 (2019).

[135] A. Nedić and J. Liu, *Distributed optimization for control*, Annual Review of Control, Robotics, and Autonomous Systems **1** (2018), 77–103.

[136] A. Nedić, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization **27** (2017), no. 4, 2597–2633.

[137] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Transactions on Automatic Control **54** (2009), no. 1, 48–61.

[138] A. Nedić, A. Ozdaglar, and P. A Parrilo, *Constrained consensus and optimization in multi-agent networks*, IEEE Transactions on Automatic Control **55** (2010), no. 4, 922–938.

[139] I. Notarnicola and G. Notarstefano, *Constraint-coupled distributed optimization: a relaxation and duality approach*, IEEE Transactions on Control of Network Systems **7** (2019), no. 1, 483–492.

[140] I. Notarnicola, A. Simonetto, F. Farina, and G. Notarstefano, *Distributed personalized gradient tracking with convex parametric models*, IEEE Transactions on Automatic Control **68** (2022), no. 1, 588–595.

[141] ———, *Distributed personalized gradient tracking with convex parametric models*, IEEE Transactions on Automatic Control (2022).

[142] G. Notarstefano, I. Notarnicola, and A. Camisa, *Distributed optimization for smart cyber-physical networks*, Foundations and Trends® in Systems and Control **7** (2019), no. 3, 253–383.

[143] K. Okuguchi and F. Szidarovszky, *The theory of oligopoly with multi-product firms*, Springer Science & Business Media, 2012.

[144] A. M Ospina, N. Bastianello, and E. Dall'Anese, *Data-based online optimization of networked systems with infrequent feedback*, arXiv preprint arXiv:2109.06343 (2021).

[145] A. M Ospina, A. Simonetto, and E. Dall'Anese, *Personalized demand response via shape-constrained online learning*, 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 2020, pp. 1–6.

[146] ———, *Time-varying optimization of networked systems with human preferences*, IEEE Transactions on Control of Network Systems (2022).

[147] D. Paccagnan, B. Gentile, F. Parise, M. Kamgarpour, and J. Lygeros, *Nash and Wardrop equilibria in aggregative games with coupling constraints*, IEEE Transactions on Automatic Control **64** (2018), no. 4, 1373–1388.

[148] Y. Pang and G. Hu, *Exact convergence of gradient-free distributed optimization method in a multi-agent system*, 2018 IEEE Conference on Decision and Control (CDC), 2018, pp. 5728–5733.

[149] Y. Pang and G. Hu, *Randomized gradient-free distributed optimization methods for a multiagent system with unknown cost function*, IEEE Transactions on Automatic Control **65** (2019), no. 1, 333–340.

[150] F. Parise, B. Gentile, and J. Lygeros, *A distributed algorithm for almost-Nash equilibria of average aggregative games with coupling constraints*, IEEE Transactions on Control of Network Systems **7** (2019), no. 2, 770–782.

[151] F. Parise, S. Grammatico, B. Gentile, and J. Lygeros, *Distributed convergence to Nash equilibria in network and average aggregative games*, Automatica **117** (2020), 108959.

[152] F. Parise and A. Ozdaglar, *Analysis and interventions in large network games*, Annual Review of Control, Robotics, and Autonomous Systems **4** (2021), 455–486.

[153] J. Popenda, *On the discrete analogy of gronwall lemma*, Demonstratio Mathematica **16** (1983), no. 1, 11–26.

[154] J. Poveda and N. Quijano, *Distributed extremum seeking for real-time resource allocation*, American Control Conf., 2013, pp. 2772–2777.

[155] J. Poveda, M Benosman, and A. Teel, *Distributed extremum seeking in multi-agent systems with arbitrary switching graphs*, IFAC-PapersOnLine **50** (2017), no. 1, 735–740.

[156] S. Pu and A. Nedić, *Distributed stochastic gradient tracking methods*, Mathematical Programming (2020), 1–49.

[157] G. Punzo, A. Tewari, E. Butans, M. Vasile, A. Purvis, M. Mayfield, and L. Varga, *Engineering resilient complex systems: the necessary shift toward complexity science*, IEEE Systems Journal **14** (2020), no. 3, 3865–3874.

[158] G. Qu and N. Li, *Harnessing Smoothness to Accelerate Distributed Optimization*, IEEE Transactions on Control of Network Systems **5** (2018), no. 3, 1245–1260.

[159] ———, *On the exponential stability of primal-dual gradient dynamics*, IEEE Control Systems Letters **3** (2018), no. 1, 43–48.

[160] M. Rabbat and R. Nowak, *Distributed optimization in sensor networks*, International symposium on information processing in sensor networks, 2004, pp. 20–27.

[161] S S. Ram, A. Nedić, and V. V Veeravalli, *Distributed stochastic subgradient projection algorithms for convex optimization*, Journal of optimization theory and applications **147** (2010), no. 3, 516–545.

[162] A. K. Sahu and S. Kar, *Decentralized zeroth-order constrained stochastic optimization algorithms: Frank–wolfe and variants with applications to black-box adversarial attacks*, Proceedings of the IEEE **108** (2020), no. 11, 1890–1905.

[163] A. K. Sahu, S. Kar, J. M. Moura, and H V. Poor, *Distributed constrained recursive nonlinear least-squares estimation: Algorithms and asymptotics*, IEEE Transactions on Signal and Information Processing over Networks **2** (2016), no. 4, 426–441.

[164] Y. B. Salamah, L. Fiorentini, and U. Ozguner, *Cooperative extremum seeking control via sliding mode for distributed optimization*, 2018 IEEE Conference on Decision and Control (CDC), 2018, pp. 1281–1286.

[165] J. A Sanders, F. Verhulst, and J. Murdock, *Averaging methods in nonlinear dynamical systems*, Vol. 59, Springer, 2007.

[166] S. Sastry, *Nonlinear systems: analysis, stability, and control*, Vol. 10, Springer Science & Business Media, 2013.

[167] G. Scutari, F. Facchinei, J.-S. Pang, and D. P Palomar, *Real and complex monotone communication games*, IEEE Transactions on Information Theory **60** (2014), no. 7, 4197–4231.

[168] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Mathematical Programming **176** (2019), no. 1-2, 497–544.

[169] S. Shahrampour and A. Jadbabaie, *Distributed online optimization in dynamic environments using mirror descent*, IEEE Transactions on Automatic Control **63** (2017), no. 3, 714–725.

[170] B. Shi, S. S Du, M. I Jordan, and W. J Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Mathematical Programming (2021), 1–70.

[171] A. Simonetto, E. Dall'Anese, J. Monteil, and A. Bernstein, *Personalized optimization with user's feedback*, Automatica **131** (2021), 109767.

[172] ———, *Personalized optimization with user's feedback*, Automatica **131** (2021), 109767.

[173] A. Simonetto and H. Jamali-Rad, *Primal recovery from consensus-based dual decomposition for distributed convex optimization*, Journal of Optimization Theory and Applications **168** (2016), no. 1, 172–197.

[174] A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro, *A class of prediction-correction methods for time-varying convex optimization*, IEEE Trans. on Signal Processing **64** (2016), no. 17, 4576–4591.

[175] E. D Sontag, *Input to state stability: Basic concepts and results*, Nonlinear and optimal control theory, 2008, pp. 163–220.

[176] W. Su, S. Boyd, and E. Candes, *A differential equation for modeling nesterov's accelerated gradient method: Theory and insights*, Advances in neural information processing systems, 2014, pp. 2510–2518.

[177] A. Sundararajan, B. Van Scoy, and L. Lessard, *A canonical form for first-order distributed optimization algorithms*, 2019 American Control Conference (ACC), 2019, pp. 4075–4080.

[178] Y. Tan, D. Nešić, and I. Mareels, *On non-local stability properties of extremum seeking control*, Automatica **42** (2006), no. 6, 889–903.

[179] Y. Tang, J. Zhang, and N. Li, *Distributed zero-order algorithms for nonconvex multi-agent optimization*, IEEE Transactions on Control of Network Systems (2020), 1–1.

[180] Y. Tang, K. Dvijotham, and S. Low, *Real-time optimal power flow*, IEEE Transactions on Smart Grid **8** (2017), no. 6, 2963–2973.

[181] T. Tatarenko and B. Touri, *Non-convex distributed optimization*, IEEE Transactions on Automatic Control **62** (2017), no. 8, 3744–3757.

[182] A. R Teel and D. Popovic, *Solving smooth and nonsmooth multivariable extremum seeking problems by the methods of nonlinear programming*, Proceedings of the 2001 american control conference.(cat. no. 01ch37148), 2001, pp. 2394–2399.

[183] A. Terpin, S. Fricker, M. Perez, M. H. de Badyn, and F. Dörfler, *Distributed feedback optimisation for robotic coordination*, 2022 American Control Conference (ACC), 2022, pp. 3710–3715.

[184] Z. J Towfic and A. H Sayed, *Adaptive penalty-based distributed stochastic convex optimization*, IEEE Transactions on Signal Processing **62** (2014), no. 15, 3924–3938.

[185] I. Vandermeulen, M. Guay, and P. J. McLellan, *Discrete-time distributed extremum-seeking control over networks with unstable dynamics*, IEEE Transactions on Control of Network Systems **5** (2018), no. 3, 1182–1192.

[186] D. Varagnolo, F. Zanella, A. Cenedese, G. Pillonetto, and L. Schenato, *Newton-Raphson consensus for distributed convex optimization*, IEEE Transactions on Automatic Control **61** (2016), no. 4, 994–1009.

[187] C. Wang and X. Xie, *Design and analysis of distributed multi-agent saddle point algorithm based on gradient-free oracle*, 2018 australian new zealand control conference (anzcc), 2018, pp. 362–365.

[188] D. Wang, M. Chen, and W. Wang, *Distributed extremum seeking for optimal resource allocation and its application to economic dispatch in smart grids*, IEEE Transactions on Neural Networks and Learning Systems **30** (2019), no. 10, 3161–3171.

[189] J. Wang and N. Elia, *Control approach to distributed optimization*, 2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2010, pp. 557–561.

[190] ———, *A control perspective for centralized and distributed convex optimization*, 2011 50th ieee conference on decision and control and european control conference, 2011, pp. 3800–3805.

[191] L. Wang, Y. Wang, and Y. Hong, *Distributed online optimization with gradient-free design*, 2019 Chinese Control Conference (CCC), 2019, pp. 5677–5682.

[192] T. Wang and P. Yi, *Distributed projection-free algorithm for constrained aggregative optimization*, arXiv preprint arXiv:2207.11885 (2022).

[193] A. Wibisono, A. C Wilson, and M. I Jordan, *A variational perspective on accelerated methods in optimization*, Proceedings of the National Academy of Sciences **113** (2016), no. 47, E7351–E7358.

[194] A. C Wilson, B. Recht, and M. I Jordan, *A Lyapunov analysis of accelerated methods in optimization*, Journal of Machine Learning Research **22** (2021), no. 113, 1–34.

[195] B. Wittenmark and A. Urquhart, *Adaptive extremal control*, Proceedings of 1995 34th ieee conference on decision and control, 1995, pp. 1639–1644.

[196] C. Xi, R. Xin, and U. A Khan, *ADD-OPT: Accelerated distributed directed optimization*, IEEE Transactions on Automatic Control **63** (2017), no. 5, 1329–1339.

[197] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms*, arXiv preprint arXiv:1708.07747 (2017).

[198] J. Xie, F R. Yu, T. Huang, R. Xie, J. Liu, and Y. Liu, *A survey on the scalability of blockchain systems*, IEEE Network **33** (2019), no. 5, 166–173.

[199] R. Xin and U. A Khan, *A linear algorithm for optimization over directed graphs with geometric convergence*, IEEE Control Systems Letters **2** (2018), no. 3, 315–320.

[200] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, *Convergence of asynchronous distributed gradient methods over stochastic networks*, IEEE Transactions on Automatic Control **63** (2017), no. 2, 434–448.

[201] S. Yang, Q. Liu, and J. Wang, *A multi-agent system with a proportional-integral protocol for distributed constrained optimization*, IEEE Transactions on Automatic Control **62** (2017), no. 7, 3461–3467.

[202] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H Johansson, *A survey of distributed optimization*, Annual Reviews in Control **47** (2019), 278–305.

[203] M. Ye and G. Hu, *Distributed extremum seeking for constrained networked optimization and its application to energy consumption control in smart grid*, IEEE Transactions on Control Systems Technology **24** (2016), no. 6, 2048–2058.

[204] M. Ye, G. Hu, L. Xie, and S. Xu, *Differentially private distributed Nash equilibrium seeking for aggregative games*, IEEE Transactions on Automatic Control **67** (2021), no. 5, 2451–2458.

[205] P. Yi and L. Pavel, *An operator splitting approach for distributed generalized Nash equilibria computation*, Automatica **102** (2019), 111–121.

[206] X. Yi, X. Li, L. Xie, and K. H Johansson, *Distributed online convex optimization with time-varying coupled inequality constraints*, IEEE Transactions on Signal Processing **68** (2020), 731–746.

[207] X. Yi, L. Yao, T. Yang, J. George, and K. H Johansson, *Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication*, IEEE Conference on Decision and Control (CDC), 2018, pp. 3397–3402.

[208] H. Yin, P. G Mehta, S. P Meyn, and U. V Shanbhag, *Synchronization of coupled oscillators is a game*, IEEE Transactions on Automatic Control **57** (2011), no. 4, 920–935.

[209] D. Yuan and D. W. C. Ho, *Randomized gradient-free method for multiagent optimization over time-varying networks*, IEEE Transactions on Neural Networks and Learning Systems **26** (2015), no. 6, 1342–1347.

[210] X. Zeng, P. Yi, and Y. Hong, *Distributed continuous-time algorithm for constrained convex optimizations via nonsmooth analysis approach*, IEEE Transactions on Automatic Control **62** (2017), no. 10, 5227–5233.

[211] Y. Zhang, R. J Ravier, M. M Zavlanos, and V. Tarokh, *A distributed online convex optimization algorithm with improved dynamic regret*, IEEE conf. on Decision and Control (CDC), 2019, pp. 2449–2454.

[212] Y. Zhang and M. M Zavlanos, *A consensus-based distributed augmented lagrangian method*, IEEE Conference on Decision and Control (CDC), 2018, pp. 1763–1768.

[213] Z. Zhao, G. Chen, and M. Dai, *Distributed event-triggered scheme for a convex optimization problem in multi-agent systems*, Neurocomputing **284** (2018), 90–98.

[214] Y. Zhou, Y. Zhang, X. Luo, and M. M Zavlanos, *Human-in-the-loop robot planning with non-contextual bandit feedback*, 2021 60th IEEE Conference on Decision and Control (CDC), 2021, pp. 2848–2853.

[215] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, *Big data analytics in intelligent transportation systems: A survey*, IEEE Transactions on Intelligent Transportation Systems **20** (2018), no. 1, 383–398.

[216] M. Zhu and S. Martínez, *Discrete-time dynamic average consensus*, Automatica **46** (2010), no. 2, 322–329.