Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF **DOCTOR OF PHILOSOPHY**

TITLE _____

_____

PRESENTED BY _____

ACCEPTED BY THE DEPARTMENTS OF

_____

_____

_____ _____

ADVISOR, MAJOR PROFESSOR DATE

_____ _____

DEPARTMENT HEAD DATE

_____ _____

DEPARTMENT HEAD DATE

APPROVED BY THE COLLEGE COUNCIL

_____ _____

DEAN DATE

Thesis Committee

Pedro Aguiar
Instituto Superior Técnico
(Advisor)

José Moura
Carnegie Mellon University

Markus Püschel
ETH Zurich & Carnegie Mellon University
(Advisor)

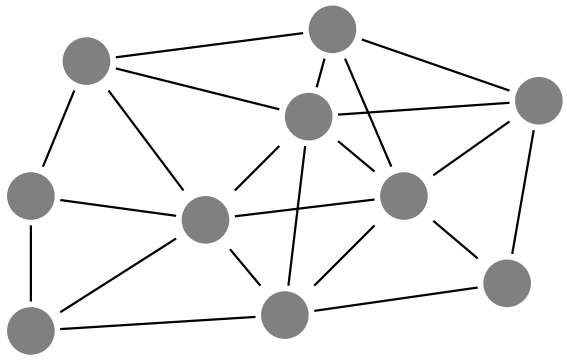
Alejandro Ribeiro
University of Pennsylvania

João Xavier
Instituto Superior Técnico
(Advisor)

Communication-Efficient Algorithms For Distributed Optimization

João F. C. Mota

September 2013



A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in

Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA
Instituto Superior Técnico, Technical University of Lisbon, Portugal

Co-advised by:

Pedro Aguiar, Instituto Superior Técnico, Technical University of Lisbon
Markus Püschel, ETH Zurich & Carnegie Mellon University
João Xavier, Instituto Superior Técnico, Technical University of Lisbon

. . . to my parents, Luís and Helena.

Abstract

This thesis is concerned with the design of distributed algorithms for solving optimization problems. The particular scenario we consider is a network with P compute nodes, where each node p has exclusive access to a cost function f_p. We design algorithms in which all the nodes cooperate to find the minimum of the sum of all the cost functions, $f_1 + \cdots + f_P$. Several problems in signal processing, control, and machine learning can be posed as such optimization problems. Given that communication is often the most energy-consuming operation in networks and, many times, also the slowest one, it is important to design distributed algorithms with low communication requirements, that is, communication-efficient algorithms. The two main contributions of this thesis are a classification scheme for distributed optimization problems of the kind explained above and a set of corresponding communication-efficient algorithms.

The class of optimization problems we consider is quite general, since we allow that each function may depend on arbitrary components of the optimization variable, and not necessarily on all of them. In doing so, we go beyond the commonly used assumption in distributed optimization and create additional structure that can be explored to reduce the total number of communications. This structure is captured by our classification scheme, which identifies particular instances of the problem that are easier to solve. One example is the standard distributed optimization problem, in which all the functions depend on all the components of the variable.

All our algorithms are distributed in the sense that no central node coordinates the network, all the communications occur exclusively between neighboring nodes, and the data associated with each node is always processed locally. We show several applications of our algorithms, including average consensus, support vector machines, network flows, and several distributed scenarios for compressed sensing. We also propose a new framework for distributed model predictive control, which can be solved with our algorithms. Through extensive numerical experiments, we show that our algorithms outperform prior distributed algorithms in terms of communication-efficiency, even some that were specifically designed for a particular application.

Acknowledgments

This thesis is the result of a complicated sequence of events. I would like to take the opportunity to thank here some of the people who, directly or indirectly, influenced, changed, or caused those events.

The direct causers of the main events were, undoubtedly, my advisors: João Xavier, Pedro Aguiar, and Markus Püschel. The three of them gave me the support, the insight, and the knowledge that made this thesis possible. I learned a lot from them, both academically and non-academically. Most importantly, no matter how busy they were, they could always find time to answer my questions, to take care of bureaucracy that involved me, and to meet in our regular meetings. Also, I want to say that I had lots of fun in those yearly (work!) trips to several towns in Portugal. Thank you for all of that!

I would like to thank my thesis committee members, José Moura and Alejandro Ribeiro, for all the insight and suggestions. During my PhD, and especially during the years I spent in CMU, José was always very supportive. On the few occasions that we discussed research, José showed me how to look at my research from a different perspective. I would also like to thank Alejandro for arranging everything when I visited him in Philadelphia.

The person who convinced me to enter the CMU/Portugal PhD program was João Paulo Costeira. He has always been in the background, doing whatever is needed to make this program great, and providing a comfortable layer between all the bureaucracy that lies under such a big program and the students (including myself). He has also put me in contact with people and projects from the real world! Another early causer of the events that led to this thesis was Victor Barroso, who invited me to participate in research meetings at ISR, and subsequently introduced me to 2/3 of my future advisors.

During my PhD, I had the opportunity to collaborate and to discuss research with several people. I would like to thank them for that. Some of these people are Michael Rabbat, João Miranda Lemos, André Martins, Bruno Sinopoli, Gabriela Hug, Mário Figueiredo, Petros Boufounos, Qing Ling, Ricardo Lima, Aurora Schmidt, Ricardo Cabral, Brian Swenson, Stephen Boyd, Dusan Jakovetić, Dragana Bajovic, Soummya Kar, Jerónimo Rodrigues, Sabina Zejnilovic, Pinar Oguz,

Dario Figueira, Paulo Oliveira, June Zhang, Claudia Soares, Qixing Liu, Matthias Althoff, Alysson Bessani, Christian Conte, Paulo Oliveira, Bruce Krogh, Stefan Richter, Marija Ilić, Pedro Guerreiro, Susana Brandão, Nicholas O'Donoughue, Nikos Arechiga, Kyri Baker, Aliaksei Sandryhaila, Marek Telgarsky, Augusto Santos, Bernardo Pires, Ceyhun Eksin, Christian Berger, Divyanshu Vats, Akshay Rajans, Ehsan Zamanizadeh, Jhi-Young Joo, Sanja Cvijic, Luís Brandão, Franz Franchetti, Xiahui Wang (Eeyore), Luca Parolini, Rohan Chabukswar, Joel Harley, Rodrigo Belo, Joya Deri, and Sérgio Pequito. Also, thank you to Daniel McFarlin and Vas Chellappa for helping me out with several issues with Linux and MPI. Some of the experiments shown in this thesis were run on a computer cluster gently provided by Florin Manolache.

Conducting research while jumping back and forth over a large ocean is not possible without proper funding and an excellent team "taking care of things." So I would like to thank the CMU/Portugal program and FCT for this opportunity, and also the staff involved at CMU, IST, and the CMU/Portugal program, especially Ana Mateus, Carolyn Patterson, Susana Santana, Alexandra Araújo, Ana Santos, Filomena Viegas, Lori Spears, Claire Bauerle, Tara Moe, Elaine Lawrence, and Samantha Goldstein.

No single piece of this thesis would be possible without the early support of both of my parents, Luís and Helena, who always encouraged me no matter what direction I chose. Their support has been a constant throughout my entire education and, for this and for other reasons, the minimum I can give back is to dedicate this thesis to them.

My sister, Renata, has also always provided constant encouragement, kept me in a good mood, and was a source of inspiration. My uncle Manuel and aunt Fátima, and my uncle Alexandre and aunt Lili, and cousin Afonso have always encouraged me in my studies and instilled in me an interest in science from an early age.

Finally, I have no words to describe my gratitude to my wife, Kate. Thank you for all your support, kindness, and love. Thank you also for making sure that, during the writing of this thesis, I had proper nutrition, rest, and also fun; thanks for proofreading some parts of the thesis also. Now I promise that I'll do my homework for our piano lessons.

Contents

Chapter 1

Introduction

Optimization theory has contributed to many fields in engineering by providing efficient algorithms that solve nontrivial real world problems. Notable examples can be found in signal processing, control engineering, and machine learning [1, 2, 3]. On the other hand, over the last years, some computation platforms on which these algorithms may be executed have become distributed. For example, computers are now equipped with several processing devices, allowing for parallel computation. Also, complex systems such as power grids or water distribution systems are composed of several interconnected components, each with some processing power, and thus are distributed by nature. In addition, the data to be processed is often generated at different locations as, for example, in sensor networks, or in the internet. All these factors ask for algorithms that process data or control systems in a distributed way. However, it is challenging to design optimization algorithms that are matched to distributed resources. Part of the reason is that efficient centralized algorithms, such as for example interior-point methods, cannot be easily adapted to distributed scenarios. The high-level goal of this thesis is to advance the design of distributed algorithms for solving optimization problems.

1.1 Overview

Figure 1.1 will help us describe the main problem addressed by this thesis. The figure shows a network with 10 nodes, where each node p holds a function f_p. Our goal is to make all nodes cooperate in order to find a minimizer of the sum of all the functions:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f_1(x_{S_1}) + f_2(x_{S_2}) + \cdots + f_P(x_{S_P})\,. \tag{P}$$

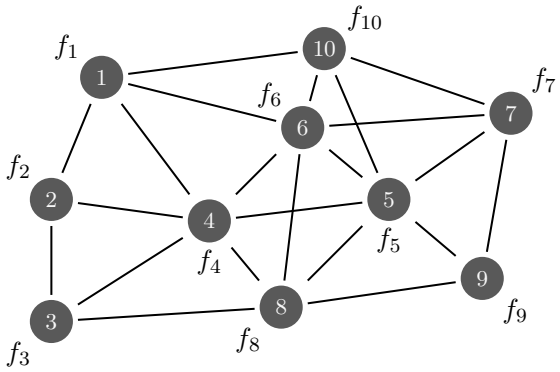
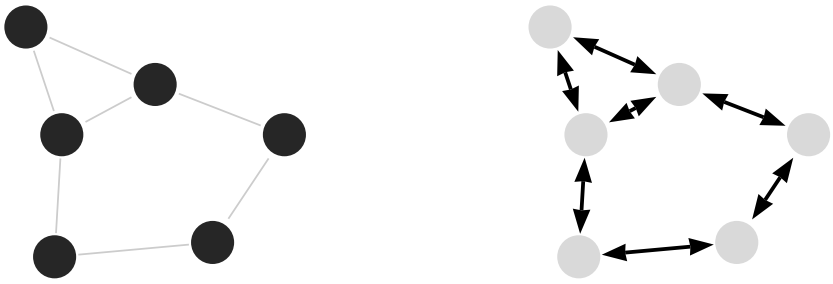


FIGURE 1.1: Illustration of the main problem of the thesis: each node in the network holds a private function and the goal is to minimize the sum of all the functions.

where $x \in \mathbb{R}^n$ is the optimization variable. Each function f_p in (P) depends on the components of the variable x that are indexed by the set $S_p \subseteq \{1, \ldots, n\}$, and we use x_{S_p} to denote those components. For example, if the function at node 3 depends on components x_1, x_5, x_8, and x_{10}, then $S_3 = \{1, 5, 8, 10\}$ and $f_3(x_{S_3}) = f_3(x_1, x_5, x_8, x_{10})$. We require each function f_p to be private to node p, i.e., no other node in the network has access to it. The edges of the network represent communication links. This means, for example, that node 3 in Figure 1.1 can communicate only with its neighbors: nodes 2, 4, and 8. Given such a network, an algorithm that solves (P) is considered *distributed* if it uses no central node, no all-to-all communications, and if the privacy requirement for each function f_p is satisfied. In this thesis, we aim to solve (P), and related problems, with distributed algorithms that are communication-efficient, i.e., that use a minimal amount of communication. Communication-efficiency is an essential requirement, for example, when the nodes are battery-operated devices, such as in sensor-networks, since communication is usually very energy-demanding.

Simple example. Consider an inference problem on a sensor network [4, 5], and suppose that each function f_p depends on all the components of the optimization variable $x \in \mathbb{R}^n$, i.e., $S_p = \{1, \ldots, n\}$, for all p or, more compactly, $\cap_{p=1}^P S_p = \{1, \ldots, n\}$. Each node in the network represents a sensor with computing abilities. The edges indicate direct sensor communication, for instance, through a wireless connection. We want to estimate a parameter $\bar{\theta} \in \mathbb{R}^n$ (e.g., a set of environmental parameters [6]), by using noisy measurements from all nodes. Let θ_p be the measurement of $\bar{\theta}$ taken at node p. Assuming the noise is independent across nodes, finding the maximum log-likelihood estimate of $\bar{\theta}$ can be written as (P) with $\cap_{p=1}^P S_p = \{1, \ldots, n\}$. For example, if the noise is Gaussian with zero mean and its covariance is the identity matrix, each $f_p(x)$ is given by $(1/2)\|x - \theta_p\|^2$, and the resulting problem is known as the *average consensus problem* [7]. In this case, the solution to (P) is simply $x^\star = (1/P)\sum_{p=1}^P \theta_p$, that is, the maximum log-likelihood



(a) Computation (b) Communication

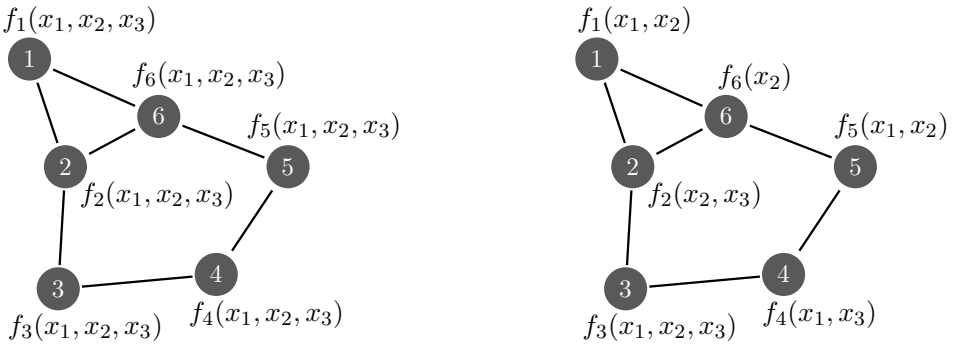
FIGURE 1.2: The two steps of a distributed algorithm. The nodes iteratively perform (a) computations and (b) broadcast the results of those computation to their neighbors.

estimation of $\bar{\theta}$ is the average of all the measurements. However, in our distributed scenario, node p is the only node who knows θ_p, and this makes computing the above average challenging. This simple example shows that to compute a solution of (P) the nodes have to communicate, either by exchanging their private data or by exchanging their estimates of the problem's solution. What they exchange and how they do it is determined by the distributed algorithm they use.

Distributed algorithms. A distributed algorithm computes a solution $x^\star$ of (P) while satisfying the requirement that each function f_p remains private to node p. Typically, each iteration of a distributed algorithm consists of the two steps shown in Figure 1.2: (a) a computation step, and (b) a communication step. In the computation step, all nodes update their estimates of the components of $x^\star$. Usually, each node p updates its estimates by combining information given by its private function f_p with information given by the estimates of its neighbors from the prior communication step. All these estimates are then exchanged in the subsequent communication step. Although all nodes in Figure 1.2 are performing each of the two steps in parallel, this is not required for a distributed algorithm. Actually, as we will see, in environments such as wireless networks it might be impossible to perform the communication step (b) in parallel, because of packet collisions. In the average consensus example given above, a popular choice for the computation step (a) is to linearly combine the estimate of node p with the estimates of its neighbors $\mathcal{N}_p$. That is, the estimate of node p, x_p, is updated as

$$x_p^{k+1} = a_{pp}\, x_p^k + \sum_{j \in \mathcal{N}_p} a_{pj}\, x_j^k\,, \tag{1.1}$$

where each a_{pj} is a positive number, $a_{pp} + \sum_{j\in\mathcal{N}_p} a_{pj} = 1$, and k denotes the iteration number. The computation scheme (1.1) implies that the nodes exchange their estimates x_p^k at each communication



(a) Global variable

(b) Non-connected variable

FIGURE 1.3: Two instances of (P) for a variable with 3 components, $x = (x_1, x_2, x_3)$. In (a), the variable is *global* (and thus connected) because all the functions depend on all the components. In (b), the variable is *non-connected* because x_1 induces a subgraph that is not connected.

step (Figure 1.2(b)). This family of algorithms for the average consensus problem has been widely studied in the literature [7, 8, 9, 10, 11, 12].

In this thesis, we propose algorithms that solve not only the average consensus problem, but the entire class (P). We will see that this class contains several other problems that are relevant in signal processing, control theory, machine learning, and other areas. Solving (P) in full generality, however, is challenging because the sets S_p are arbitrary. Our approach consists of identifying particular cases of (P) that are easier to solve, design algorithms for those cases, and then generalize them to the most difficult cases. To do that, we introduce a scheme to classify instances of (P), as overviewed next. The outcome of our approach will be an algorithm solving (P) in full generality. Despite its generality, our algorithm achieves performances better than prior distributed algorithms, even including some that were designed for a particular application.

Classification scheme. The most popular instance of (P) is illustrated in Figure 1.3(a): each function depends on *all* the components of the variable, $\cap_{p=1}^{P} S_p = \{1, \dots, n\}$. Rewriting (P) for this case, we have

$$\underset{x}{\text{minimize}} \;\; f_1(x) + f_2(x) + \cdots + f_P(x)\,, \tag{G}$$

which is the instance of (P) for which most distributed algorithms have been designed. In our classification scheme, formally introduced later in Section 1.3 and visualized in Figure 1.7, we say that problem (G) has a *global variable*. Although many applications can be written as (G), many others are instances of (P) with a non-global variable. In fact, our main motivation for considering the generic problem (P) stems from its ability to model problems where each node is interested only in a subset of the problem's parameters or variables, rather than in all of them. This is typical in large-scale systems, for example, in large plants, in the power grid, and in the internet.

A fundamental assumption we make is that if node p depends on components x_{S_p}, then that node is interested in computing the optimal value for those components only, and not for any of the other components. For example, node 4 in Figure 1.3(b) depends on components x_1 and x_3, which means that it will compute the optimal value for these components, but not for x_2. The flexibility introduced in (P) by the sets S_p, however, produces instances that are difficult to solve, given the previous assumption. Figure 1.3(b) shows an example: the component x_1 appears in the functions of nodes 1, 3, 4, and 5, but not in the functions of nodes 2 and 6. This means that node 1 is "isolated" from all the other nodes that also depend on x_1; indeed, nodes 2 and 6 are not interested in computing an optimal value for x_1, let alone exchanging estimates of it. In other words, the subgraph of the nodes that depend on x_1 is not connected and, for this reason, we say that the variable in this case is *non-connected*. Of course, computing an optimal solution of (P) in this case will invariably require selecting one of the nodes 2 or 6 to retransmit estimates of x_1, so that all the nodes depending on this component can agree on an optimal value for it. In the small example of Figure 1.3(b), it is indifferent to select either node 2 or node 6 for this task, but in larger networks, and for arbitrary sets S_p, we should select the nodes in such a way that the total number of communications is minimized. Our solution for this problem involves computing Steiner trees and is explained in Chapter 4.

The concepts of global variable and non-connected variable are concepts of the classification scheme we introduce in this thesis. These concepts and the ones of connected, mixed, and star-shaped variable will be formally defined in Section 1.3, but their relation can be visualized in Figure 1.7. Roughly, the variable of (P) is divided into two classes: connected and non-connected. These are, in fact, the most relevant classes in our classification scheme for two reasons: they form a partition of the all the instances of the variable of (P), and addressing them requires completely different techniques. These two classes thus comprise the first level of our classification scheme. The second level consists of the following subclasses: global, star-shaped, and mixed. These subclasses neither are mutually disjoint nor do they cover all instances of the variable of (P). However, they are relevant both because they are much simpler instances of (P), and because they have been solved with several distributed algorithms. Most of the algorithms that solve these subclasses, however, cannot be easily generalized to solve the entire connected and non-connected classes. In this thesis, we propose an algorithm that solves (P) for all classes and subclasses of variables.

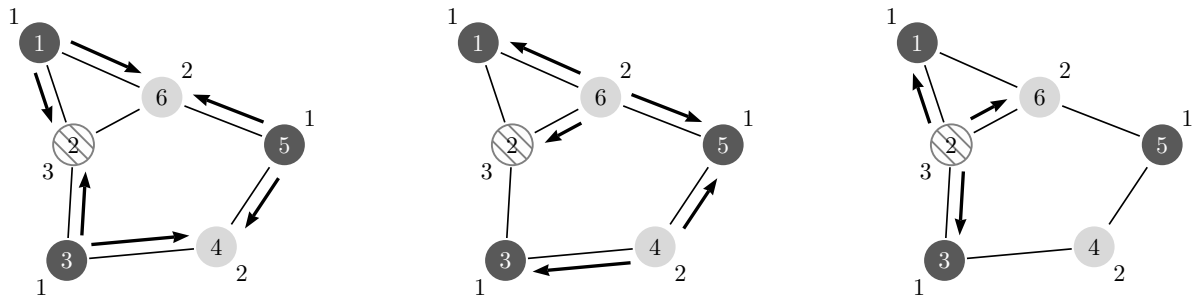
Overview of some applications. In this thesis we will consider several applications that arise in distributed contexts and that can be written as instances of (P). The recent field of compressed sensing [13, 14] provides a rich collection of such problems: basis pursuit (BP) [15], basis pursuit denoising (BPDN) [15], and the least absolute shrinkage and selection operator (lasso) [16], among others. These compressed sensing problems are convex and provide heuristics for finding sparse

solutions of linear systems. Although finding the sparsest solution of a linear system is NP-hard, compressed sensing theory establishes conditions under which the previous problems find an optimal (i.e., sparsest) solution. There is an increasing interest in solving compressed sensing in distributed scenarios, where either the columns or the rows of the matrix defining the linear system are spread over several nodes. We reformulate the above compressed sensing problems as (P), some with a global variable and others with a mixed one; some of these reformulations are novel and are presented in this thesis for the first time.

We will see that training a support vector machine (SVM)[17, Ch.7] requires solving an optimization problem that can be easily recast as (G). Roughly, given a database with two classes of datapoints, the goal in training an SVM is to find the hyperplane that best separates the two classes of datapoints. When the datapoints are distributed among several sites, training an SVM arises naturally as a distributed optimization problem. Therefore, solving this problem with a distributed algorithm has the advantages of not requiring the transmission of the private databases to a remote location, and of providing more robustness (if one node fails, the remaining nodes can still train the SVM, yet, with less data).

Many systems can be modeled as networked dynamical systems [18]. Specifically, each system is seen as the node of a network and has associated a state, a control input, or both. The state of a given node is influenced not only by its own state and control input (or simply, input), but also by the states and inputs of its neighbors. An effective control strategy for this type of systems is distributed model predictive control (D-MPC) [19], which consists of the following. First, at each time instant, each node senses its own state; then, the nodes collectively solve an optimization problem that finds the best set of control inputs for a future time-horizon. These inputs are computed in such a way that their application to the systems will lead the nodes' states to a given goal and, at the same time, they will minimize some "energy function." Although the nodes know an optimal set of inputs for all the time instants in the time-horizon, they will only use the input for the next time instant. The reason is to mitigate the impact of modeling and sensing errors. So, in the next time instant, after applying the previously computed input, each node senses its state and cooperates with the other nodes to solve the D-MPC optimization problem, now with new data. This procedure is repeated at each time instant. In this thesis, we provide a new framework for formulating D-MPC problems, and also communication-efficient algorithms to solve them.

We also mention that several network flow problems can be recast as (P) with a star-shaped variable. These are optimization problems formulated on directed networks where physical items can flow through the edges of the network. As a consequence, certain conservation laws have to be satisfied and are typically written as problem constraints. Network flow problems arise in several contexts [20], for example, in determining best energy policies in the power grid. After



(a) Step 1 (b) Step 2 (c) Step 3

FIGURE 1.4: Illustration of how the algorithms operate according to the coloring of the network. The coloring scheme has three colors: nodes 1, 3, and 5 have color 1, nodes 4 and 6 have color 2, and node 2 has color 3.

some reformulations, network flow problems can be recast as (P) and, hence, can be solved with the algorithms we propose here.

Overview of the proposed algorithms. Problem reformulation plays a key role in the design of distributed optimization algorithms. In fact, we will see throughout this thesis that it impacts significantly the final algorithm. Our strategy for solving instances of (P), and ultimately (P) in full generality, consists of reformulating those instances into a format such that well-known centralized optimization algorithms become naturally distributed.

Our reformulations make use of a concept that has rarely appeared in high-level distributed algorithms, such as the ones considered in this thesis. That concept is *network coloring*, an assignment of colors to the nodes of a network such that no two neighboring nodes have the same color (for convenience, instead of colors, we just use natural numbers). Assuming that a coloring scheme is available beforehand is realistic in many distributed scenarios, especially in wireless networks. For example, wireless networks require protocols known as media access control (MAC) to avoid packet collisions, i.e., that one node receives two messages at the same time and in the same frequency (assuming there is only one receive antenna). Some MAC protocols, such as time division multiple access (TDMA), rely on network coloring.

Figure 1.4 shows how the algorithms we propose work as a function of the coloring scheme. The network in this figure has three colors: nodes 1, 3, and 5 have color 1, nodes 4 and 6 have color 2, and node 2 has color 3. The algorithms we propose are iterative, and each iteration is divided into a number of steps equal to the number of colors. Figure 1.4 thus has 3 subfigures, each one corresponding to a step. In each step, all the nodes with the same color perform the same tasks in parallel, as illustrated in subfigures 1.4(a), 1.4(b), and 1.4(c). These subfigures show the communication pattern occurring in each step. From an high-level point of view, the tasks

performed by node p consist of:

1. finding new estimates for the components that f_p depends on, by solving

$$\text{minimize } f_p(\cdot) + \text{“quadratic term,”}$$

that is, node p minimizes the sum of f_p and a quadratic term. That quadratic term depends on the network structure as well as on previous estimates of the neighbors of node p. Solving the above optimization problem corresponds to evaluating the proximity operator of the function f_p and, many times, this can be done in a simple way.

2. sending the new estimates to the neighboring nodes.

Finally, we note that the concept of network coloring required by our algorithms coincides with the concept of network coloring commonly used in low-level communication protocols, namely, MAC protocols [21, Ch.6]. The goal of MAC protocols is to avoid packet collisions due to the hidden node and the exposed node problems [21, §6.2.2]. For example, in Figure 1.4(a), node 6 is receiving simultaneous messages from nodes 1 and 5. If the messages are in the same frequency and node 6 has one antenna only, this results in a packet collision and the nodes have to retransmit their messages. Time division multiple access (TDMA), for example, is a MAC protocol that avoids packet collisions by using a second-order coloring scheme: each node cannot have the same color as its neighbors and as its neighbors' neighbors. Such a coloring scheme works for our algorithms as well and, for this reason, the high-level structure of our algorithms is not altered by low-level protocols when they are implemented in networks that use TDMA as a MAC protocol.

The strategy we use to derive our distributed algorithms consists of reformulating the problems we want to solve in such a way that we can apply well known centralized optimization algorithms. Regarding our choice for these algorithms, we will focus on the *Alternating Direction Method of Multipliers* (ADMM), more specifically on an extended version of it: the multi-block, or extended, ADMM [22]. ADMM was proposed in the seventies by [23, 24] to solve linearly constrained optimization problems, using a “divide-and-conquer” approach. In the eighties and nineties, ADMM was shadowed by the popular interior point methods, which solve small- and medium-sized problems very efficiently, but in a centralized way. Lately, ADMM has regained attention from the optimization community, because of its wide applicability and its ability to deal with large-scale and distributed scenarios. Notably, ADMM has been applied to solve some of the problems addressed in this thesis, in particular, instances of (P) with a global variable and with a star-shaped variable. In spite of that, little is still known about its behavior. For instance, partial results on the convergence rate of ADMM, or a proof of the convergence of the multi-block ADMM, were established only very recently.

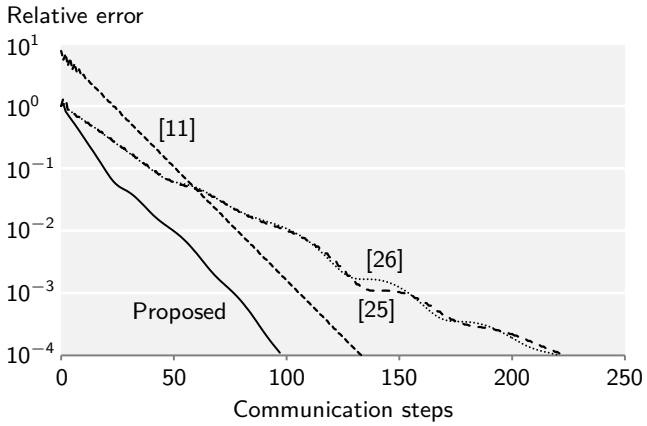


FIGURE 1.5: Comparison of the performance of a proposed algorithm with prior algorithms for the average consensus problem. The network is a randomly generated geometric network with 2000 nodes.

Example of performance results. Figure 1.5 shows as example the performance of the algorithm we propose for (G) when applied to the average consensus problem (scalar case, i.e., $n = 1$). The plot shows the relative error of the solution estimates versus the number of communication steps. We say that a communication step has occurred whenever all the nodes have updated their estimates and transmitted them to their neighbors. This is equivalent to saying that the number of communication steps is the total number of communications divided by $2E$, twice the number of edges. A communication, in this case, is defined as an "edge usage." For example, in Figure 1.4, there are 6 communications in (a), 5 communications in (b), and 3 communications in (c). And one communication step in that figure is comprised of steps 1, 2, and 3. The number of communication steps, therefore, provides a direct measure of communication-efficiency. The network we used in the experiments of Figure 1.5 has $P = 2000$ nodes and was randomly generated as a geometric network with parameter $\sqrt{\log(P)/P} \simeq 0.06$. The figure shows that, among all algorithms, the proposed one required the least amount of communications (i.e., communication steps) to achieve any error between 10^0 and 10^{-4}. The other algorithms in the figure are [25, 26], which solve the entire global problem class (G), and [11], which is considered the most efficient consensus algorithm [9], but it can only solve the average consensus problem and not any other problem in the class (G). Actually, if we consider the convergence rate, i.e., the slopes of the error lines, the proposed algorithm and [11] have roughly the same performance. In fact, they have the same slope, but [11] exhibits an offset, since it requires a special initialization. All the other algorithms were initialized with zeros.

This experimental result reveals the surprising fact that, although the algorithms we propose solve an entire problem class, they can sometimes achieve the same performance as the best algorithms for a particular application. This is particularly surprising for the average consensus problem, since it is the simplest and the most thoroughly studied distributed problem.

1.2 Goals of the thesis

We next summarize the goals of the thesis and then we explain each requirement in detail.

We aim to design, analyze, and implement algorithms that solve optimization problems of the form (P) on networks. The algorithms should be

Distributed: no node has complete knowledge about the problem data and no central node is allowed; also, each node communicates only with its neighbors;

Communication-efficient: the number of communications they use is minimized;

Network-independent: the algorithms run on networks with arbitrary topology and their output is independent of the network.

Distributed. A distributed algorithm only makes sense in an environment where both the data and the computing power are distributed. In such an environment, an algorithm is considered distributed if it fulfills three requirements. First, local data to a given node should remain private to that node. This enforces local computations, since any computation involving a piece of data has to be performed at the node where that data belongs to. In our problem (P), data of node p is encoded in the function f_p and, thus, we require f_p to be private to node p; this means that no other node has full knowledge of f_p at any time during and before the execution of the algorithm.

The second requirement is that there should not exist any central or special node. Such a node would coordinate all the other nodes and would make the data at a given node reachable to any other node in a very small number of hops. Actually, an algorithm that satisfies the first requirement but not the second one is usually called a *parallel algorithm* [27]. Indeed, according to [27], parallel algorithms run on systems where computing devices are at a small distance of each other and may be controlled by a central entity. Distributed algorithms, in contrast, run on systems where computing devices are located far apart, making centralized coordination inconvenient; in the latter, there is also little control on the network topology.

Finally, the third requirement is that each node communicates only with neighboring nodes. Although this is equivalent to forbidding a central node, we explicitly state this requirement in order to exclude platforms that allow all-to-all communications. For example, an algorithm running on a computer cluster and using function calls from a message passing interface (MPI) [28] implementation, such as `MPI_Bcast` or `MPI_Reduce`, cannot be considered distributed; at most, it is parallel. We mention that sometimes distributed algorithms are also referred to as *decentralized algorithms*.

Communication-efficient. In a centralized algorithm, the execution time and the closely related floating-point operation (FLOP) count are the most common performance metrics: the lower these metrics are, the more efficient an algorithm is. In distributed scenarios, however, other metrics arise. For example, computing accurate solutions is challenging in scenarios where there is communication noise. In that case, slower algorithms that are noise-resilient may be preferable to faster algorithms that are noise-sensitive. Another example is energy consumption. In many distributed scenarios, e.g., sensor networks, nodes rely on batteries and therefore have a limited source of energy. In these situations, increasing the lifespan of the network becomes the main priority. As communication in battery-operated devices is currently the most energy-consuming operation [6, 29], this priority translates into having algorithms with low communication requirements. The performance metric adopted in this thesis will then be the number of communications: the lower the number of communications an algorithm uses, the more efficient that algorithm will be. Hence, our goal will be to design distributed optimization algorithms that use the fewest communications possible.

Network-independent. The last requirement we impose on distributed algorithms is *network independence*. This simply means that the output of the algorithm, i.e., the estimate of the solution returned by the algorithm, should be independent of the network topology. For instance, the algorithms should output the same solution estimate whether they are run on a densely or on a sparsely connected network. Naturally, the performance of the algorithms, i.e., the number of iterations or communications they use to compute that estimate, will in general vary with the network topology.

1.3 A classification scheme for distributed optimization

Our strategy for achieving the goals of this thesis is a divide-and-conquer one: first, we identify instances of (P) that are easier to solve; then, we combine the solutions we designed for the simpler instances to solve (P) in full generality. For convenience, we reproduce (P) here:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f_1(x_{S_1}) + f_2(x_{S_2}) + \cdots + f_P(x_{S_P})\,. \tag{P}$$

In this section, we formally introduce our classification scheme for the variable of problem (P). Before doing that, however, we need the concept of *communication network*.

1.3.1 Communication network

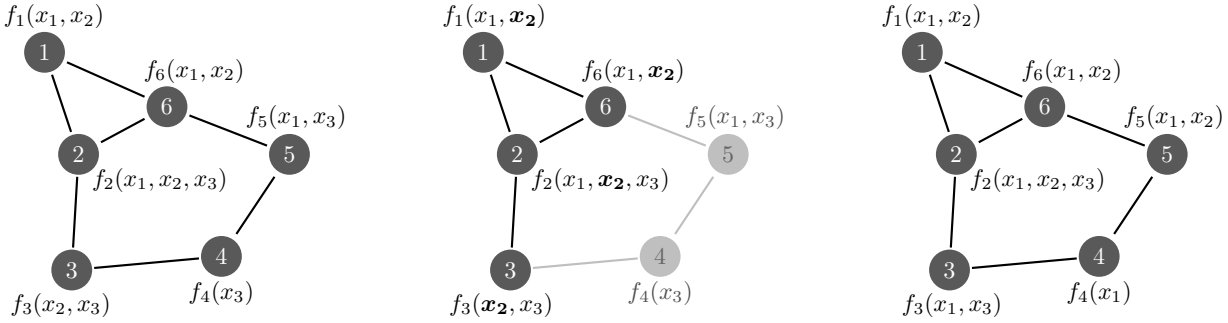
The communication network is the physical network through which the computing devices, seen as network nodes, communicate. We represent the communication network with an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the set of nodes and the set of edges, respectively. The cardinality of these sets, i.e., the number of nodes and the number of edges, will be denoted with $P = |\mathcal{V}|$ and $E = |\mathcal{E}|$, respectively. Figure 1.1 shows an example of a graph representing a communication network with $P = 10$ nodes and $E = 21$ edges. An edge belongs to the communication network, say $(i, j) \in \mathcal{E}$, if and only if nodes i and j communicate directly. For example, nodes 1 and 10 in Figure 1.1 are neighbors: this means they can exchange messages with each other, because there is a communication link connecting them. We use the following convention: if $(i, j) \in \mathcal{E}$, then $i < j$. Throughout this thesis, we will assume that the communication network $\mathcal{G}$ is connected and that its topology does not vary with time.

Functions associated to nodes. Associated with each node, there is a function depending on the components of a variable $x \in \mathbb{R}^n$. The function at node p is denoted with $f_p : \mathbb{R}^{n_p} \to \mathbb{R} \cup \{+\infty\}$, where n_p is the cardinality of the set S_p. As explained before, we use $S_p \subseteq \{1, \ldots, n\}$ to denote the components of $x \in \mathbb{R}^n$ that function f_p depends on. Of course, $1 \leq n_p = |S_p| \leq n$. We assume that each node p is interested in computing the optimal value only of the components of x that are indexed by S_p. We will see that in some situations, however, it is difficult, or even impossible, to solve instances of (P) without forcing some nodes to receive and transmit components of x that are not indexed by their sets S_p. To make our problem well-defined, we assume that each component of the variable appears in at least one of the nodes, that is, $\cup_{p=1}^{P} S_p = \{1, \ldots, n\}$.

Unless otherwise stated, we that assume f_p is closed and convex [1, 3, 2, 30, 31], and not identically $+\infty$. Note that our definition for each f_p allows it to take the value $+\infty$; as a consequence, node p can impose constraints on the variable x implicitly, via indicator functions. An indicator function of a given set $S \subset \mathbb{R}^n$ is defined as $\mathrm{i}_S : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$,

$$\mathrm{i}_S(x) = \begin{cases} 0 & , \ x \in S \\ +\infty & , \ x \notin S \, . \end{cases}$$

Including an indicator function $\mathrm{i}_S(x)$ in the objective of a minimization problem forces $x \in S$, since otherwise the optimal (minimal) value is $+\infty$. Each function f_p is *private*, i.e., at all times during and before the execution of the algorithm, only node p knows f_p. As explained before, this privacy rule formalizes our wish to derive a distributed algorithm by enforcing local computations; namely, all computations involving f_p have to be done at node p. This makes sense in scenarios where each f_p encodes a database that should be known only at node p, or simply to make use of all the



(a) Generic connected variable (b) Subgraph induced by x_2 (c) Mixed connected variable

FIGURE 1.6: Example of (a) a generic connected variable and (c) a mixed connected variable. (b) highlights the subgraph induced by the component x_2 in (a). The communication network is the same in all cases. In (c), the component x_1 is global, and x_2 and x_3 induce connected subgraphs.

distributed computing resources as, for example, in a sensor network, where each sensor has some processing power available for computation.

1.3.2 Variable classification

Although each function is uniquely associated to a single node, the same does not happen for each component of $x \in \mathbb{R}^n$, the optimization variable. This creates an additional structure and motivates our classification scheme. Essential to our classification scheme is the concept of induced subgraph.

Induced subgraph. Let $x_l \in \mathbb{R}$ denote the lth component of the optimization variable $x \in \mathbb{R}^n$. We define the subgraph induced by x_l similarly to how [32, Ch.1] defines the subgraph induced by a set of nodes. In our case, these nodes are the ones whose functions depend on x_l. To be more concrete, given a communication network $\mathcal{G}$, the *subgraph induced by* x_l is the subgraph $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l) \subseteq \mathcal{G}$, where $\mathcal{V}_l$ is the set of nodes whose functions depend on x_l, and an edge (i, j) belongs to $\mathcal{E}_l$ only if $(i, j) \in \mathcal{E}$ and both nodes i and j belong to $\mathcal{V}_l$. As an example, Figure 1.6(b) highlights the subgraph induced by the component x_2 in the setting of Figure 1.6(a): the set of nodes and the set of edges of this induced subgraph $\mathcal{G}_2$ are, respectively, $\mathcal{V}_2 = \{1, 2, 3, 6\}$ and $\mathcal{E}_2 = \{(1, 2), (1, 6), (2, 3), (2, 6)\}$. Note that neither f_4 nor f_5 depend on x_2.

Component-wise classification of $\boldsymbol{x}$. We classify each component x_l according to its induced subgraph $\mathcal{G}_l$ the following way: x_l is

- *connected* if $\mathcal{G}_l$ is a connected subgraph, and is *non-connected* otherwise;
- *global* if its induced subgraph coincides with the communication network, i.e., $\mathcal{G}_l = \mathcal{G}$;
- *star-shaped* if $\mathcal{G}_l$ is a star graph.

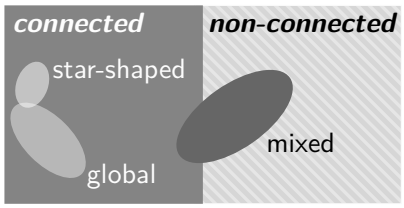


FIGURE 1.7: Our classification scheme for the variable $x \in \mathbb{R}^n$ of problem (P). The variable is either connected or non-connected. Global and star-shaped variables are particular instances of a connected variable, and a mixed variable can be connected or non-connected.

By star graph we mean a graph in which there exists a node who is a neighbor of all the other nodes; the remaining nodes can also be neighbors between themselves. For example, the subgraph induced by variable x_2 in Figure 1.6(b) is a star, because every node is a neighbor of node 2; therefore, x_2 is star-shaped. If a component is star-shaped, it can be handled in a centralized way, since the node in the center of the star can act as a central node. It can be checked that component x_1 in Figure 1.6(a) is also star-shaped, with node 6 in the center, but component x_3 is not. All the components of the variable in that figure, however, are connected, since the respective subgraphs are connected. Naturally, a star-shaped variable is always connected. An example of a global component is given in Figure 1.6(c): the subgraph induced by x_1 coincides with communication graph and, thus, x_1 is global. In other words, all the functions in Figure 1.6(c) depend on x_1. Again, a global component is always connected, since its induced subgraph coincides with the communication network, which we assume connected. Unless the communication network is a star, a global variable is never star-shaped. We had already illustrated a non-connected component in Figure 1.3(b): the subgraph induced by x_1 in that network is not connected, and thus x_1 non-connected.

Classification of x. With the component-wise classification of components, we are now in conditions to classify the full optimization variable $x \in \mathbb{R}^n$ in (P). The proposed classification scheme is shown in Figure 1.7. There, the variable x is either

- *connected* if all the components x_l of $x \in \mathbb{R}^n$ are connected, for $l = 1, \ldots, n$; or
- *non-connected* if x has at least one non-connected component.

For example, while the variable in Figure 1.6(a) is connected, because x_1, x_2, and x_3 are connected, the variable in Figure 1.3(b) is non-connected, because x_1 is non-connected (in spite of x_2 and x_3 being connected). Note that the connected and non-connected classes partition the entire class of the variable x (see Figure 1.7). This distinction between a connected and a non-connected variable is the most important one in our classification scheme. In fact, we will see in Chapter 4 that they have to be addressed with different techniques.

To the best of our knowledge, no algorithm has ever been designed (purposefully) to solve (P) with a generic connected or non-connected variable. However, there are algorithms solving it with global, mixed (connected), and star-shaped variables, defined as follows. The variable $x \in \mathbb{R}^n$ is

- *global* if all its components are global: $\cap_{p=1}^{P} S_p = \{1, \dots, n\}$;
- *mixed* if it has at least one global component and at least one non-global component: $\cap_{p=1}^{P} S_p \neq \Big\{\emptyset, \{1, \dots, n\}\Big\}$;
- *star-shaped* if all its components are star-shaped.

We have already seen an example of a global variable in Figure 1.3(a). When (P) has a global variable, it can be written simply as

$$\underset{x}{\text{minimize}} \;\; f_1(x) + f_2(x) + \cdots + f_P(x)\,. \tag{G}$$

Because of the assumption that the communication network is always connected, a global variable is also always connected. A mixed variable, in turn, can be either connected or non-connected. It is connected if all the non-global components are connected, and non-connected if at least one of the non-global components is non-connected. Figure 1.6(c) shows an example of a mixed variable that is connected, since the non-global components x_2 and x_3 are connected. Problem (P) with a mixed variable can be written as

$$\underset{x=(y,z)\in\mathbb{R}^n}{\text{minimize}} \quad f_1(y, z_{S_1}) + f_2(y, z_{S_2}) + \cdots + f_P(y, z_{S_P})\,, \tag{M}$$

where the variable x was decomposed into its global components y and into its non-global components z. Finally, all the components of a star-shaped variable are like x_2 in Figure 1.6(b). Note that in Figure 1.7 the star-shaped class intersects with the global class; this happens when the variable is global and the communication network is a star.

Summarizing, our classification scheme partitions the variable of problem (P) into two classes, shown as rectangles of Figure 1.7: connected and non-connected. These classes are the most fundamental ones, since they require different solution methods. The subclasses shown as ellipsoids in Figure 1.7 identify easier instances of (P). Also, each one of these subclasses has been addressed with prior distributed optimization algorithms. In reality, while several algorithms have been proposed for the global and the star-shaped subclasses, we only found one distributed algorithm, in [33], solving an instance of (P) with a mixed variable. That instance is actually a very particular one: the variable is connected and all the non-global components are star-shaped. The classification scheme of Figure 1.7 will also guide us throughout the thesis: we first address the global subclass,

which is not only the subclass for which most of the distributed optimization algorithms have been proposed, but also the simplest one; in particular, the notation required to handle problems in this subclass is simpler, since we do not need the indexing sets S_p. Then, we address the connected class, by generalizing the algorithm for the global class. And, finally, we generalize the connected class algorithm to handle both a connected and a non-connected variable, that is, to handle any instance of (P).

1.4 Contributions

We list the main contributions of this thesis:

- We provide a classification scheme for the class (P) of distributed optimization problems. Although it borrows some aspects from factor graphs [34] (actually the same aspects that are used in [35]), it establishes a relation between the (abstract) optimization problem to be solved and the (concrete) computational platform, in our case, the communication network. This classification scheme plays a fundamental role in the thesis, not only by providing a framework to develop our algorithms, but also by allowing us to organize prior work and applications.

- We develop a set of distributed algorithms to solve (P) that are communication-efficient. The order in which we present our algorithms in the next chapters goes from the most specific to the most general, a pattern that corresponds to the order in which they were developed. More specifically, we first present an algorithm for the global class (G), then we present an algorithm for the connected class and, lastly, we present an algorithm that solves any instance of (P). All these algorithms are distributed, network-independent, and communication-efficient. In particular, we will see that they usually require less communications to converge than prior distributed algorithms.

- We apply our algorithms to several application problems from engineering and computer science. Some of these applications are novel, i.e., to the best of our knowledge, they have never been solved with distributed algorithms. This includes several compressed sensing problems and distributed model predictive control (D-MPC). Actually, we propose a new framework for D-MPC that considerably extends the modeling capability of the prior framework.

- To assess the performance of our algorithms, we provide extensive benchmarks with prior algorithms. This required an implementation of all the algorithms, including prior distributed optimization algorithms, and also algorithms that are specific to a given application.

1.5 Organization

The remainder of the thesis is organized as follows.

- In Chapter 2, we provide background on distributed and parallel algorithms for optimization including, for example, decomposition methods and the alternating direction method of multipliers. Although these methods are not distributed, they work as building blocks for distributed algorithms, which are presented subsequently. We also discuss prior distributed algorithms, organized according to the subclasses defined by our classification scheme.

- In Chapter 3, we present our algorithm for the global class, which is not only the most common class, but also the simplest one conceptually and notationally. This will allow us to introduce our main ideas without complicated notation. The chapter starts by stating the problem formally and discussing the general assumptions we make. Then, several applications are given, some of which are novel. Finally, the algorithm is derived and experimental results are shown.

- Chapter 4 has a similar structure, but now addresses our main problem (P) in full generality. It starts with restating the problem and discussing the assumptions. Next, several potential applications are shown, including a new framework for D-MPC. Also, we present in that chapter the only algorithm we found in the literature that, after some adaptations, can also solve our problem in full generality. After that, we derive our algorithm, first assuming a general connected variable, and then moving to a non-connected one. The chapter ends with the presentation of several experimental results.

- Our conclusions and possible directions for future work are presented in Chapter 5. There, we also restate our major contributions and discuss current limitations of our algorithms.

Chapter 2

Background and Related Work

Distributed and parallel algorithms not only are relevant in the real world, but also are challenging to design. They provide several advantages over their centralized counterparts, for example, the ability to process distributed data and, notably, significant computational speed-ups. In the context of optimization, parallel algorithms, including decomposition methods, date back to the sixties with the works of Dantzig and Wolfe [36], Benders [37], and Everett [38]. Research on distributed optimization algorithms started later, in the mid-eighties, with the work of Tsitsiklis, Bertsekas, and Athans [39] and has boomed in the past ten years, motivated by the widespread of sensor networks [4]. Nowadays, distributed optimization finds application in sensor networks (localization [40], clustering [41], etc), in cognitive radio [42], in machine learning [43, 44, 45, 35], and in the control of complex systems such as irrigation canals [46] and the power grid [47, 48, 49, 50, 51].

In this chapter, we use the classification scheme developed in the previous chapter to organize existing work on distributed optimization. Since most of this work builds upon parallel methods, we first overview relevant work on parallel methods, with special emphasis on the *Alternating Direction Method of Multipliers* (ADMM), since it will play a key role in this thesis.

2.1 Building blocks: non-distributed, parallel algorithms

We start by reviewing some methods that, although not distributed, work as building blocks of distributed algorithms. There are three subsections: one dedicated to decomposition methods, another dedicated to block-coordinate minimization methods, and the last one dedicated to augmented Lagrangian methods, which includes ADMM.

2.1.1 Decomposition methods

Decomposition methods are the precursors of distributed optimization methods. Their goal, as the name indicates, is to decompose a complex problem into smaller, simpler ones. Yet, they are not considered distributed, because they generally require a master node coordinating several slave nodes. The prototypical problem they solve is

$$\begin{array}{ll} \underset{x_1,\dots,x_P}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} & A_1x_1 + A_2x_2 + \cdots + A_Px_P = b\,, \end{array} \tag{2.1}$$

where the variable is $x = (x_1, \dots, x_P) \in \mathbb{R}^n$, with $x_p \in \mathbb{R}^{n_p}$, and $n_1 + \cdots + n_P = n$. Each function $f_p : \mathbb{R}^{n_p} \to \mathbb{R}$ is assumed convex. Problem (2.1) is coupled through its constraint $Ax = \begin{bmatrix} A_1 & A_2 & \cdots & A_P \end{bmatrix} x = b \in \mathbb{R}^m$, which is always assumed feasible.

Similarly to distributed methods, decomposition methods solve (2.1) by assigning a pair (f_p, A_p) to one device (or node) but, in contrast to distributed methods, all devices (or nodes) are controlled by a master node. Occasionally, the structure of the matrix A allows discarding the master node and the decomposition method becomes distributed. Decomposition methods are divided into primal and dual methods, and comprehensive references on the topic are [2, §6.4], [27, Ch.3], and [52].

Primal decomposition. To solve (2.1) through primal decomposition, we rewrite it as

$$\begin{array}{ll} \underset{y_1,\dots,y_P}{\text{minimize}} & \phi_1(y_1) + \phi_2(y_2) + \cdots + \phi_P(y_P) \\ \text{subject to} & y_1 + y_2 + \cdots + y_P = b\,, \end{array} \tag{2.2}$$

where each $y_p \in \mathbb{R}^m$ is a new variable, and each function $\phi_p : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$\begin{array}{rll} \phi_p(y_p) := & \inf\limits_{x_p} & f_p(x_p) \\ & \text{s.t.} & y_p = A_px_p\,. \end{array} \tag{2.3}$$

For simplicity, we assume that each A_p has full row rank, which implies that ϕ_p is defined over all $\mathbb{R}^m$. Given a master node and P slave nodes, the master node solves the master problem (2.2) and delegates to slave node p the task of handling computations involving ϕ_p. Typically, the master problem (2.2) is solved with a first-order minimization method, such as the projected subgradient method. It can be shown that the subgradient of ϕ_p at a point y_p is given by $-\lambda_p$, where λ_p is the (optimal) dual variable associated to the constraint of (2.3); see sections 5.4.4 and 6.4.2 of [2] for more details. Therefore, in primal decomposition, the master node updates $y = (y_1, \dots, y_P)$ as

$$y^{k+1} = \left[y^k + \alpha_k\lambda^k\right]_{\{1_n^\top y = b\}}\,, \tag{2.4}$$

where $[\cdot]_{\{1_n^\top y=b\}}$ denotes the projection onto the set $\{y \in \mathbb{R}^n : 1_n^\top y = b\}$, α_k is a positive stepsize, $1_n \in \mathbb{R}^n$ is a vector of ones, and $\lambda^k = (\lambda_1^k, \ldots, \lambda_P^k)$ is the vector of dual variables at iteration k. At each iteration, the master node sends y_p^k to slave node p, who then solves the problem in (2.3) and returns λ_p^k to the master node. The master node, in turn, updates y as in (2.4) and moves on to the next iteration. According to our previous definitions, primal decomposition is not distributed since it requires a master node playing the role of a central node.

Dual decomposition. Dual decomposition methods, rather than solving (2.1) directly, solve its dual problem instead:

$$\underset{\lambda}{\text{minimize}} \;\; f_1^\star(A_1^\top \lambda) + f_2^\star(A_2^\top \lambda) + \cdots + f_P^\star(A_P^\top \lambda) - b^\top \lambda \,, \tag{2.5}$$

where $\lambda \in \mathbb{R}^m$ is the dual variable and $f_p^\star : \mathbb{R}^m \to \mathbb{R}$ is the convex conjugate of f_p, defined as

$$f_p^\star(\lambda) = \sup_{x_p} \;\; \lambda^\top x_p - f_p(x_p) \,. \tag{2.6}$$

While (2.1) is coupled through its constraint, (2.5) is coupled through its objective (since all conjugate functions depend on λ). As in the primal decomposition, given a master node and P slave nodes, the master node solves the master problem (2.5) and delegates to slave node p the task of handling $f_p^\star$. Whenever each function f_p is strictly convex, there is only one minimizer $x_p(\lambda)$ of the problem in (2.6) for a given λ. Hence, in this case, after the master node finds a dual solution $\lambda^\star$ to (2.5), the pth block of the optimal primal solution of (2.1) can be found in the pth slave node as $x_p(\lambda^\star)$. In other words, when each function f_p is strictly convex, a primal solution is immediately available after solving the dual problem (2.5). Again, the master problem (2.5) can be solved with first-order minimization methods, such as the subgradient method. The subgradient of $f_p^\star \circ A_p^\top$, where $\circ$ denotes composition, at a point λ is $A_p x_p(\lambda)$, where $x_p(\lambda)$ solves the problem in (2.6) [2, Prop.B.25(b)]. Hence, in dual decomposition, the master node updates λ as

$$\lambda^{k+1} = \lambda^k - \alpha_k (A_1 x_1(\lambda^k) + A_2 x_2(\lambda^k) + \cdots + A_P x_P(\lambda^k) - b) \,, \tag{2.7}$$

where $\alpha_k > 0$ is the stepsize at iteration k. At each iteration, the master node sends λ^k to all slave nodes and each slave node p, in turn, returns $A_p x_p(\lambda^k)$ to the master node. Similarly to primal decomposition, dual decomposition also requires a central node (the master node) and, therefore, it is not distributed. Other dual decomposition methods are the Dantzig-Wolfe decomposition [36],[2, §6.4.1] and the Benders decomposition [37].

Whenever each f_p is strongly convex with parameter μ, $f_p^\star$ is differentiable and its gradient $\nabla f_p^\star$ is Lipschitz continuous with constant $1/\mu$ [30, Th. 4.2.2], [53]. In that case, a faster algorithm can

be applied, for example, the gradient method or even Nesterov's fast gradient method. These are explained next.

First-order minimization methods. Decomposition methods generally use first-order methods to solve the master problem, i.e., methods that use only first-order (sub)derivatives. Consider, for example,

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & x \in X\,, \end{array} \tag{2.8}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. If f is not differentiable, an appropriate method to find a minimizer of (2.8) is the projected subgradient method:

$$x^{k+1} = \left[x^k - \alpha_k d^k\right]_X\,, \tag{2.9}$$

where x^k is the estimate at iteration k, d^k is the subgradient of f at the point x^k, i.e., $d^k \in \partial f(x^k)$,[1] $[\cdot]_X$ is the projection operator onto the set X, and $\alpha_k > 0$ is the stepsize at iteration k. The projected subgradient method is non-descent, that is, it does not guarantee that the cost function $f(x^k)$ decreases at every iteration. However, under the assumption that f is Lipschitz continuous, i.e., that there exists $L > 0$ such that $\|f(y) - f(x)\| \leq L\|y - x\|$ holds for all x, y, and under an appropriate choice for the stepsize sequence $\{\alpha_k\}_{k=0}^{\infty}$, the best cost function estimate $f_{\text{best}}^k := \min_{0\leq l\leq k} f(x^l)$ converges to the optimal value $f^\star$ of (2.8). This is guaranteed, for example, by a square summable but not summable stepsize sequence, for instance, $\alpha_k = 1/(1+k)$. A constant stepsize sequence $\alpha_k = \alpha$, for all k, in contrast, only guarantees that f_{best}^k converges to a neighborhood of $f^\star$. Even when convergence is guaranteed, the method is rather slow, since $f_{\text{best}}^k - f^\star$ converges to zero at rate $O(1/\sqrt{k})$. Extensive information about subgradient methods can be found in [54, §8.2],[31, Ch.3] [55, 56].

When the function f is continuously differentiable and its gradient ∇f is Lipschitz-continuous with constant L, i.e., $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ for all $x, y \in \mathbb{R}^n$, more efficient methods can be applied. In fact, for a differentiable function, $\partial f(x) = \{\nabla f(x)\}$, and the iterations (2.9) become the projected gradient method. In contrast with subgradient methods, gradient methods are descent and converge even with a fixed stepsize $\alpha_k = \alpha \in (0, 1/L]$, for all k. Moreover, $f(x^k) - f^\star$ converges to zero at rate $O(1/k)$. Gradient methods are studied extensively in [2, Ch.1,2][31, Ch.1,2][57, 58, 59].

Surprisingly, a small modification of the projected gradient yields a method whose error $f(x^k) - f^\star$ decreases at rate $O(1/k^2)$, as discovered by Nesterov. The problem assumptions are the same

[1] The subdifferential ∂f of a convex function f at a point x is defined as $\partial f(x) = \{d : f(y) \geq f(x) + d^\top(y-x)\,, \forall_y\}$. Any point d belonging to the subdifferential $\partial f(x)$ is called subgradient of the function f at the point x.

as in the projected gradient method. An instance of Nesterov's method is

$$\begin{aligned} x^{k+1} &= \left[y^k - \alpha_k \nabla f(y^k)\right]_X \\ y^{k+1} &= x^{k+1} + \tfrac{k-1}{k+2}(x^{k+1} - x^k)\,, \end{aligned} \tag{2.10}$$

which requires no significant additional computation with respect to (2.9). Yet, it not only has better bounds on the rate of convergence, but it also converges much faster in practice. For more information about accelerated first-order methods, see [31, 58, 59, 60, 61, 62, 63, 64].

2.1.2 Block-coordinate minimization methods

Block-coordinate methods are appropriate when fixing some of the variables in an optimization problem makes the problem easier to solve. Consider, for example,

$$\begin{array}{ll} \underset{x=(x_1,\ldots,x_P)}{\text{minimize}} & f(x_1, x_2, \ldots, x_P) \\ \text{subject to} & x \in X_1 \times X_2 \times \cdots \times X_P\,, \end{array} \tag{2.11}$$

where the variable is $x = (x_1, \ldots, x_P) \in \mathbb{R}^n$ with $x_p \in \mathbb{R}^{n_p}$ and $n_1 + \cdots + n_P = n$. The function $f : \mathbb{R}^n \to \mathbb{R}$ is assumed convex, and each set $X_p \subseteq \mathbb{R}^{n_p}$ is assumed closed and convex. Block-coordinate minimization methods solve (2.11) via a sequence of minimization problems with respect to one block variable while the other blocks are fixed, i.e.,

$$\begin{array}{ll} \underset{\xi}{\text{minimize}} & f(x_1, \ldots, x_{p-1}, \xi, x_{p+1}, \ldots, x_P) \\ \text{subject to} & \xi \in X_p\,. \end{array}$$

Two important types of block-coordinate methods are nonlinear Jacobi and nonlinear Gauss-Seidel.

Nonlinear Jacobi. The nonlinear Jacobi method is defined as

$$\begin{array}{rll} x_p^{k+1} = & \underset{x_p}{\arg\min} & f(x_1^k, \ldots, x_{p-1}^k, x_p, x_{p+1}^k, \ldots, x_P^k)\,, \qquad p = 1, \ldots, P\,, \\ & \text{s.t.} & x_p \in X_p \end{array} \tag{2.12}$$

where the pth minimization is taken with respect to x_p. Since updating x_p^k to x_p^{k+1} requires all the other block components to be fixed at x_j^k, for $j \neq p$, which were found in the previous iteration, the updates can be carried out in parallel. Convergence of the nonlinear Jacobi method to a minimizer of (2.11) is guaranteed whenever f is differentiable and the mapping $x - \gamma \nabla f(x)$ is a contraction for any $\gamma > 0$ [27, Prop.3.10]. Another Jacobi-type method requiring milder assumptions is the diagonal quadratic approximation [65, 66].

Nonlinear Gauss-Seidel. The nonlinear Gauss-Seidel method is defined as

$$x_p^{k+1} = \begin{array}{ll} \arg\min\limits_{x_p} & f(x_1^{k+1}, \ldots, x_{p-1}^{k+1}, x_p, x_{p+1}^{k}, \ldots, x_P^{k}) \,, \\ \text{s.t.} & x_p \in X_p \end{array} \qquad p = 1, \ldots, P \,. \tag{2.13}$$

In contrast with Jacobi methods, updating x_p at iteration k requires knowing the current estimates of the first $p-1$ blocks, i.e., x_j^{k+1} for $j < p$. Hence, all updates have to be carried out sequentially. The order of the sequence, however, can change from iteration to iteration, and the convergence to a minimizer of (2.11) is guaranteed whenever each problem in (2.13) has a unique solution and a regularity condition is satisfied [67]. For example, differentiability and strict convexity of f implies that regularity condition is satisfied (see the errata of proposition 2.7.1 of [2], available at `http://www.athenasc.com/nlperrata.pdf`).

2.1.3 Augmented Lagrangian methods

Augmented Lagrangian methods are important tools for distributed optimization, even though they were not designed for that purpose. They date back to penalty methods, where a constrained problem is solved via a sequence of unconstrained problems. Although relying on duality, augmented Lagrangian methods are guaranteed to find a primal solution even when the cost function is not strictly convex. This gives them a clear advantage over "simple" duality-based methods, such as dual decomposition. On the other hand, they do not distribute as easily as "simple" duality-based methods, because of the augmented term in the augmented Lagrangian.

Method of multipliers. Discovered independently by Hestenes [68] and by Powell [69], the method of multipliers solves the constrained problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = b \,, \end{array} \tag{2.14}$$

where the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed and convex, $b \in \mathbb{R}^m$, and the linear system $Ax = b$ is feasible. Using $\lambda \in \mathbb{R}^m$ to denote the dual variable, the augmented Lagrangian of (2.14) is

$$L_\rho(x; \lambda) = f(x) + \lambda^\top (Ax - b) + \frac{\rho}{2} \|Ax - b\|^2 \,,$$

where $\rho > 0$ is the augmented Lagrangian parameter. Note that the augmented Lagrangian differs from the ordinary Lagrangian in the augmented term $(\rho/2)\|Ax - b\|^2$. The method of multipliers solves (2.14) by minimizing the augmented Lagrangian with respect to x, keeping the dual variable λ

fixed at λ^k, and then by updating λ in a gradient-based way. That is, it iterates

$$x^{k+1} = \arg\min_x \; L_\rho(x; \lambda^k) \tag{2.15}$$

$$\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)\,. \tag{2.16}$$

Note that (2.16) is indeed a gradient iteration: the dual function $L_\rho(\lambda) := \inf_x L_\rho(x; \lambda)$ is differentiable and its gradient is given by $Ax(\lambda) - b$, where $x(\lambda)$ minimizes $L_\rho(\cdot; \lambda)$.[2] Furthermore, it can be shown that the gradient $Ax(\lambda) - b$ is Lipschitz continuous with constant $1/\rho$ [30, Th. 4.2.2], [53]. Rockafellar [70] showed that the iterations (2.15)-(2.16) are actually an application of the proximal minimization algorithm to the dual problem of (2.14) (see also [27, §3.4.4] and [71, Ch.3]). Therefore, the conditions under which the method of multipliers converges are very mild; see [72, 73, 74] [27, §3.4.4][2, §4.2] for a detailed analysis and for related methods. Nevertheless, the optimization problem in (2.15) is usually nonseparable, because of the augmented term $(\rho/2)\|Ax - b\|^2$. This makes the method of multipliers difficult to apply in distributed optimization. We next present an alternative that, while preserving the good convergence properties of the method of multipliers, it suits distributed optimization better.

Alternating Direction Method of Multipliers. The Alternating Direction Method of Multipliers (ADMM) is an augmented Lagrangian method introduced in the mid-seventies by Glowinski and Marrocco [23] and by Gabay and Mercier [24]. It solves

$$\begin{array}{ll} \underset{x_1, x_2}{\text{minimize}} & f_1(x_1) + f_2(x_2) \\ \text{subject to} & A_1 x_1 + A_2 x_2 = b\,, \end{array} \tag{2.17}$$

where $f : \mathbb{R}^{n_1} \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^{n_2} \to \mathbb{R} \cup \{+\infty\}$ are closed convex functions, and $A_1 \in \mathbb{R}^{m \times n_1}$ and $A_2 \in \mathbb{R}^{m \times n_2}$ are full column rank matrices. The augmented Lagrangian of (2.17) is

$$L_\rho(x_1, x_2; \lambda) = f_1(x_1) + f_2(x_2) + \lambda^\top (A_1 x_1 + A_2 x_2 - b) + \frac{\rho}{2}\|A_1 x_1 + A_2 x_2 - b\|^2\,,$$

where the parameter $\rho > 0$ is assumed fixed. ADMM minimizes L_ρ first with respect to x_1, then with respect to x_2, and it finally updates the dual variable λ as in the method of multipliers:

$$x_1^{k+1} = \arg\min_{x_1} \; L_\rho(x_1, x_2^k; \lambda^k) \tag{2.18}$$

[2] This is true even when A does not have full column rank. To see that, write $L_\rho(\lambda)$ as $L_\rho(\lambda) = \inf_z \Psi(z) + \lambda^\top z + \frac{\rho}{2}\|z\|^2$, where $\Psi(z) := \inf_x\{f(x) : z = Ax - b\}$ is a convex function [1, §3.2.5]. The quadratic term $\|z\|^2$ makes the objective strictly convex and, therefore, the problem defining $L_\rho(\lambda)$ in terms of z has a unique minimizer $z(\lambda)$ for each λ. It follows that the subdifferential of L_ρ is the singleton $\{z(\lambda)\}$. For each λ, there can be several $x(\lambda)$'s solving the problem defining $\Psi(z(\lambda))$.

$$x_2^{k+1} = \arg\min_{x_2} \; L_\rho(x_1^{k+1}, x_2; \lambda^k) \tag{2.19}$$

$$\lambda^{k+1} = \lambda^k + \rho(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b)\,. \tag{2.20}$$

ADMM can be seen as the application of the method of multipliers to problem (2.17), where the minimization with respect to the primal variable (x_1, x_2) consists of just one Gauss-Seidel pass. Surprisingly, it solves (2.17) with the same accuracy level as the method of multipliers does, by using a few more iterations; see [75] for a detailed comparison between ADMM and the method of multipliers. Curiously, both methods are instances of the proximal point algorithm [76, 77, 78]: while the method of multipliers results from applying iteratively the resolvent operator to the subdifferential of the dual function of (2.14) [70], ADMM results from applying iteratively the Douglas-Rachford operator [79, 80] to the subdifferential of the dual function of (2.17), as discovered by Gabay [81]. An excellent account on this topic, including an introduction to monotone operator theory, is given by Eckstein [71, Ch.3] (see also [82]). Alternative proofs for the convergence of ADMM that do not use any monotone operator theory include [27, §3.4.4] and [35, 83]. Roughly, ADMM converges whenever f and g are closed and convex, (2.17) is solvable, and strong duality holds. When A_1 and A_2 do not have full column rank, the sequence (x_1^k, x_2^k) might not converge, even though $f_1(x_1^k) + f_2(x_2^k)$ and λ^k converge [27, p.260]. Regarding the augmented Lagrangian parameter ρ, the proofs of the convergence hold for any positive, fixed ρ. Since, in practice, the value of ρ significantly affects the performance of the algorithm, it is common to use heuristics to adapt ρ along the iterations [48, 35]. These heuristics, however, cannot be easily implemented in distributed environments, because they require information from all the nodes at each iteration.

Until recently, the known proofs for the convergence of ADMM did not allow to derive a convergence rate. It was known, however, that ADMM converged linearly for linear programs [84]. More recently, a series of works has derived bounds for the convergence rate of ADMM, many times, under assumptions stronger than the ones required to prove plain convergence. For example, [85] proved that the primal and the dual variables converge in an ergodic sense at rate of $O(1/k)$. The same rate was established in [86] in a non-ergodic sense. The work [87] proved that the cost function of the dual problem converges to the optimal value at rate $O(1/k)$ and, as a consequence, the square of the primal and dual residuals [35] converge to zero at the same rate. It is assumed, however, that at least one of the functions f_1 or f_2 is strongly convex. Inspired by Nesterov's gradient method, [87] also proposes a modification to ADMM whose dual cost function converges at rate $O(1/k^2)$. Note that both $O(1/k)$ and $O(1/k^2)$ are sublinear rates.[3] When ADMM is applied to the average consensus problem (after a suitable reformulation to make it distributed, as we will

[3]We say that a sequence $\{x^k\}$ converges linearly (more appropriately, R-linearly) to $x^\star$ if there exists $M > 0$ and $c > 1$ such that $\|x^k - x^\star\| \leq \frac{M}{c^k}$, for a sufficiently large k.

see later), linear convergence can be proved [9]. For general quadratic problems, [88] conjectured that linear convergence also holds, which was later proved in [89]. More recently, Deng and Yin [90] showed that a generalized version of ADMM converges linearly in terms of the primal and the dual estimates when at least one of the functions f_1 or f_2 is strongly convex, differentiable, and has a Lipschitz continuous gradient. Work that establishes convergence rates for modified versions of ADMM includes [91, 92, 93, 94].

The recent stream of theoretical work on ADMM in recent years has been motivated by its application in many areas. For example, ADMM has been applied to image processing [95], to localization [40], and to several statistical and machine learning problems [96, 35]. Reference [35], in particular, provides a survey on ADMM from an optimization perspective and describes many applications in statistics and machine learning.

Multi-block ADMM. The multi-block ADMM is a natural generalization of ADMM when, instead of the variable being partitioned into two blocks, x_1 and x_2, as in (2.17), it is partitioned into a finite number C. Sometimes this method is also known as generalized ADMM or extended ADMM. Since there are other methods named generalized ADMM, we will refer to it as multi-block ADMM or as extended ADMM. More specifically, the multi-block ADMM solves

$$\begin{array}{ll} \underset{x_1,\ldots,x_C}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_C(x_C) \\ \text{subject to} & A_1x_1 + A_2x_2 + \cdots + A_Cx_C = b\,, \end{array} \tag{2.21}$$

by iterating

$$x_1^{k+1} = \underset{x_1}{\arg\min}\ L_\rho(x_1, x_2^k, \ldots, x_C^k; \lambda^k) \tag{2.22}$$

$$x_2^{k+1} = \underset{x_2}{\arg\min}\ L_\rho(x_1^{k+1}, x_2, x_3^k, \ldots, x_C^k; \lambda^k) \tag{2.23}$$

$$\vdots \tag{2.24}$$

$$x_C^{k+1} = \underset{x_C}{\arg\min}\ L_\rho(x_1^{k+1}, x_2^{k+1}, \ldots, x_{C-1}^{k+1}, x_C; \lambda^k) \tag{2.25}$$

$$\lambda^{k+1} = \lambda^k + \rho \sum_{c=1}^{C} A_c x_c^{k+1}\,. \tag{2.26}$$

Note that (2.21) is the same problem as (2.1), the problem solved by decomposition methods. In this case, the augmented Lagrangian is

$$L_\rho(x_1, x_2, \ldots, x_C; \lambda) = \sum_{c=1}^{C} f_c(x_c) + \lambda^\top \Big(\sum_{c=1}^{C} A_c x_c - b\Big) + \frac{\rho}{2}\Big\|\sum_{c=1}^{C} A_c x_c - b\Big\|^2\,.$$

It is assumed that each function $f_c : \mathbb{R}^{n_c} \to \mathbb{R} \cup \{+\infty\}$ is closed and convex, and that each matrix $A_c \in \mathbb{R}^{m \times n_c}$ has full column rank. When $C = 2$, the multi-block ADMM (2.22)-(2.26) becomes the 2-block ADMM (2.18)-(2.20). The only known proof of convergence of the multi-block ADMM is due to Han and Yuan [22] and it assumes that all functions $f_1, \ldots, f_C$ are strongly convex. The following theorem summarizes the known convergence results for the multi-block ADMM, including its particular version, the 2-block ADMM.

Theorem 2.1 ([83, 22])**.**

Let $f_c : \mathbb{R}^{n_c} \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function over $\mathbb{R}^{n_c}$, not identically $+\infty$, and let A_c be an $m \times n_c$ matrix, for $c = 1, \ldots, C$. Assume that (2.21) is solvable and that either

(a) $C = 2$ and each A_c has full column-rank, or

(b) $C \geq 2$, each f_c is strongly convex with modulus μ_c and

$$0 < \rho < \min_{c=1,\ldots,C} \frac{2\mu_c}{3(C-1)\,\sigma_{\max}^2(A_c)}\,, \tag{2.27}$$

where $\sigma_{\max}(\cdot)$ denotes the largest singular value of a matrix.

Then, the sequence $\{(x_1^k, \ldots, x_C^k, \lambda^k)\}$ generated by (2.22)-(2.26) converges to $(x_1^\star, \ldots, x_C^\star, \lambda^\star)$, where $(x_1^\star, \ldots, x_C^\star)$ solves (2.21) and $\lambda^\star$ solves the dual problem of (2.21): $\min_\lambda b^\top \lambda + \sum_{c=1}^C f_c^\star(-A_c^\top \lambda)$, where $f_c^\star$ is the convex conjugate of f_c, $c = 1, \ldots, C$.

A proof for case (a) can be found in [83], which generalizes the proofs of [27, 35]. A proof for case (b) can be found in [22]. It is believed that the multi-block ADMM (2.22)-(2.26) still converges for any finite $C > 2$ whenever each function f_c is closed and convex and each matrix A_c has full column rank, i.e., that the generalization of Theorem 2.1 under case (a) still holds. This belief is based on empirical evidence [97], but its proof remains still an open problem. So far, there are only proofs of convergence for similar algorithms that are either slower [98] or that cannot be implemented (at least, straightforwardly) in distributed scenarios [97]. In fact, [98] proves that a modification of the iterates (2.22)-(2.26) converges linearly when each function f_c is strictly convex, differentiable, and has a Lipschitz continuous gradient. That modification consists of changing the stepsize ρ in (2.26) to a smaller number. When that number is sufficiently small, linear convergence can be proved. However, in practice, reducing the stepsize makes the algorithm slower. We note that the distributed algorithms proposed in this thesis are based on the multi-block ADMM, and that we started using them [99] even before there was a proof of convergence [22].

2.2 Distributed algorithms

To the best of our knowledge, the problem we aim to solve, (P), has been considered before only with the following types of variable: global, star-shaped, and mixed (where all non-global components are star-shaped); see Figure 1.7 from Chapter 1 for a visualization of the relation between these types of variables. We next review distributed algorithms that were designed for these types of variables, or for applications that can be written as (P) with such variables.

We mention that [35, §7.2] proposes an algorithm based on the 2-block ADMM for solving (P) with a generic variable. However, it either requires a platform supporting all-to-all communications (equivalently, a central node), or running, at each iteration, a consensus algorithm on each induced subgraph [35, §10.1]. This makes that algorithm not distributed in our sense. Actually, that algorithm becomes distributed only when the variable is star-shaped. We also mention that we found only one distributed algorithm in the literature that can solve (P) when the variable is non-global and non-star-shaped, but still connected. That algorithm, also based on the 2-block ADMM, was proposed in [47] for state estimation of power systems, a problem formulated as (P) with a star-shaped variable. In Chapter 4, we generalize that algorithm for a generic connected variable, and then for a non-connected variable. This means that the algorithm in [47] can also solve (P) in full generality, after proper modifications. Our experimental results, however, show that it always requires more communications to converge to a solution of (P) than the algorithm we propose.

2.2.1 Global class

Among all the classes, the global problem class (G) is the most well studied. For convenience, we recall that (G) is written as

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f_1(x) + f_2(x) + \cdots + f_P(x)\,, \tag{G}$$

where all functions depend on all the components of the variable x. Although several applications can be posed naturally as (G), the application that triggered the interest on the design of distributed algorithms for (G) was the average consensus [7]. Indeed, this was the motivating application in [4], which designed probably the first distributed algorithm for the class (G), an incremental subgradient algorithm. Other important pioneer work includes gradient- and subgradient-based algorithms by Nedić, Ozdaglar, and collaborators [100, 101, 102, 103, 104], whose work was inspired by [39, 105]. At the same time, the first distributed, ADMM-based algorithm was proposed by Schizas, Ribeiro, and Giannakis [25].

We next review four categories of algorithms for (G): incremental, (sub)gradient-based, double-looped, and ADMM-based. Special emphasis will be given to the latter, since the algorithms we propose are also based on ADMM.

Incremental methods. An incremental (sub)gradient method solves problems with the format (G) with the same scheme as the (sub)gradient method (2.9). However, instead of using the (sub)gradient of the entire objective $f_1 + \cdots + f_P$, it only uses the (sub)gradient of one function f_p at a time. More concretely, it consists of

$$x^{k+1} = \left[x^k - \alpha_k \nabla \tilde{f}_{i_k}(x^k)\right]_{X_p}, \tag{2.28}$$

where we decomposed each $f_p = \tilde{f}_p + \mathrm{i}_{X_p}$ into its real-valued part $\tilde{f}_p$ and into its infinity-valued (or constraint-enforcing) part i_{X_p}. To simplify notation, we assumed in (2.28) that each function $\tilde{f}_p$ is differentiable; if not, just replace $\nabla \tilde{f}_p(x^k)$ by any subgradient of $\tilde{f}_p$ at the point x^k. The sequence $\{i_k\}$ takes values in $\{1, 2, \ldots, P\}$ and determines the order of the updates, which can be deterministic or randomized. Surveys about incremental methods, including convergence analysis, can be found in [2, §1.5.2,§6.3.2] and [5]. Roughly, incremental (sub)gradient methods progress faster than their non-incremental counterparts far from the solution, but are slower near the solution [5]. Since they use the (sub)gradient of only one function at each iteration, they can be implemented naturally in a distributed scenario, with a single node performing the update (2.28) at each time instant, in a round-robin fashion. This was done in [4, 106] for a deterministic sequence $\{i_k\}$ and in [107] for a randomized one. The work [5] surveys these methods and, in addition, presents an unified view of incremental (sub)gradient methods, incremental proximal methods, and their combination. In general, incremental methods have slow convergence rates; and, in distributed optimization, they have the disadvantage of making just a single node active at each time instant. The algorithms we propose here, besides exhibiting faster convergence rates, have a higher degree of parallelism, even though not all nodes are active at the same time, i.e., they are not fully parallel.

(Sub)gradient-based. If we apply the (sub)gradient algorithm (2.9) directly to problem (G), the resulting algorithm is non-distributed, since updating x at iteration k requires the (sub)gradients of all the functions at the point x^k. Therefore, using (sub)gradient algorithms to solve (G) in a distributed way requires either reformulating (G) into another equivalent problem, or changing the (sub)gradient algorithm.

The first option was taken in [108], where (G) was rewritten as

$$\begin{array}{ll} \underset{x_1,\ldots,x_P}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} & x_p - x_j \geq 0\,, \quad j \in \mathcal{N}_p\,, \quad p = 1, \ldots, P\,, \end{array} \tag{2.29}$$

where each variable $x_p \in \mathbb{R}^n$, held at node p, is a clone of the original variable $x \in \mathbb{R}^n$. Recall that $\mathcal{N}_p$ denotes the set of neighbors of node p. This reformulation increases the size of the optimization variable in (G) from n to Pn and adds $2E$ constraints (two constraints per each edge $(i,j) \in \mathcal{E}$: $x_i - x_j \geq 0$ and $x_j - x_i \geq 0$, which implies $x_i = x_j$ and, thus, the equivalence between (G) and (2.29)). Note that problem (2.29) has the same format as (2.1), the problem that decomposition methods solve. Indeed, [108] then applies the dual decomposition method described in Subsection 2.1.1 (the generalization of the dual decomposition from equality-constrained problems to inequality-constrained ones is straightforward). If we rewrite the constraints of (2.29) in matrix form, the matrices corresponding to each A_p in (2.1) have a special format: the nonzero entries correspond either to the variable of node p or to the variables of its neighbors. This is makes dual decomposition yield a distributed algorithm. Since in [108] each function is assumed strictly convex, the dual function is differentiable and the gradient algorithm can be applied to solve the dual problem.

The second option (of changing the (sub)gradient algorithm) was taken in a series of works, including [100, 101, 102, 103]. These works study the convergence of a (sub)gradient algorithm coupled with a consensus scheme:

$$x_p^{k+1} = \Big[\sum_{j \in \mathcal{N}_p \cup \{p\}} a_{pj}^k x_j^k - \alpha_k \nabla f_p(x^k) \Big]_{X_p}, \tag{2.30}$$

where we used the same simplifications as in (2.28). In (2.30), $a_{pj}^k > 0$ models the influence node j exerts on node p at iteration k. So, at each iteration k, node p receives the estimates x_j^k from its neighbors $\mathcal{N}_p$, averages them with its own estimate x_p^k, and then performs a projected (sub)gradient step, where the (sub)gradient that is used is the one given by its private function f_p. It is generally assumed that $\sum_j a_{ij}^k = 1$. When all functions f_p are zero and all the sets X_p are the full space $\mathbb{R}^n$, (2.30) becomes the familiar consensus scheme (1.1); and when the network is reduced to a single node, (2.30) becomes the familiar (sub)gradient algorithm (2.9). The linearity of the algorithm (2.30) and the nonexpansiveness property of the projection operator allow an extensive study of the algorithm. In particular, there are proofs of convergence even when the network edges appear and disappear randomly over time. The resulting algorithm, however, inherits the slow convergence properties of the (sub)gradient algorithm, making it communication-inefficient. Variations of (2.30) have also been explored [109, 110, 111, 112]. For example, [109] considers the update $x_p^{k+1} = [\sum_{j \in \mathcal{N}_p \cup \{p\}} a_{pj}^k (x_j^k - \alpha_k \nabla f_j(x^k))]_{X_p}$ and, thus, the (sub)gradient update occurs before transmission; the work [110, 111] applies (2.30) to the dual of a constrained optimization problem.

After noticing that (2.30) is the application of the (sub)gradient algorithm (2.9) to a problem related (but not equivalent) to (G), [113] proposed an improvement based on Nesterov's fast gradient

algorithm (2.10). The problem algorithm (2.30) actually solves is

$$\underset{x=(x_1,\ldots,x_P)}{\text{minimize}} \; f_1(x_1) + \cdots + f_P(x_P) + \frac{\alpha\, k}{2} \sum_{(i,j)\in\mathcal{E}} \|x_i - x_j\|^2 , \tag{2.31}$$

where $x_p \in \mathbb{R}^n$ is the copy of x held by node p, $\alpha > 0$ is a constant, and k is the iteration number. The first term of the objective of (2.31) is the original objective (G), where the variable x was replaced by its copy x_p at the pth function f_p; the second term is a consensus-inducing term, in the sense that different values of the copies between neighbors are penalized. As the iterations go on, the second term becomes more important, forcing the nodes to achieve a consensus on their copies. Note that problems (G) and (2.31) are not equivalent and that this provides an additional reason why algorithms based on (2.30) usually converge slowly. The algorithms we propose in this thesis reformulate (G) into problems that are equivalent to the original one and, thus, do not have this drawback.

Other distributed algorithms that solve (G) with algorithms based on (sub)gradient methods include [114], which hinges on a dual averaging algorithm by Nesterov [115], and [104], which studies a gossip-based version of (2.30), i.e., only two neighboring nodes communicate at each time instant. The work [116] solves (2.5), i.e., the dual of (2.1), using Polyak's heavy-ball method [117].

Double-looped algorithms. A reformulation of (G) similar to (2.29), but that uses half the constraints, is

$$\begin{array}{ll} \underset{x_1,\ldots,x_P}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} & x_i = x_j \,, \quad (i,j) \in \mathcal{E} \,, \end{array} \tag{2.32}$$

where the copies associated to each node are enforced to be the same through the edges of the network. Similarly to what we saw for (2.29), if we apply dual decomposition to (2.32), the result is a distributed algorithm. Unless it is assumed that each function f_p is strictly convex, it is not possible, however, to recover a primal solution after having solved the dual problem. An alternative is to use augmented Lagrangian methods, for example, the method of multipliers (2.15)-(2.16). The augmented term, however, precludes the minimization (2.15) from being carried out in a distributed way. A known workaround is to use an additional loop: an iterative algorithm such as the nonlinear Jacobi (2.12) or the nonlinear Gauss-Seidel (2.13). In fact, this has been done for solving problem (2.1) in [118] (method of multipliers concatenated with the diagonal quadratic approximation) and in [65] (method of multipliers concatenated with the nonlinear Jacobi method). In our work [119], which is not included in this thesis, we applied Nesterov's gradient algorithm (2.10) to both loops, for solving basis pursuit, a problem that can be written as (G), as we will see in the next chapter. Another relevant work is [120], which solves (G) with the method of multipliers

concatenated with a randomized nonlinear Gauss-Seidel method, and uses a reformulation identical to (2.32); see [121] for related work. A difficulty that arises when implementing double-looped algorithms is determining a distributed, robust stopping criterion for the inner loop. Implementing double-looped algorithms in a communication-efficient manner is therefore very challenging.

ADMM-based. If we apply ADMM to the reformulations (2.29), (2.32), and similar ones, we get, in general, distributed algorithms that do not suffer the lack of parallelism of incremental methods, the slow rates of convergence of (sub)gradient-based methods, and the cumbersome two loops of double-looped algorithms.

Algorithm 1 [25]

Initialization: Choose $\rho \in \mathbb{R}$; for all $p \in \mathcal{V}$, set $x_p^0 = \mu_p^0 = \eta_p^0 = 0_n \in \mathbb{R}^n$ and $\tau_p = 1/(\rho(D_p+1))$; set $k = 0$

1: **repeat**

2: **for all** $p \in \mathcal{V}$ [in parallel] **do**

3: Compute $z_p^{k+1} = \tau_p \, \mu_p^k + \frac{1}{D_p+1} \sum_{j \in \mathcal{N}_p^+} x_j^k$ and exchange z_p^{k+1} with neighbors $\mathcal{N}_p$

4: Compute $x_p^{k+1} = \operatorname{prox}_{\tau_p f_p}\left(\frac{1}{D_p+1} \sum_{j \in \mathcal{N}_p^+} z_j^{k+1} - \tau_p \, \eta_p^k \right)$ and exchange x_p^{k+1} with neighbors $\mathcal{N}_p$

5: Update the dual variables:

$$\mu_p^{k+1} = \mu_p^k + \frac{1}{\tau_p}\left(\frac{1}{D_p+1} \sum_{j \in \mathcal{N}_p^+} x_j^{k+1} - z_p^{k+1} \right)$$

$$\eta_p^{k+1} = \eta_p^k + \frac{1}{\tau_p}\left(x_p^{k+1} - \frac{1}{D_p+1} \sum_{j \in \mathcal{N}_p^+} z_j^{k+1} \right)$$

6: $k \leftarrow k+1$

7: **end for**

8: **until** some stopping criterion is met

As said before, the first distributed algorithm based on ADMM was proposed in [25], for solving a particular instance of (G) in the context of estimation. That algorithm, however, can be easily generalized to solve the entire class (G) and is shown as Algorithm 1, explained later. Appendix A shows the derivation of Algorithm 1: we show this derivation for completeness and because, to our best knowledge, there is no reference in the literature where the algorithm is derived to solve the entire class (G). The derivation applies the 2-block ADMM to the following reformulation of (G):

$$\begin{array}{ll} \underset{\bar{x},\bar{z}}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} & x_p = z_j \,, \quad j \in \mathcal{N}_p^+ \,, \quad p = 1, \ldots, P \,, \end{array} \tag{2.33}$$

where each node p has two copies of x: $x_p \in \mathbb{R}^n$ and $z_p \in \mathbb{R}^n$.[4] The optimization variable is $(\bar{x}, \bar{z}) = (x_1, \ldots, x_P, z_1, \ldots, z_P) \in (\mathbb{R}^n)^{2P}$, which makes problem (2.33) have $2P$ times more variables than the original problem (G). In (2.33), we used $\mathcal{N}_p^+ = \mathcal{N}_p \cup \{p\}$ to denote the extended neighborhood of node p, i.e., its set of neighbors $\mathcal{N}_p$ and itself. Problem (2.33) then has $2E + P$ constraints, since there are 2 constraints per edge $(i, j) \in \mathcal{E}$, $x_i = z_j$ and $x_j = z_i$, and each node p constrains $x_p = z_p$. Regarding Algorithm 1, it is fully parallel, as all nodes perform the same tasks at the same time. In the initialization, ρ is the augmented Lagrangian parameter and is assumed fixed and known by all the nodes. At each node p, there is an auxiliary variable τ_p that depends on ρ and on $D_p = |\mathcal{N}_p|$, the number of neighbors of node p. The algorithm consists of three operations, in two of each there is a communication step. Specifically, in step 3 (resp. 4) each node updates z_p (resp. x_p) and exchanges it with its neighbors. Note that updating x_p in step 4 requires the variables z_j from the neighbors $j \in \mathcal{N}_p$. Note also that while the update of z_p is linear and independent of the function f_p, the update of x_p involves the prox operator of a scaled version of f_p. The prox operator of a closed convex function $f : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$\text{prox}_f(x) = \arg\min_y \; f(y) + \frac{1}{2}\|y - x\|^2 . \tag{2.34}$$

This operator, introduced in [122], arises in ADMM-based algorithms, since each ADMM subproblem (cf. (2.18)-(2.19)) is a quadratic problem that can always be written in terms of the prox operator. The prox operator has many properties; see [123] for an extensive list. After performing step 4 in Algorithm 1, node p updates two dual variables, μ_p and η_p, using the new values of x_j and z_j, for $j \in \mathcal{N}_p^+$. The convergence of the algorithm is guaranteed by the convergence results for the 2-block ADMM (2.18)-(2.20).

Algorithm 2 [26]

Initialization: Choose $\rho \in \mathbb{R}$; for all $p \in \mathcal{V}$, set $x_p^0 = \mu_p^0 = 0_n \in \mathbb{R}^n$ and $\tau_p = 1/(2\rho D_p)$; set $k = 0$

1: **repeat**
2: **for all** $p \in \mathcal{V}$ [in parallel] **do**
3: Compute $x_p^{k+1} = \text{prox}_{\tau_p f_p}\Big(\frac{1}{2D_p}\sum_{j\in\mathcal{N}_p}(x_p^k + x_j^k) - \tau_p \mu_p^k\Big)$ and exchange x_p^{k+1} with neighbors $\mathcal{N}_p$
4: Update the dual variable $\mu_p^{k+1} = \mu_p^k + \frac{1}{2\tau_p}\big(x_p^{k+1} - \frac{1}{D_p}\sum_{j\in\mathcal{N}_p} x_j^{k+1}\big)$
5: $k \leftarrow k + 1$
6: **end for**
7: **until** some stopping criterion is met

[4] As pointed out in [25], if there are cliques in the network and, in each clique, only one node is chosen to have the second copy of x, say z_p, problems (G) and (2.33) are still equivalent. In that case, we can even go further and reduce each clique to one node. Since this is a very specific case, we will ignore it and assume that there are no cliques or, if there are, that each node has two copies of x anyway.

The second distributed algorithm based on ADMM was proposed in [26] to solve the average consensus problem, in the context of channel decoding. As [25], it can also be easily generalized to solve the entire class (G). Indeed, that algorithm was used in [9, 42, 43, 124] to solve several other problems in signal processing and machine learning that can be recast as (G). Algorithm 2 shows an adaptation of the algorithm proposed in [26] to solve the entire class (G); its derivation is shown in Appendix A.3. As in Algorithm 1, the derivation applies the 2-block ADMM, but to a different reformulation of (G). Namely, starting with the equivalent problem (2.32), [26] adds E new variables, each one associated to an edge $(i,j) \in \mathcal{E}$ of the network, and rewrites (2.32) as

$$\begin{aligned} \underset{\bar{x},\bar{z}}{\text{minimize}} \quad & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} \quad & x_i = z_{ij}\,, \quad (i,j) \in \mathcal{E} \\ & x_j = z_{ij}\,, \quad (i,j) \in \mathcal{E}\,, \end{aligned} \tag{2.35}$$

where $(\bar{x},\bar{z}) = (x_1,\ldots,z_P,\ldots,z_{ij},\ldots) \in (\mathbb{R}^n)^{P+E}$ is the optimization variable. Problem (2.35) then has $P+E$ more variables (each of size n) than (G) and introduces $2E$ constraints. In Appendix A.3, we show how the application of the 2-block ADMM (2.18)-(2.20) to (2.35) yields Algorithm 2. Note that the application of the same algorithm to a different problem reformulation yields a different, yet more efficient, algorithm. In particular, Algorithm 2 has only one communication step per iteration, whereas Algorithm 1 has two. The communication step occurs in step 3, where each node p updates its estimate x_p by computing the prox operator of $\tau_p f_p$, and then broadcasts the new estimate to its neighbors $\mathcal{N}_p$. Note that $\bar{z}$, the variable that was introduced in (2.35), is absent of Algorithm 2 since, as shown in Appendix A.3, it can be eliminated. The notable work [9] provides a thorough analysis of Algorithms 1 and 2 applied to the average consensus problem. Namely, it establishes linear convergence, proposes a scheme to select the augmented Lagrangian parameter ρ, and studies the factors that influence their convergence. More recently, the work [125, 126], based on the results of [90], establishes the linear convergence of Algorithm 2 whenever each function f_p is strongly convex, differentiable, and its gradient is Lipschitz continuous. It also studies the factors that influence the convergence rate of the algorithm and, based on that study, proposes a scheme to select the augmented Lagrangian parameter ρ. Although that scheme gives a reasonable value for ρ, it does not give the optimal one, i.e., it is usually possible to select a better one by trial-and-error. This partly explains why in the experimental results presented in this thesis we always try several values for ρ, through grid search, and select the one that yields the best result.

The algorithm we propose for (G), rather than using the 2-block ADMM, applies the multi-block ADMM (2.22)-(2.26) directly to reformulation (2.32). Although we cannot establish a convergence rate (since that is still an open problem for the multi-block ADMM), we show through extensive

experimental results that the resulting algorithm outperforms both Algorithms 1 and 2 in terms of the number of communications.

Other splitting methods. We already mentioned that ADMM is an application of the Douglas-Rachford splitting operator to finding the zeros of a given monotone operator. Besides ADMM, other splitting algorithms can be applied and yield distributed optimization algorithms. One example is in [127], which applies a parallel splitting scheme directly to reformulation (2.32), as the algorithm we propose. Our experimental results show, however, that our algorithm outperforms the algorithm proposed in [127] in terms of the number of communications.

We also mention that [128] proposed an asynchronous distributed algorithm for (G) using a randomized version of the Douglas-Rachford operator. Their experimental results show, however, that the resulting algorithm requires more communications to converge than by using the synchronous version. A gossip-based distributed ADMM-based algorithm has been recently proposed in [94] and has been shown to converge with rate $O(1/k)$.

2.2.2 Star-shaped class

Somehow differently from the global class (G), distributed algorithms for the star-shaped class have been motivated mainly by specific applications, and not by the goal of solving an entire class of optimization problems. Such a motivating applications include network utility maximization (NUM), network flow problems, state estimation in power systems, and distributed model predictive control (D-MPC). For this reason, we will organize this section application-wise rather than algorithm-wise. Some applications, most notably D-MPC, arise naturally in scenarios where the variable is non-global and non-star-shaped. In fact, one of the contributions of this thesis is a new framework for D-MPC that uses a generic connected, or even non-connected, variable; this will be addressed in Chapter 4.

Network utility maximization. Consider a network whose edges have a finite transmission capacity and whose nodes are either packet sources, packet sinks, or packet re-transmitters. Each source sends packets to one sink through a specific, pre-chosen route along the network. Associated to each source s there is an utility function U_s (increasing and concave) that depends on x_s, the rate at which source s sends packets. The goal of network utility maximization (NUM), proposed in [129, 130], is to maximize the sum of the utilities of all the sources, while satisfying the link capacity constraints:

$$\begin{array}{ll} \underset{x=(x_1,\ldots,x_S)}{\text{maximize}} & \sum_{s=1}^{S} U_s(x_s) \\ \text{subject to} & Rx = c \\ & x \geq 0\,, \end{array} \tag{2.36}$$

where the lth row of the routing matrix R has ones in entries corresponding to sources that use link l and zeros elsewhere. The lth entry of vector c has the capacity of link l. Note that problem (2.36) is a particular instance of (2.1). It has been used to model congestion control on the Internet [129, 131, 132] and scheduling problems [133]; see also the surveys [52, 134]. If we build an auxiliary network indicating which links are used by each source then, as we will see in Chapter 4, a dual problem of (2.36) can be written as (P) with a star-shaped variable. Actually, if we apply a gradient or a subgradient method directly to that dual problem, we obtain a distributed algorithm because all the induced subgraphs are stars. This is done in [131], which proposes and analyzes synchronous and asynchronous versions of the gradient method for a dual problem of (2.36); curiously, the TCP/IP Vegas protocol, which was designed as an ad hoc congestion control protocol, is interpreted in [132] as a gradient method solving that dual problem. With the goal of improving the speed to convergence, Newton-like methods have also been proposed, for example, a diagonally scaled version of the gradient method with Hessian information in the diagonal [135], and a Newton method where the descent direction is computed approximately [136, 137, 138]. More recently, [139] took advantage of the strong concavity of typical utility functions, which implies that their conjugate is differentiable with Lipschitz continuous gradients, and proposed applying Nesterov's gradient method (2.10) with a choice for a Lipschitz constant that does not require knowing all the utilities at a central location. Then, it proved that the primal estimates converge at rate $O(1/k)$ to their optimal values.

In all these methods, the communication between the source nodes and the used links can be done implicitly, i.e., without sending additional numbers over the network: only by increasing or decreasing the sending rate at which each source sends its packets, and by discarding or not packets that arrive to a given link, an implicit communication can be established. The algorithm we propose for (P), in contrast, requires explicit communication between the source nodes and the links; however, it exhibits faster convergence to the equilibrium.

Distributed model predictive control. Model predictive control (MPC), also known as receding horizon control, is an efficient control scheme for discrete-time systems. Dating back to the early sixties [140, 141], MPC became very popular in the petro-chemical industry in the early eighties, as surveyed in [142]. The interest in applying MPC to distributed systems, however, arose later, in the nineties [143, 144]. The setting is a network of systems, each of which has associated a state, a control input, or both. Each system interacts with neighboring systems in two ways: through system dynamics and through communication. Interaction through system dynamics means that the state of each system is influenced by the states and control inputs of neighboring systems; sometimes, neighboring systems also have coupled goals (or efficiency measures). Interaction through communication refers to the ability that each system has to exchange messages with neighboring

systems and, thus, it corresponds to what we call communication network. MPC in this scenario is usually referred to as distributed MPC (D-MPC). The goal in each instance of D-MPC is to make the systems cooperate to find an optimal set of inputs, i.e., control inputs that drive the state of each system from an initial (measured) state to a predefined goal, while minimizing the energy to do so. This can be cast as an optimization problem with the format of (P), as we will see in Chapter 4. To the best of our knowledge, all prior work on D-MPC has assumed that interaction through dynamics coincides with interaction through communication. That is, if two systems have coupled dynamics, i.e., the state of one of them is influenced by the state or input of the other, then they necessarily communicate directly. According the classification scheme introduced in Chapter 1, the variable in this case is star-shaped. In this thesis, we introduce a new framework for D-MPC, where coupled systems do not necessarily need to communicate directly. We also present potential applications for this new framework.

Early work on D-MPC has focused on studying stability and performance of heuristics whose solutions are not guaranteed to be optimal. For example, [145] proposes a one-step scheme where each system solves a local optimization problem that incorporates state predictions from its neighbors; this is preceded by a communication step, where state predictions are exchanged between neighboring nodes. For related methods, see [19, 146, 147].

D-MPC has also been tackled with optimization-based algorithms, not always completely distributed, that find exact solutions. For example, [144] proposes an augmented Lagrangian method where the augmented term is linearized, a method now known as split inexact Uzawa method in the image processing community [148, 149]. The resulting algorithm is not distributed, since it requires a central node. Distributed algorithms for D-MPC include dual decomposition with the subgradient method [150] (as described in Subsection 2.1.1), distributed interior-point methods [151], and more recently, fast gradient methods [46] and ADMM [46, 152]. In particular, [46, 152] apply the ADMM method proposed in [35], which becomes distributed whenever the variable is star-shaped. This is, in fact, the case since, as mentioned before, all prior work on D-MPC assumes that interaction through dynamics coincides with interaction through communication.

The algorithms we propose for D-MPC require less communications to achieve convergence than all these algorithms. In addition, they solve D-MPC in scenarios that have never been considered before: problems with a connected variable that is neither global nor star-shaped, and problems with a non-connected variable. Both cases model systems that are coupled through their dynamics, but cannot communicate directly.

Network flows. Beyond NUM and D-MPC, there is an extensive literature on network flow problems, some of which can be formulated as (P) as well. In a typical network flow problem, each component of the optimization variable is associated to an edge of the network, and the function

at each node depends on the variables associated to its incident edges. Hence, the variable is star-shaped; actually, each induced subgraph is very simple: it consists of two nodes and an edge connecting them. The first optimization algorithm solving a network flow problem was Dantzig's simplex method [153, Ch.19-20]. Extensive information about network flows, including specialized algorithms (most of them centralized), can be found in the surveys [154, 155] and in the books [20, 156].

Regarding distributed algorithms for network flows, dual decomposition methods generally yield distributed algorithms. For example, by assuming strict convexity on the cost functions, [157] computes the dual of a network flow problem and proposes to solve it with an asynchronous Gauss-Seidel method. The application of a subgradient method to a similar problem is analyzed in [158]. More recently, [159] proposed a double-looped algorithm, where the outer loop uses the proximal minimization algorithm (to overcome the lack of strict convexity of the primal objective) and the inner loop uses the gradient method. The work [160, 161, 162, 136] proposes a distributed algorithm for network flows based on Newton's method, where the Newton direction is computed approximately. Although the resulting method requires the cost functions to be strongly convex and twice differentiable, it is proven to converge superlinearly to a neighborhood of the problem's solution. This contrasts with the algorithm we propose for (G), which only requires the cost functions to be convex, possibly non-differentiable. Our algorithm thus requires assumptions much less restrictive that the assumptions of methods based on dual decomposition or on Newton's algorithm. Additionally, as will be shown in Chapter 4, the algorithm we propose requires less communications to converge than the algorithm in [160, 161, 162].

2.2.3 Mixed class

The mixed problem class (M), reproduced here for convenience,

$$\underset{x=(y,z)\in\mathbb{R}^n}{\text{minimize}} \quad f_1(y, z_{S_1}) + f_2(y, z_{S_2}) + \cdots + f_P(y, z_{S_P}), \tag{M}$$

has rarely appeared in literature, despite its generality and possible applications. One instance of (M) has appeared in [33] (see also [52, §IV-B]) as a dual of a NUM problem with coupled objectives. We will look at this problem with more detail in Chapter 4. Such a problem can model cooperative systems, e.g., systems where the rate allocated to one source depends on the rate allocated to the cluster that source belongs to, or competitive systems, e.g., wireless power control or digital subscribed line (DSL) spectrum management where signal-to-interference ratios (SIR) are dependent on transmit powers of other users. The method proposed in [33] is distributed and consists of solving that dual problem (which has the format of (M)) with a gradient method.

Actually, the application of the gradient method to (M) in [33] yields a distributed algorithm, because the non-global components, z in (M), are star-shaped.

We will also use the framework of (M) to solve in a distributed way a compressed sensing problem with a data partitioning that has never been considered before. More concretely, basis pursuit denoising (BPDN), and a related problem that we call *reversed lasso* have been solved in a distributed way with a row partition [163, 42, 124] and with a column partition [163], respectively. The reverse cases, i.e., BPDN with a column partition and reversed lasso with a row partition, have never been solved before. We will show in Chapter 4 that reversed lasso with a row partition can be formulated as (M), and therefore can be solved with the algorithms we propose here.

Chapter 3

Global Class

This chapter addresses the global class (G) and is based on the publications [164, 165, 166, 83, 163]. The chapter is organized into four sections. In Section 3.1, the problem is formally stated and the assumptions are clearly identified. In Section 3.2, we describe some applications that can be written as (G). Special emphasis is given to Subsection 3.2.2, since it contains novel contributions, such as writing some distributed compressed sensing problems as (G). Then, in Section 3.3, we propose our algorithm for the global class (G) and analyze it. Finally, in Section 3.4, we show the performance of the proposed algorithm against prior algorithms by running extensive simulations. These show that, while solving the entire class (G), our algorithm is as efficient as algorithms that were specifically designed for particular applications and, often, it is even better.

3.1 Problem statement

The global problem class (G) consists of minimizing the sum of P functions where each function depends on all the components of x. For convenience, let us rewrite (G) here:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f_1(x) + f_2(x) + \cdots + f_P(x)\,. \tag{G}$$

We make the following assumptions:

Assumption 3.1. *Each function $f_p : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed and convex over $\mathbb{R}^n$ and is not identically $+\infty$.*

Assumption 3.2. *Problem* (G) *is solvable, i.e., it has at least one solution $x^\star$.*

In Assumption 3.1 we use the concept of an extended real-valued function f, which can take infinite values and is defined over all $\mathbb{R}^n$. Such a function is closed and convex if its epigraph

$\operatorname{epi} f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \,:\, f(x) \leq r\}$ is closed and convex, respectively [54, §1.2], [30, §B.1]. Alternatively, a function is closed if it is lower semicontinuous or if all its sublevel sets are closed [54, Prop.1.2.2], [30, Prop.1.2.2]. Considering extended real-valued functions simplifies the notation without losing generality: as explained before, each node p can constrain variable x to belong to a given set S, i.e., $x \in S$, through an indicator function i_S, defined as $\mathrm{i}_S(x) = 0$ if $x \in S$, and $\mathrm{i}_S(x) = +\infty$ if $x \notin S$.

We associate problem (G) to a communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $P = |\mathcal{V}|$ nodes and $E = |\mathcal{E}|$ edges: the pth node of the network is the only node who knows function f_p or, in other words, function f_p is private to node p. Regarding the network, we assume:

Assumption 3.3. *The network is connected and its topology does not vary with time.*

Assumption 3.4. *A coloring scheme $\mathcal{C}$ of the network is available; each node knows its own color and the color of its neighbors.*

The concept of network coloring was explained in Section 1.1: it is an assignment of numbers, called colors, to the nodes such that no neighboring nodes have the same color. Formally, each node is assigned a color in $\mathcal{C} = \{1, \ldots, C\}$, where $C := |\mathcal{C}|$ is the total number of colors, and $\mathcal{C}(p)$ denotes the color of node p. The coloring scheme $\mathcal{C}$ is called *proper* (or valid) if $\mathcal{C}(i) \neq \mathcal{C}(j)$, for all $(i, j) \in \mathcal{E}$. Our goal is to *design a distributed algorithm that solves* (G) *while keeping the function f_p private to node p.* Recall that a distributed algorithm is one that uses no central or special node and no all-to-all communications.

Discussion of the assumptions. Compared to prior algorithms for the global class (G), the problem Assumptions 3.1 and 3.2 are very general, while the network Assumptions 3.3 and 3.4 are more restrictive, yet realistic in some scenarios. In fact, what Assumption 3.1 asks is the problem to be convex, a minimal requirement to guarantee that we can find a global minimizer of (G). In Assumption 3.2, we require that the problem is well-posed by having at least one solution.

Regarding the network assumptions, assuming a fixed network topology as in Assumption 3.3 is a common first step in distributed optimization. Some algorithms, however, are proven to converge under intermittent link failures, e.g., [108, 102, 120]. These algorithms, in turn, require more assumptions on the functions in (G). In fact, there seems to be a curious tradeoff between the problem assumptions and the network assumptions: the algorithms that relax the network assumptions usually require more restrictive problem assumptions, and vice-versa. Regarding Assumption 3.4, this assumption is new in the context of distributed optimization and will underlie the construction of our algorithm. Recall that finding the minimum number of colors a network can be colored with is NP-hard [167], except for bipartite networks. The minimum number of colors required to color a network $\mathcal{G}$ is called the *chromatic number* and is represented with $\chi(\mathcal{G})$. Assuming that $\chi(\mathcal{G})$ is known

and that $\chi(\mathcal{G}) > 2$ (i.e., the network is not bipartite), coloring $\mathcal{G}$ with $\chi(\mathcal{G})$ colors is NP-hard as well. Given its importance in wireless networks, there are several approximation algorithms to compute coloring schemes of networks, some of which are distributed [168, 169, 170, 171]. For example, [168] proposes a coloring scheme that uses $O(D_{\max})$ colors while requiring $O(D_{\max}/\log^2(D_{\max})+\log^\star(P))$ iterations to compute them, where $D_{\max} := \max\{D_p : p \in \mathcal{V}\}$ is the maximum degree of a node in the network. Another coloring scheme using less iterations, but more colors, more specifically, $O(\log^\star(P))$ iterations and $O(D^2_{\max})$ colors, is proposed in [171]. In this thesis, we assume that a coloring scheme with C colors is given and we will ignore how it was obtained. Consequently, the additional number of communications to obtain the scheme will also be ignored in the comparison with other algorithms. Although all the other algorithms use no coloring scheme (all nodes work in parallel), the comparison is fair for two reasons: first, if an algorithm is run several times on the same network, for example, for different data, coloring the network just needs to be done once, before the first instantiation; after running the algorithm several times, the coloring cost becomes diluted. The second, and perhaps more important, reason is that in networks where the transmission medium is shared, for example, in wireless networks or even in Ethernet cables, the nodes cannot communicate in parallel without using a medium access control (MAC) protocol [172, Ch.5-6],[21]. For example, in wireless networks, one node cannot receive two different messages from its neighbors at the same time and at the same frequency (unless it uses more than one receive antenna [173]). This creates the hidden and the exposed node problems [21, §6.2.2], which are prevented by the use of MAC protocols. For data-intensive algorithms, such as the ones considered in this thesis, schedule-based MAC protocols are the most energy-efficient [21, §6.7]. Time division multiple access (TDMA) is such a protocol which, in addition, is also based on network coloring. The particular coloring scheme used by TDMA can also be used for the algorithms we propose; thus, our algorithms integrate naturally with TDMA. Prior algorithms for distributed optimization, in contrast, assume no particular MAC protocol. The second part of Assumption 3.4 will be discussed when we introduce our algorithm; briefly, it allows discarding a centralized entity controlling all the nodes that have the same color (recall that they are not neighbors) and, because of that, it is essential in making our algorithm distributed.

3.2 Applications

There are many engineering problems that can be written as (G). Here, we will focus on problems that arise in networks and, consequently, that can be solved via distributed algorithms. We address two types of problems: inference problems, which include average consensus and support vector

machines (SVMs), and sparse solutions of linear systems, which include several compressed sensing problems.

3.2.1 Inference problems

Average consensus. Consider the scalar version of the inference problem described in Chapter 1: a sensor network composed of P nodes is deployed to estimate a parameter $\bar{\theta} \in \mathbb{R}$. The estimation uses measurements from all the sensors, which are assumed noisy. Let θ_p denote the measurement at node p. When the noise is independent across nodes, Gaussian, with zero mean, and identity covariance matrix, the maximum log-likelihood estimation of $\bar{\theta}$ is given by *average consensus* [7]:

$$\underset{x}{\text{minimize}} \quad \tfrac{1}{2}(x-\theta_1)^2 + \tfrac{1}{2}(x-\theta_2)^2 + \cdots + \tfrac{1}{2}(x-\theta_P)^2 \,. \tag{3.1}$$

Average consensus has been widely studied in the literature, and many distributed algorithms have been proposed to solve it [174, 8, 175, 176, 12, 11, 177, 10]. Curiously, most of these algorithms are not optimization-based, in the sense that they do not view the consensus problem as the distributed optimization problem (3.1); rather, they simply solve it with a linear update scheme, such as (1.1). Work that has addressed average consensus by devising a distributed optimization algorithm for (3.1) includes [4, 106, 25, 26, 9]. In particular, [9] analyzes Algorithms 1 and 2, described in Chapter 2, applied to consensus. Despite the vast quantity of algorithms for the average consensus, we will see that the algorithm we propose for the global class (G) has a performance similar to the most efficient algorithms, if not better.

Support vector machine (SVM). Another important inference problem is a support vector machine (SVM) [17, Ch.7]. Training an SVM consists of finding the parameters $(s, r) \in \mathbb{R}^{n-1} \times \mathbb{R}$ of an hyperplane $\{x \in \mathbb{R}^{n-1} : s^\top x = r\}$ that best separates two classes of points. These points are given as $(x_k, y_k) \in \mathbb{R}^{n-1} \times \mathbb{R}$, where $y_k \in \{-1, 1\}$ indicates the class of the point x_k. Finding these parameters usually involves solving an optimization problem, for example,

$$\begin{array}{ll} \underset{s,r,\xi}{\text{minimize}} & \tfrac{1}{2}\|s\|^2 + \beta\, 1_K^\top \xi \\ \text{subject to} & y_k(s^\top x_k - r) \geq 1 - \xi_k \,, \quad k = 1, \ldots, K \\ & \xi \geq 0 \,, \end{array} \tag{3.2}$$

where K is the total number of points, $\beta > 0$ is a tradeoff parameter, and $\xi \in \mathbb{R}^K$ is a vector of slack variables. In a network scenario, we assume each node knows m_p points, but all the nodes

cooperate to solve the global problem (3.2). This problem can be written as (G) by setting

$$\begin{aligned} f_p(s,r) = \; & \inf_{\bar{\xi}_p} \quad \tfrac{1}{2P}\|s\|^2 + \beta\, 1_{m_p}^\top \bar{\xi}_p \\ & \text{s.t.} \quad Y_p(X_p s - r 1_{m_p}) \geq 1_{m_p} - \bar{\xi}_p \\ & \qquad \bar{\xi}_p \geq 0\,, \end{aligned} \tag{3.3}$$

where Y_p is a diagonal matrix with the labels y_k of the points of node p in the diagonal, and X_p is an $m_p \times n$ matrix with each row containing $x_k^\top$, ordered the same way as Y_p. The variable x in (G) corresponds to (s, r), since the slack variables $\bar{\xi}_p$ are internal to each node. This distributed SVM problem has been solved in [43] with Algorithm 2. See [178] for a related message-passing method.

3.2.2 Sparse solutions of linear systems

Another application we consider is finding sparse solutions of distributed linear systems. This is mainly motivated by the recent field of compressed sensing [13, 14], which establishes a new paradigm for signal acquisition and sampling. Surveys on the topic include [179, 180, 181, 182]. While acquisition of signals in compressed sensing is usually simple, reconstructing them afterwards is more complicated and it involves solving an optimization problem. In noiseless scenarios, the most common problem is *basis pursuit* (BP) [15]:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = b\,, \end{aligned} \tag{3.4}$$

where $x \in \mathbb{R}^n$ is the variable and $\|x\|_1$ denotes the ℓ_1-norm of x, defined as $\|x\|_1 := \sum_{i=1}^n |x_i|$. The matrix $A \in \mathbb{R}^{m\times n}$ and the vector $b \in \mathbb{R}^m$ are associated to the acquisition process, and we assume they are given. The linear system $Ax = b$ is underdetermined, i.e., $m < n$, and the matrix A is usually assumed full rank, so that the linear system is feasible for any b. This is common in compressed sensing, since the entries of A are usually drawn randomly and in an independent way. In noisy scenarios other problems are used. An example is *basis pursuit denoising* (BPDN) [15]:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \tfrac{1}{2}\|Ax - b\|^2 + \beta\|x\|_1\,, \tag{3.5}$$

where $\beta > 0$ is a tradeoff parameter and $\|z\|$ denotes the ℓ_2-norm of $z \in \mathbb{R}^q$, i.e., $\|z\| = \sqrt{\sum_{i=1}^q z_i^2}$. There is also a problem that we will call *reversed lasso* [183, 184, 185]:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & \|x\|_1 \\ \text{subject to} \quad & \|Ax - b\| \leq \sigma\,, \end{aligned} \tag{3.6}$$

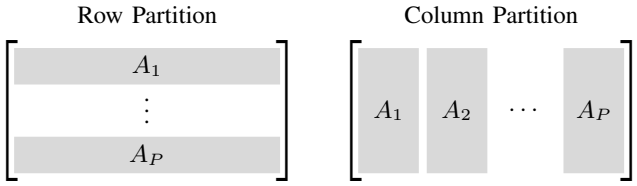


FIGURE 3.1: Row partition and column partition of A into P blocks. A block in the row (resp. column) partition is a set of rows (resp. columns).

where $\sigma > 0$ is a known bound on the noise magnitude, and a problem called the *least absolute shrinkage and selection operator* (lasso) [16]:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\|Ax-b\|^2 \\ \text{subject to} & \|x\|_1 \leq \gamma\,, \end{array} \tag{3.7}$$

where $\gamma > 0$ is a known parameter. Problems (3.4)-(3.7) provide heuristics to find sparse solutions of the linear system $Ax = b$. In fact, it was established in [186] that finding a sparsest solution of that linear system is NP-hard. Such a problem would be written as (3.4) with the cost function replaced by the cardinality of the vector x, $\text{card}(x)$. We thus see that (3.4) approximates the non-continuous, non-convex function $\text{card}(x)$ by the convex function $\|x\|_1$, resulting in a convex (and hence easier) problem. The same approximation motivates problems (3.5)-(3.7), but in the scenario where b may not be expressed exactly as a linear combination of the columns of A. The theory of compressed sensing establishes conditions on the matrix A under which approximating $\text{card}(x)$ by $\|x\|_1$ in, for example, BP (3.4) yields an exact approximation: this means that the NP-hard problem obtained from (3.4) by replacing $\|x\|_1$ with $\text{card}(x)$ has the same solution as the convex problem BP (3.4). Surprisingly, some types of random matrices satisfy those conditions with overwhelming probability. For more details see, for example, [187, 188, 189, 190].

Problems (3.5), (3.6), and (3.7) are all related through duality and, therefore, are equivalent in some sense, provided their parameters β, σ, and γ are chosen appropriately. Among these problems, (3.6) is the one to which compressed sensing results apply directly [189, 185], in spite of never have been coined a specific name. Apparently, sometimes it is also called lasso [180], but we avoid that name to prevent confusion with the original lasso (3.7). Instead, we will call it *reversed lasso* since, compared to lasso, its objective and constraints are reversed. Note that when $\sigma = 0$, the reversed lasso becomes BP (3.4). We are interested in solving the compressed sensing problems (3.4)-(3.7) in the distributed scenarios described next.

Distributed scenarios: row and column partition. We consider two different scenarios for splitting the data in matrix A and vector b among the nodes of a network with P nodes.

These are called *row partition* and *column partition*, and are visualized in Figure 3.1. In the row (resp. column) partition, each node stores a block of rows (resp. columns) of A. While in the row partition vector b is partitioned similarly to A, with each node storing the corresponding subblock, in the column partition we assume all nodes know the full vector b. More specifically, node p knows $(A_p, b_p) \in \mathbb{R}^{m_p \times n} \times \mathbb{R}^{m_p}$ in the row partition and knows $(A_p, b) \in \mathbb{R}^{m \times n_p} \times \mathbb{R}^m$ in the column partition. Naturally, we have $m = \sum_{p=1}^{P} m_p$ and $n = \sum_{p=1}^{P} n_p$.

The row partition scenario arises naturally when applying compressed sensing in a sensor network. For instance, suppose the nodes of the network are interested in estimating a high-dimensional but sparse vector $x \in \mathbb{R}^n$, for example, an ultra-wide band but spectrally sparse radio signal. Each node in the network is equipped with a low bandwidth antenna and, hence, any signal acquisition has to be done at a rate far below the Nyquist rate. By using a random demodulator [180, 191], compressed sensing can be applied, and each row of the linear system $Ax = b$ represents one measurement (performed at a low acquisition rate). Therefore, if node p takes m_p linear measurements of x, we have exactly the row partition scenario. This setting appeared in [192, 193], where several applications are described. It is assumed there, however, that the signal reconstruction, i.e., solving one of the problems (3.5)-(3.7) is done in a centralized way, in a fusion center. The algorithms we propose in this thesis allow reconstructing the signal on the network, without using any fusion center. Furthermore, all nodes will know the signal when the algorithm finishes. Other applications include distributed target localization [194] and distributed field reconstruction [195].

One application of the column partition is described in [196], in the context of forward modeling in geological applications. The goal is to find the Green's function, represented by a vector x, of a model of the earth's surface. The authors of [196] propose deploying a set of sources and a set of receivers over some geographical area and have all the sources emit a signal simultaneously. The receivers capture a linear superposition of all the emitted signals. The proposed way to find x is by solving BP (3.4), where a set of columns of A is associated to a source. This is clearly our column partition scenario. The distance between all the devices in this application makes a distributed solution convenient, such as the ones provided by our algorithms.

We will see next how BP, BPDN, and lasso with a row partition are naturally recast as (G). Then, we will consider the less trivial case of a column partition, for all the problems (3.4)-(3.7). The only problem that will be missing is reversed lasso with a row partition. However, in Chapter 4, we will be able to recast it as (P), not with a global variable, but with a mixed one.

Row partition: BP, BPDN, and lasso. Consider a row partition as shown in Figure 3.1. Then, BP (3.4) can be written as (G) by setting as the function of node p

$$f_p(x) = \frac{1}{P}\|x\|_1 + \mathrm{i}_{A_p x = b_p}(x)\,, \tag{3.8}$$

where $\mathrm{i}_{A_p x = b_p}(x)$ is the indicator function of the set $\{x : A_p x = b_p\}$. Similarly, BPDN (3.5) can be written as (G) by setting as the function of node p

$$f_p(x) = \frac{1}{2}\|A_p x - b_p\|^2 + \frac{\beta}{P}\|x\|_1 \,. \tag{3.9}$$

Note that the parameter β and the number of nodes P is assumed to be known by all nodes. Lasso (3.7) can also be written easily as (G) by setting

$$f_p(x) = \frac{1}{2}\|A_p x - b_p\|^2 + \mathrm{i}_{\|x\|_1 \leq \gamma}(x) \tag{3.10}$$

as the function of node p. Here, the parameter γ is also assumed to be known at all nodes. Each function f_p in (3.8)-(3.10) contains data that is known only by node p: namely, the pair (A_p, b_p). All these functions are closed and convex. Furthermore, the extended real-valued function (3.8) (resp. (3.10)) is not identically $+\infty$ whenever A_p has full rank (resp. γ is positive). BPDN with a row partition was solved in [42, 124] with Algorithm 2, viewing it as an instance of (G) with (3.9).

Column partition: duality and regularization. We now turn into a column partition and recast all the problems (3.4)-(3.7) as (G). We will need duality to do this. However, plain duality will not be enough to recover primal solutions from dual solutions, since the problems we dualize have cost functions that are not strictly convex. We will thus use regularization and, in the case of BP, the concept of *exact regularization*. We introduce this concept together with a result by Friedlander and Tseng [197]. Consider the following conic program

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^\top x \\ \text{subject to} & x \in \mathcal{K} \\ & Ax = b \,, \end{array} \tag{3.11}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$ are given, and $\mathcal{K} \subseteq \mathbb{R}^n$ is a nonempty, closed, convex cone. Problem (3.11) is assumed to have a nonempty solution set $\mathcal{S}$. Consider now a regularization function $\phi : \mathbb{R}^n \to \mathbb{R}$ such that all sublevel sets of $\mathcal{S}$, i.e., $\{x \in \mathcal{S} : \phi(x) \leq \alpha\}$, are bounded for all α. A result in [197], more specifically in corollary 2.3 of [197], states that when $\mathcal{K}$ is polyhedral, i.e., $\mathcal{K} = \{x : v_i^\top x \leq 0 \,, i = 1, \ldots, q\}$ for some set of q vectors $v_i \in \mathbb{R}^n$, then the regularization of (3.11) with ϕ is exact. This means that there exists a $\bar{\delta} > 0$ such that the set of solutions of the regularized problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^\top x + \frac{\delta}{2}\phi(x) \\ \text{subject to} & x \in \mathcal{K} \\ & Ax = b \end{array} \tag{3.12}$$

is contained in the set of solutions $\mathcal{S}$ of (3.11), for all $0 \leq \delta \leq \bar{\delta}$ [197, Cor.2.3]. As mentioned in [197], this is a generalization of exact regularization results for linear programs [198, 199]. Experimental results in [197] suggest that the above result is true even when $\mathcal{K}$ is not polyhedral, namely, when $\mathcal{K}$ is the Lorenz cone $\{(x,t) \,:\, \|x\| \leq t\}$ (also known as the ice-cream cone and as the second-order cone). That cone will actually arise in some of our problems for which, inspired by the results in [197], we will perform the above regularization. For BP, we will use the exact regularization result, since BP is equivalent to a linear program, which is the simplest instance of a conic program. Regarding the choice of δ, we are unaware of any method that finds $\bar{\delta}$ without first solving the unregularized problem (3.11). Therefore, we will choose δ based on trial-and-error. According to our experiments, $\delta \in [10^{-3}, 10^{-1}]$ allows computing an optimal solution with reasonable accuracy most of the times. In [197, §7], it is reported that $\delta = 10^{-4}$ yielded an optimal solution in 85% of their experiments.

Column partition: BP. We start with BP (3.4). Consider the regularization function $\phi(x) = (1/4)\|x\|^2$ and the regularized problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \frac{\delta}{2}\|x\|^2 \\ \text{subject to} & Ax = b\,. \end{array} \tag{3.13}$$

Then, by the previous discussion, the following theorem follows.

Theorem 3.5.

Problem (3.13) *is an exact regularization of BP* (3.4)*, i.e., there exists a* $\bar{\delta} > 0$ *such that the solution of* (3.13)*, with* $0 \leq \delta \leq \bar{\delta}$*, is always a solution of BP.*

Proof.

We use the exact regularization results of [197, 198, 199], as explained before. First, we recast BP as a problem with the same format as (3.11):

$$\begin{array}{ll} \underset{x,t}{\text{minimize}} & 1_n^\top t \\ \text{subject to} & Ax = b \\ & -t \leq x \leq t\,, \end{array} \tag{3.14}$$

where $t \in \mathbb{R}^n$ is an epigraph variable and $1_n \in \mathbb{R}^n$ is the vector of ones. Problem (3.14) has the same format as (3.11) by making the correspondence $c = (0_n, 1_n)$ and $\mathcal{K} = \{(x,t) \,:\, -t \leq x \leq t\}$, which is a polyhedral cone that is nonempty, closed, and convex. The corresponding regularized

problem (3.12) with $\phi(z) = (1/4)\|z\|^2$ is

$$\begin{array}{lll} \underset{x,t}{\text{minimize}} & 1_n^\top t + \frac{\delta}{4}\|x\|^2 + \frac{\delta}{4}\|t\|^2 & \iff \\ \text{subject to} & Ax = b & \\ & -t \le x \le t & \end{array} \quad \begin{array}{lll} \underset{x}{\text{minimize}} & \frac{\delta}{4}\|x\|^2 \; + \; \inf\limits_{t} & 1_n^\top t + \frac{\delta}{4}\|t\|^2 \\ \text{subject to} & Ax = b \quad \text{s.t.} & -t \le x \le t \end{array}$$

$$\iff \quad \begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \frac{\delta}{2}\|x\|^2 \\ \text{subject to} & Ax = b\,, \end{array}$$

which is (3.13). In the last equivalence, we used the fact that, for a fixed x,

$$\begin{array}{lll} \inf\limits_{t} & 1_n^\top t + \frac{\delta}{4}\|t\|^2 & = \quad \|x\|_1 + \dfrac{\delta}{4}\|x\|^2\,. \\ \text{s.t.} & -t \le x \le t\,. & \end{array} \tag{3.15}$$

□

Now, let $\lambda \in \mathbb{R}^m$ be a dual variable associated to the constraint of (3.13). The dual problem is

$$\underset{\lambda}{\text{maximize}} \;\; b^\top\lambda + \inf_x \left[\|x\|_1 + \frac{\delta}{2}\|x\|^2 - \lambda^\top A x\right] \tag{3.16}$$

$$\iff \quad \underset{\lambda}{\text{maximize}} \;\; b^\top\lambda + \sum_{p=1}^{P} \inf_{x_p} \left[\|x_p\|_1 + \frac{\delta}{2}\|x_p\|^2 - \lambda^\top A_p x_p\right] \tag{3.17}$$

$$\iff \quad \underset{\lambda}{\text{minimize}} \;\; -b^\top\lambda + \sum_{p=1}^{P} \sup_{x_p} \left[(A_p^\top\lambda)^\top x_p - \left(\|x_p\|_1 + \frac{\delta}{2}\|x_p\|^2\right)\right] \tag{3.18}$$

$$\iff \quad \underset{\lambda}{\text{minimize}} \;\; \sum_{p=1}^{P} \left(h_p^\star(A_p^\top\lambda) - \frac{1}{P}b^\top\lambda\right), \tag{3.19}$$

which has the format of (G) with the function at node p given by $f_p(\lambda) = h_p^\star(A_p^\top\lambda) - (1/P)b^\top\lambda$. From (3.16) to (3.17), we used the column partition and the fact that all terms inside the infimum decouple. From (3.17) to (3.18), we switched from a maximization problem to a minimization one. And, in (3.19), we defined $h_p^\star$ as being the convex conjugate of the function $h_p(x_p) = \|x_p\|_1 + (\delta/2)\|x_p\|^2$, for each p. Note that the global variable is the dual variable λ; also, after an optimal value $\lambda^\star$ has been found (or better, agreed by all the nodes), the pth component of the corresponding primal solution $x^\star$ is available at the pth node. Each component is given by soft-thresholding:

$$x_i = \begin{cases} \frac{1}{\delta}\Big((A_p^\top\lambda)_i - 1\Big)_i & , \; (A_p^\top\lambda)_i > 1 \\ \frac{1}{\delta}\Big((A_p^\top\lambda)_i + 1\Big)_i & , \; (A_p^\top\lambda)_i < -1 \\ 0 & -1 \le (A_p^\top\lambda)_i \le 1\,, \end{cases} \tag{3.20}$$

for i belonging to the indices of the columns of A_p; see Appendix B for the derivation of (3.20).

Column partition: BPDN. We now move to BPDN with a column partition. We will also use regularization but, this time, we will not have an exact regularization result. To regularize BPDN (3.5) the same way as BP, we first rewrite it with the format of (3.11):

$$\begin{array}{ll} \underset{x,t,u,v}{\text{minimize}} & \frac{1}{2}v + \beta\, 1_n^\top t \\ \text{subject to} & \|u\|^2 \le v \\ & -t \le x \le t \\ & u = Ax - b\,, \end{array} \tag{3.21}$$

where $t \in \mathbb{R}^n$ and $v \in \mathbb{R}$ are epigraph variables, and $u \in \mathbb{R}^m$ is an auxiliary variable. Problem (3.21) has the same structure as (3.11), since its objective is linear, the last two constraints are also linear, and the cone $\mathcal{K}$ is the Cartesian product $\mathcal{K} = \mathcal{K}_{x,t} \times \mathcal{K}_{u,v}$, where $\mathcal{K}_{x,t} = \{(x,t) \,:\, -t \le x \le t\}$ is polyhedral, but $\mathcal{K}_{u,v} = \{(u,v) \,:\, \|u\|^2 \le v\}$ is not. Using the function $\phi(z) = (1/4)\|z\|^2$ to regularize (3.21), we obtain

$$\begin{array}{ll} \underset{x,t,u,v}{\text{minimize}} & \frac{1}{2}v + \beta 1_n^\top t + \frac{\delta}{4}\Big(\|x\|^2 + \|t\|^2 + \|u\|^2 + v^2\Big) \\ \text{subject to} & \|u\|^2 \le v \\ & -t \le x \le t \\ & u = Ax - b\,, \end{array} \tag{3.22}$$

$$\iff \quad \underset{x}{\text{minimize}} \quad \tfrac{\delta}{4}\|x\|^2 + \tfrac{\delta}{4}\|Ax-b\|^2 \;+\; \inf_{t} \;\; \beta 1_n^\top t + \tfrac{\delta}{4}\|t\|^2 \;\text{ s.t. }\; -t \le x \le t \;+\; \inf_{v} \;\; \tfrac{1}{2}v + \tfrac{\delta}{4}v^2 \;\text{ s.t. }\; \|Ax-b\|^2 \le v \tag{3.23}$$

$$\iff \quad \underset{x}{\text{minimize}} \quad (\tfrac{1}{2}+\tfrac{\delta}{4})\|Ax-b\|^2 + \beta\|x\|_1 + \tfrac{\delta}{2}\|x\|^2 + \tfrac{\delta}{4}\|Ax-b\|^4\,. \tag{3.24}$$

From (3.22) to (3.23), we replaced u by $Ax - b$. From (3.23) to (3.24), we used (3.15) with the weight β and eliminated the epigraph variable v.

Although our next steps are also valid for (3.24), we will discard the last term of its objective, for simplicity. That is, we will solve instead:

$$\underset{x}{\text{minimize}} \quad (\tfrac{1}{2}+\tfrac{\delta}{4})\|Ax-b\|^2 + \beta\|x\|_1 + \tfrac{\delta}{2}\|x\|^2\,. \tag{3.25}$$

We now introduce an auxiliary variable $y \in \mathbb{R}^m$ and write (3.25) equivalently as

$$\begin{array}{ll} \underset{x}{\text{minimize}} & (\frac{1}{2}+\frac{\delta}{4})\|y\|^2 + \beta\|x\|_1 + \frac{\delta}{2}\|x\|^2 \\ \text{subject to} & Ax = b + y\,. \end{array} \tag{3.26}$$

Associate a dual variable $\lambda \in \mathbb{R}^m$ to the constraint of (3.26) and compute the dual problem:

$$\underset{\lambda}{\text{maximize}} \;\; b^\top \lambda + \inf_y \left((\frac{1}{2} + \frac{\delta}{4}) \|y\|^2 + \lambda^\top y \right) + \inf_x \left(\beta \|x\|_1 + \frac{\delta}{2} \|x\|^2 - \lambda^\top A x \right) \tag{3.27}$$

$$\iff \quad \underset{\lambda}{\text{minimize}} \;\; b^\top \lambda + \frac{1}{2+\delta} \|\lambda\|^2 + \sum_{p=1}^{P} \bar{h}_p^\star (A_p^\top \lambda) \tag{3.28}$$

$$\iff \quad \underset{\lambda}{\text{minimize}} \;\; \sum_{p=1}^{P} \left[\bar{h}_p^\star (A_p^\top \lambda) + \frac{1}{P} b^\top \lambda + \frac{1}{(2+\delta)P} \|\lambda\|^2 \right], \tag{3.29}$$

which has the format of (G) with $f_p(\lambda) = \bar{h}_p^\star(A_p^\top \lambda) + (1/P) b^\top \lambda + (1/((2+\delta)P)) \|\lambda\|^2$ as the function of each node p. From (3.27) to (3.28), we switched from a maximization problem to a minimization one, and used the fact that the infimum problem in y has a closed-form expression. Also, the infimum in x was decomposed into blocks, and $\bar{h}_p^\star$ denotes the convex conjugate of the function $\bar{h}_p(x_p) = \beta \|x_p\|_1 + (\delta/2) \|x_p\|^2$. From (3.28) to (3.29), we just grouped terms. Note that solving (3.29) is not equivalent to solving BPDN for two reasons: first because we used regularization for which there are no exactness results and, second, because we ignored the quartic term in (3.24).

Column partition: reversed lasso. Regarding reversed lasso (3.6), we will also regularize it and, again, we will not have any exact regularization guarantee. To do the regularization the same way as before, we first rewrite it with the format of (3.11):

$$\begin{array}{ll} \underset{x,t,u,v}{\text{minimize}} & 1_n^\top t \\ \text{subject to} & \|u\| \leq v \\ & -t \leq x \leq t \\ & u = Ax - b \\ & v = \sigma \,, \end{array} \tag{3.30}$$

where $t \in \mathbb{R}^n$ is, again, an epigraph variable, and $u \in \mathbb{R}^m$ and $v \in \mathbb{R}$ are auxiliary variables, introduced to make a cone appear. Problem (3.30) has indeed the same structure as (3.11), since the objective is linear, the last two constraints are linear equalities, and the cone $\mathcal{K}$ is the Cartesian product of two cones: $\mathcal{K} = \mathcal{K}_{x,t} \times \mathcal{K}_{u,v}$, where $\mathcal{K}_{x,t} = \{(x,t) \,:\, -t \leq x \leq t\}$ is polyhedral, and $\mathcal{K}_{u,v} = \{(u,v) \,:\, \|u\| \leq v\}$ is the Lorenz cone and, thus, not polyhedral. By regularizing (3.30) with the function $\phi(z) = (1/4)\|z\|^2$, we obtain

$$\begin{array}{ll} \underset{x,t,u,v}{\text{minimize}} & 1_n^\top t + \frac{\delta}{4} \left(\|x\|^2 + \|t\|^2 + \|u\|^2 + v^2 \right) \\ \text{subject to} & \|u\| \leq v \,, \quad v = \sigma \\ & -t \leq x \leq t \\ & u = Ax - b \end{array} \tag{3.31}$$

$$\begin{array}{lll} \Longleftrightarrow & \underset{x}{\text{minimize}} & \frac{\delta}{4}\|x\|^2 \;+\; \underset{t}{\inf} \;\; 1_n^\top t + \frac{\delta}{4}\|t\|^2 \;+\; \underset{u}{\inf} \;\; \frac{\delta}{4}\|u\|^2 \\ & & \qquad\quad \text{s.t.} \;\; -t \le x \le t \qquad\quad \text{s.t.} \;\; u = Ax - b \\ & & \qquad\qquad\qquad\qquad\qquad\quad \|u\| \le \sigma \end{array} \tag{3.32}$$

$$\begin{array}{lll} \Longleftrightarrow & \underset{x}{\text{minimize}} & \|x\|_1 + \frac{\delta}{2}\|x\|^2 + \frac{\delta}{4}\|Ax-b\|^2 \\ & \text{subject to} & \|Ax - b\| \le \sigma\,. \end{array} \tag{3.33}$$

From (3.31) to (3.32), we used the constraint $v = \sigma$. From (3.32) to (3.33), we used (3.15) and the fact that

$$\begin{array}{lll} \underset{u}{\inf} & \frac{\delta}{4}\|u\|^2 & = \;\; \mathrm{i}_{\|Ax-b\|\le\sigma}(x) + \dfrac{\delta}{4}\|Ax-b\|^2\,. \\ \text{s.t.} & u = Ax - b & \\ & \|u\| \le \sigma & \end{array}$$

In contrast with BP and similarly to BPDN, there is no proof that (3.33) is an exact regularization of reversed lasso, although experimental results in [197] suggest that exact regularization might occur for the Lorenz cone. In our experimental results, discussed later, we solved (3.33) using $\delta = 10^{-2}$ and the corresponding solutions never differed more than 0.5% from the "true" solution.

We next introduce an auxiliary variable $y \in \mathbb{R}^m$ in (3.33), yielding

$$\begin{array}{ll} \underset{x,y}{\text{minimize}} & \|x\|_1 + \frac{\delta}{2}\|x\|^2 + \frac{\delta}{4}\|y\|^2 \\ \text{subject to} & \|y\| \le \sigma \\ & y = Ax - b\,. \end{array} \tag{3.34}$$

Now, associate a dual variable $\lambda \in \mathbb{R}^m$ to the last constraint of (3.34) and compute the dual problem (without dualizing the first constraint). This gives

$$\begin{array}{lll} & \underset{\lambda}{\text{maximize}} & b^\top\lambda + \;\; \underset{y}{\inf} \;\; \lambda^\top y + \frac{\delta}{4}\|y\|^2 \;+\; \underset{x}{\inf} \;\; \Big[\|x\|_1 + \dfrac{\delta}{2}\|x\|^2 - \lambda^\top A x\Big] \\ & & \qquad\quad \text{s.t.} \;\; \|y\| \le \sigma \end{array} \tag{3.35}$$

$$\Longleftrightarrow \quad \underset{\lambda}{\text{maximize}} \;\; b^\top\lambda - \sigma\|\lambda\| + \;\; \underset{x}{\inf} \;\; \Big[\|x\|_1 + \frac{\delta}{2}\|x\|^2 - \lambda^\top A x\Big] \tag{3.36}$$

$$\Longleftrightarrow \quad \underset{\lambda}{\text{minimize}} \;\; \sum_{p=1}^{P}\Big(h_p^\star(A_p^\top\lambda) + \frac{\sigma}{P}\|\lambda\| - \frac{1}{P}b^\top\lambda\Big)\,. \tag{3.37}$$

From (3.35) to (3.36), we noticed that the problem in y has a closed-form solution that can be computed by solving its dual problem. Namely, its optimal objective is $(\delta/2)\sigma^2 - \sigma\|\lambda\|$. From (3.36) to (3.37), we made the column partition explicit and took exactly the same steps as in the manipulations (3.16)-(3.19), since the problem in x is exactly the same as in (3.16). In fact, notice that by setting $\sigma = 0$ in (3.37) we obtain (3.19), exactly the same way we obtain BP from reversed

lasso in the primal domain. Problem (3.37) has the format of (G) and, similarly to BP, the pth block-component of the primal solution of (3.33) can be obtained at node p after solving the dual problem (3.37): the expression for each component is (3.20), the same as for BP. However, for the reversed lasso, we do not have the theoretical guarantee that, for a small enough δ, the solution of the regularized problem (3.33) is also a solution of the original (3.6).

Column partition: lasso. Finally we address lasso. As with BPDN and the reversed lasso, the regularization we use here is not proven to be exact. Again, we start by rewriting (3.7) as (3.11):

$$\begin{array}{lll} \underset{x}{\text{minimize}} & \frac{1}{2}\|Ax-b\|^2 & \iff \quad \underset{x}{\text{minimize}} \quad \|Ax-b\| \\ \text{subject to} & \|x\|_1 \le \gamma & \text{subject to} \quad \|x\|_1 \le \gamma \end{array}$$

$$\iff \quad \begin{array}{ll} \underset{x,t,u,v}{\text{minimize}} & v \\ \text{subject to} & \|u\| \le v \\ & \|x\|_1 \le t \\ & u = Ax-b \\ & t = \gamma\,, \end{array} \tag{3.38}$$

where, for simplicity, we did not represent the constraint $\|x\|_1 \le t$ as a set of linear inequalities. This can indeed be done by writing 2^n inequalities of the form $r_i^\top x \le t$, where each $r_i \in \mathbb{R}^n$ has ± 1 in its entries; there are 2^n such vectors. Therefore, (3.38) has the same format as (3.11), where the objective is linear, the last two constraints are linear equations, and the first two constraints represent the cone $\mathcal{K}$, which is the Cartesian product of a polyhedral closed convex cone $\mathcal{K}_{x,t} = \{(x,t) \,:\, r_i^\top x \le t,\, i = 1,\dots,2^n\}$ and the Lorenz cone $\mathcal{K}_{u,v} = \{(u,v) \,:\, \|u\| \le v\}$. We now regularize problem (3.38) the same way we regularized the previous problems:

$$\begin{array}{ll} \underset{x,t,u,v}{\text{minimize}} & v + \frac{\delta}{4}\Big(\|x\|^2 + t^2 + \|u\|^2 + v^2\Big) \\ \text{subject to} & \|u\| \le v \\ & \|x\|_1 \le t \\ & u = Ax-b \\ & t = \gamma \end{array} \tag{3.39}$$

$$\iff \quad \begin{array}{lll} \underset{x}{\text{minimize}} & \frac{\delta}{4}\|x\|^2 + \frac{\delta}{4}\|Ax-b\|^2 \;+ & \inf\limits_{v} \;\; \frac{\delta}{4}v^2 + v \\ \text{subject to} & \|x\|_1 \le \gamma & \text{s.t.} \;\; \|Ax-b\| \le v \end{array} \tag{3.40}$$

$$\iff \quad \begin{array}{ll} \underset{x}{\text{minimize}} & \|Ax-b\| + \frac{\delta}{2}\|Ax-b\|^2 + \frac{\delta}{4}\|x\|^2 \\ \text{subject to} & \|x\|_1 \le \gamma\,. \end{array} \tag{3.41}$$

From (3.39) to (3.40), we eliminated the linear constraints. From (3.40) to (3.41), we used the

fact that the optimal value of the problem in v, for a fixed x, is $(\delta/4)\|Ax-b\|^2+\|Ax-b\|$. Now, introduce an auxiliary variable $y\in\mathbb{R}^m$ in (3.41):

$$\begin{array}{ll}\underset{x,y}{\text{minimize}} & \|y\|+\frac{\delta}{2}\|y\|^2+\frac{\delta}{4}\|x\|^2\\ \text{subject to} & \|x\|_1\leq\gamma\\ & y=Ax-b\,,\end{array}$$

and compute the dual problem by dualizing both constraints (μ and λ will be the dual variables associated to the first and second constraints, respectively). We get

$$\begin{array}{lll} & \underset{\lambda,\mu}{\text{maximize}} & b^\top\lambda-\gamma\mu\ +\inf_y\left[\|y\|+\dfrac{\delta}{2}\|y\|^2+\lambda^\top y\right]+\inf_x\left[\mu\|x\|_1+\dfrac{\delta}{4}\|x\|^2-\lambda^\top Ax\right]\\ & \text{subject to} & \mu\geq 0\\ \iff & \underset{\lambda,\mu}{\text{minimize}} & \gamma\mu-b^\top\lambda+g^\star(-\lambda)\ +\sup_x\left[(A^\top\lambda)^\top x-\mu\|x\|_1-\dfrac{\delta}{4}\|x\|^2\right] \qquad (3.42)\\ & \text{subject to} & \mu\geq 0\\ \iff & \underset{\lambda,\mu}{\text{minimize}} & \gamma\mu-b^\top\lambda+g^\star(-\lambda)\ +\displaystyle\sum_{p=1}^{P}\sup_{x_p}\left[(A_p^\top\lambda)^\top x_p-\mu\|x_p\|_1-\dfrac{\delta}{4}\|x_p\|^2\right] \qquad (3.43)\\ & \text{subject to} & \mu\geq 0\\ \iff & \underset{\lambda,\mu}{\text{minimize}} & \displaystyle\sum_{p=1}^{P}\left(\frac{1}{P}(\gamma\mu-b^\top\lambda+g^\star(-\lambda))+l_p(A_p^\top\lambda,\mu)+\mathrm{i}_{\{\mu\geq 0\}}(\mu)\right). \qquad (3.44)\end{array}$$

In (3.42), $g^\star$ is the convex conjugate of $g(y)=\|y\|+(\delta/2)\|y\|^2$. We show in Appendix B that

$$g^\star(\eta)=\sup_x\ \left(\eta^\top x-\|x\|-\frac{\delta}{2}\|x\|^2\right)=\begin{cases}0 & ,\ \|\eta\|\leq 1\\ \frac{1}{2\delta}\Big(\|\lambda\|^2-2\|\lambda\|+1\Big) & ,\ \|\eta\|>1\,.\end{cases} \qquad (3.45)$$

From (3.42) to (3.43), we just made the column partition explicit and, in (3.44), we defined

$$l_p(\eta,\mu)=\sup_{x_p}\left[\eta^\top x_p-\mu\|x_p\|_1-\frac{\delta}{4}\|x_p\|^2\right]. \qquad (3.46)$$

Note that (3.44) has the same format as (G). Because of regularization, the objective in the supremum problem in (3.46) is strictly concave, which means that, after the nodes agree on an optimal dual solution $(\lambda^\star,\mu^\star)$, the pth component of the primal solution of (3.41) will be available at the pth node; see Appendix B for the particular expression.

3.3 Algorithm derivation

We now present our algorithm for the global class (G). As mentioned before, our strategy consists of reformulating (G) as (2.32) and then we applying the multi-block ADMM. For convenience, we recall reformulation (2.32)

$$\begin{array}{ll} \underset{x_1,\ldots,x_P}{\text{minimize}} & f_1(x_1)+f_2(x_2)+\cdots+f_P(x_P) \\ \text{subject to} & x_i = x_j\,, \quad (i,j)\in\mathcal{E}\,, \end{array} \tag{3.47}$$

where $x_p \in \mathbb{R}^n$ is the copy of the original variable $x \in \mathbb{R}^n$ and is held by node p. The optimization variable is now the collection of all the copies: $\bar{x} = (x_1, \ldots, x_P) \in (\mathbb{R}^n)^P$. All these copies are forced to be equal through the constraints of (3.47), which state that, for each edge (i, j) in the network, the copies of nodes i and j are equal. Since by Assumption 3.3 the network is assumed connected, there are no isolated nodes and, hence, all the copies are equal. Consequently, problems (G) and (3.47) are equivalent.

Matrix representation. Recall that, according to Assumption 3.4, we assume the network has a coloring scheme $\mathcal{C}$ with $C = |\mathcal{C}|$ colors. We use $\mathcal{C}_c \subset \mathcal{V}$ to denote the set of nodes that have color $c \in \mathcal{C}$, and $\mathcal{C}(p)$ to denote the color of node p. Also, the number of nodes with color c is represented with $C_c = |\mathcal{C}_c|$. Without loss of generality and to simplify our derivation, we will assume that the nodes are numbered according to this coloring scheme as: $\mathcal{C}_1 = \{1, 2, \ldots, C_1\}$, $\mathcal{C}_2 = \{C_1+1, C_1+2, \ldots, C_1+C_2\}$, $\ldots$, i.e., the first C_1 nodes have color 1, the next C_2 nodes have color $\mathcal{C}_2$, and so on. Now, notice that the constraints in problem (3.47) can be written in matrix format as $(B^\top \otimes I_n)\bar{x} = 0$, where $B \in \mathbb{R}^{P\times E}$ is the node-arc incidence matrix, $\otimes$ is the Kronecker product, and I_n is the identity matrix in $\mathbb{R}^n$. In the node-arc incidence matrix, each column is associated to an edge of the network $(i, j) \in \mathcal{E}$, with 1 in the ith entry, -1 in the jth entry, and zeros in the remaining entries. Given our assumption on the ordering of the nodes and the coloring scheme, we can write $(B^\top \otimes I_n)\bar{x} = (B_1^\top \otimes I_n)\bar{x}_1 + (B_2^\top \otimes I_n)\bar{x}_2 + \cdots + (B_C^\top \otimes I_n)\bar{x}_C$, where $\bar{x}_c$ collects the copies of the nodes in $\mathcal{C}_c$, i.e.,

$$\bar{x} = (\underbrace{x_1, \ldots, x_{C_1}}_{\bar{x}_1}, \underbrace{x_{C_1+1}, \ldots, x_{C_1+C_2}}_{\bar{x}_2}, \ldots, \underbrace{x_{P-C_p+1}, \ldots, x_P}_{\bar{x}_C})\,,$$

and the matrix B is partitioned by rows accordingly. Therefore, (3.47) can be written as

$$\begin{array}{ll} \underset{(x_1,\ldots,x_P)}{\text{minimize}} & \sum_{p\in\mathcal{C}_1} f_p(x_p) + \cdots + \sum_{p\in\mathcal{C}_C} f_p(x_p) \\ \text{subject to} & (B_1^\top \otimes I_n)\bar{x}_1 + (B_2^\top \otimes I_n)\bar{x}_2 + \cdots + (B_C^\top \otimes I_n)\bar{x}_C = 0\,, \end{array} \tag{3.48}$$

where we also grouped the terms in the objective according to the colors of the nodes. We next apply the multi-block ADMM to (3.48).

Applying the multi-block ADMM. We introduced the multi-block ADMM in Subsection 2.1.3. Our reformulations of (G) resulted in problem (3.48), which has the format of (2.21), the problem the multi-block ADMM solves. If we apply the multi-block ADMM (2.22)-(2.26) directly to (3.48), we will see that the update of $\bar{x}_c$ yields C_c independent problems which can consequently be solved in parallel. For example, the first block variable $\bar{x}_1$ is updated as

$$\bar{x}_1^{k+1} = \underset{\bar{x}_1=(x_1,\dots,x_{C_1})}{\arg\min} \sum_{p\in\mathcal{C}_1} f_p(x_p) + {\lambda^k}^\top (B_1^\top \otimes I_n)\bar{x}_1 + \frac{\rho}{2}\left\|(B_1^\top \otimes I_n)\bar{x}_1 + \sum_{c=2}^{C}(B_c^\top \otimes I_n)\bar{x}_c^k\right\|^2, \quad (3.49)$$

where the terms not depending $\bar{x}_1$ were dropped. Developing the quadratic term in (3.49),

$$\begin{aligned}\left\|(B_1^\top \otimes I_n)\bar{x}_1 + \sum_{c=2}^{C}(B_c^\top \otimes I_n)\bar{x}_c^k\right\|^2 \\ = \bar{x}_1^\top (B_1B_1^\top \otimes I_n)\bar{x}_1 + 2\,\bar{x}_1^\top\left(\sum_{c=2}^{C}(B_1B_c^\top \otimes I_n)\bar{x}_c^k\right) + \left\|\sum_{c=1}^{C}(B_c^\top \otimes I_n)\bar{x}_c^k\right\|^2. \quad (3.50)\end{aligned}$$

In the first term of (3.50), $B_1B_1^\top$ is the first diagonal block (of size $C_1 \times C_1$) of the network Laplacian. Because the first C_1 nodes have the same color and, hence, cannot be neighbors, the matrix $B_1B_1^\top$ is diagonal. The pth entry in the diagonal is the degree D_p of node p. Therefore, the first term of (3.50) can be written as $\bar{x}_1^\top (B_1B_1^\top \otimes I_n)\bar{x}_1 = \sum_{p\in\mathcal{C}_1} D_p\|x_p\|^2$. In the second term, $B_1B_c^\top$ is an off-diagonal block of the Laplacian and depicts the links between the nodes with color 1 and the nodes with color c. Namely, if node i has color 1 and node j has color c and they are neighbors, i.e., $(i,j) \in \mathcal{E}$, then the ijth entry of $B_1B_c^\top$ will be -1. Therefore, the second term is written equivalently as $2\,\bar{x}_1^\top\left(\sum_{c=2}^{C}(B_1B_c^\top \otimes I_n)\bar{x}_c^k\right) = -2\sum_{p\in\mathcal{C}_1}\sum_{j\in\mathcal{N}_p} x_p^\top x_j^k$. Finally, the last term of (3.50) does not depend on $\bar{x}_1$ and hence can be dropped. These simplifications render problem (3.49) equivalent to

$$\bar{x}_1^{k+1} = \underset{\bar{x}_1=(x_1,\dots,x_{C_1})}{\arg\min} \sum_{p\in\mathcal{C}_1} f_p(x_p) + \left(\gamma_p^k - \rho\sum_{j\in\mathcal{N}_p} x_j^k\right)^\top x_p + \frac{\rho D_p}{2}\|x_p\|^2, \quad (3.51)$$

where $\gamma_p^k := \sum_{j\in\mathcal{N}_p}\lambda_{pj}^k$ was obtained from the second term of (3.49) as

$${\lambda^k}^\top (B_1^\top \otimes I_n)\bar{x}_1 = ((B_1 \otimes I_n)\lambda^k)^\top \bar{x}_1 = \sum_{p\in\mathcal{C}_1}\underbrace{\sum_{j\in\mathcal{N}_p}{\lambda_{pj}^k}^\top}_{\gamma_p^{k\top}} x_j^k. \quad (3.52)$$

In the last equality in (3.52), we used the fact that the pth entry of the vector $(B_1 \otimes I_n)\lambda^k$ is given by $\sum_{j\in\mathcal{N}_p} \lambda_{pj}^k$. Note that we decomposed the dual variable as $(\ldots, \lambda_{ij}, \ldots)$, where λ_{ij} is associated to the constraint $x_i = x_j$, i.e., the edge between node i and node j. Given our convention that $(i,j) \in \mathcal{E}$ implies that $i < j$ (see Subsection 1.3.1), λ_{ij} is only defined for $i < j$. It is clear that problem (3.51) decomposes into C_1 problems that can be solved in parallel. Namely, node p updates its copy x_p as

$$\begin{aligned} x_p^{k+1} &= \arg\min_{x_p} \; f_p(x_p) + \Big(\gamma_p^k - \rho \sum_{j\in\mathcal{N}_p} x_j^k\Big)^\top x_p + \frac{\rho D_p}{2}\|x_p\|^2 \\ &= \mathrm{prox}_{\tau_p f_p}\Big(\frac{1}{D_p}\sum_{j\in\mathcal{N}_p} x_j^k - \tau_p\gamma_p^k\Big), \end{aligned} \tag{3.53}$$

where the prox operator was defined in (2.34) and $\tau_p = 1/(\rho D_p)$. The problems with respect to the other block variables can be decomposed into parallel problems the same way. The only difference is the definition of γ_p^k, which is different due to the nodes' ordering. Its general definition is

$$\gamma_p^k = \sum_{\substack{j\in\mathcal{N}_p \\ p<j}} \lambda_{pj}^k - \sum_{\substack{j\in\mathcal{N}_p \\ p>j}} \lambda_{jp}^k . \tag{3.54}$$

Note that, from (3.53), each node p needs to know the aggregate sum γ_p^k, but not the individual λ_{ij}'s. According to the multi-block ADMM iterations, namely (2.26), each λ_{ij}, for $(i,j) \in \mathcal{E}$, is updated as $\lambda_{ij}^{k+1} = \lambda_{ij}^k + \rho\,(x_i^{k+1} - x_j^{k+1})$. Replacing this update in the definition of γ_p^k in (3.54), we get

$$\begin{aligned} \gamma_p^{k+1} &= \sum_{\substack{j\in\mathcal{N}_p \\ p<j}} \lambda_{pj}^{k+1} - \sum_{\substack{j\in\mathcal{N}_p \\ p>j}} \lambda_{jp}^{k+1} \\ &= \sum_{\substack{j\in\mathcal{N}_p \\ p<j}} \lambda_{pj}^k + \rho \sum_{\substack{j\in\mathcal{N}_p \\ p<j}} \Big(x_p^{k+1} - x_j^{k+1}\Big) - \sum_{\substack{j\in\mathcal{N}_p \\ p>j}} \lambda_{jp}^k - \rho \sum_{\substack{j\in\mathcal{N}_p \\ p>j}} \Big(x_j^{k+1} - x_p^{k+1}\Big) \\ &= \underbrace{\sum_{\substack{j\in\mathcal{N}_p \\ p<j}} \lambda_{pj}^k - \sum_{\substack{j\in\mathcal{N}_p \\ p>j}} \lambda_{jp}^k}_{=\gamma_p^k} + \rho \sum_{\substack{j\in\mathcal{N}_p \\ p<j}} \Big(x_p^{k+1} - x_j^{k+1}\Big) + \rho \sum_{\substack{j\in\mathcal{N}_p \\ p>j}} \Big(x_p^{k+1} - x_j^{k+1}\Big) \\ &= \gamma_p^k + \rho \sum_{j\in\mathcal{N}_p} \Big(x_p^{k+1} - x_j^{k+1}\Big). \end{aligned}$$

D-ADMM: algorithm for the global class. The resulting algorithm is shown as Algorithm 3, which we named D-ADMM in [163], after Distributed-ADMM. Algorithm 3 solves (3.48), and hence (G), by creating C groups of nodes according to the coloring scheme. The nodes within

Algorithm 3 Algorithm for the global class (D-ADMM)

Initialization: Choose $\rho \in \mathbb{R}$; for all $p \in \mathcal{V}$, set $x_p^0 = \gamma_p^0 = 0_n \in \mathbb{R}^n$ and $\tau_p = 1/(\rho D_p)$; set $k = 0$

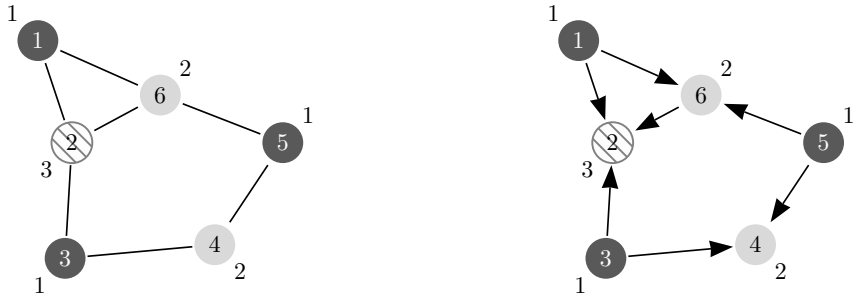
1: **repeat**
2: **for all** $c = 1, \ldots, C$ **do**
3: **for all** $p \in \mathcal{C}_c$ [in parallel] **do**
4: Compute the average

$$z_p^k = \frac{1}{D_p}\Bigg(\sum_{\substack{j \in \mathcal{N}_p \\ \mathcal{C}(j) < \mathcal{C}(p)}} x_j^{k+1} + \sum_{\substack{j \in \mathcal{N}_p \\ \mathcal{C}(j) > \mathcal{C}(p)}} x_j^k \Bigg)$$

5: Update $x_p^{k+1} = \operatorname{prox}_{\tau_p f_p}\left(z_p^k - \tau_p \gamma_p^k\right)$ and send x_p^{k+1} to neighbors $\mathcal{N}_p$
6: **end for**
7: **end for**
8: **for all** $p \in \mathcal{V}$ [in parallel] **do**
9: Update the dual variable $\gamma_p^{k+1} = \gamma_p^k + \rho \sum_{j \in \mathcal{N}_p}\left(x_p^{k+1} - x_j^{k+1}\right)$
10: **end for**
11: $k \leftarrow k + 1$
12: **until** some stopping criterion is met

each group perform the same tasks in parallel, as illustrated before in Figure 1.4. These tasks consist of computing the average of the solution estimates by the neighbors (step 4), computing the prox of the scaled function $\tau_p f_p$ at the point indicated in step 5, and then sending the new solution estimate to the neighbors. Note that in the computation of the average z_p^k of a given node p, in step 4, there are two kinds of estimates: ones that were computed in the current iteration k, i.e., x_j^{k+1} and ones that were computed in the previous iteration $k-1$, i.e., x_j^k. The first kind are estimates of the neighbors with a color smaller than the color of node p, that is, $\mathcal{C}(j) < \mathcal{C}(p)$. Node p has access to these estimates because the nodes with smaller colors have performed steps 4 and 5 before. The second kind are estimates of the neighbors with a color larger than the color of node p, $\mathcal{C}(j) > \mathcal{C}(p)$, and were transmitted in the previous iteration. Note that, in contrast with its derivation, Algorithm 3 does not assume that the nodes are ordered according to their colors, thanks to the use of inequalities $\mathcal{C}(j) \lesseqgtr \mathcal{C}(p)$ instead of $j \lesseqgtr p$. After all nodes perform step 5, the dual variables γ_p are updated simultaneously at all nodes, as described in step 9.

Apparently, Algorithm 3 needs some kind of central coordination to perform steps 4 and 5, because nodes with the same color, not being neighbors, should perform the same tasks in parallel. But, provided Assumption 3.4 holds, i.e., that each node knows its own color and the colors of its neighbors, no central coordination is required. In that case, steps 4 and 5 need not be performed exactly in parallel: as soon as node p has received the copies x_j^{k+1} from the neighbors with smaller



(a) Undirected graph (b) Directed graph

FIGURE 3.2: Construction of a directed graph (in (b)) from the coloring scheme of an undirected graph (in (a)). The coloring scheme is $\mathcal{C}_1 = \{1, 3, 5\}$, $\mathcal{C}_2 = \{4, 6\}$, and $\mathcal{C}_3 = \{2\}$. From (a) to (b), each edge gets assigned a direction, from the node with the smallest color to the node with the largest color.

colors, it can perform steps 4 and 5 immediately. Figure 3.2 illustrates an alternative way to see this. Figure 3.2(a) shows a communication network and its coloring scheme: nodes 1, 3, and 5 have color 1, nodes 4 and 6 have color 2, and node 2 has color 3. From these colors, we can assign directions to the edges of the network, as shown in Figure 3.2(b): the edge $(i, j) \in \mathcal{E}$ is assigned the direction $i \to j$ if the color of node i is smaller than the color of node j, i.e., $\mathcal{C}(i) < \mathcal{C}(j)$, and the direction $i \leftarrow j$ otherwise. For example, node 6, with color 2 has incoming edges from nodes 1 and 5, both with color 1, and an outgoing edge to node 2, with color 3. Whenever node 6 receives, at each iteration, estimates from neighbors 1 and 5, it can immediately perform steps 4 and 5 without "talking" at all with the nodes that have the same color; in this case, that is just node 4. This makes the algorithm distributed, since there is no central or coordinating node, the function f_p is only known at node p, and there are no all-to-all communications. Furthermore, the algorithm is independent of the network. Regarding its convergence, we use Theorem 2.1 to prove:

Theorem 3.6.

Let Assumptions 3.1-3.4 hold. Then, Algorithm 3 produces a sequence $(x_1^k, \ldots, x_P^k)$ *convergent to* $(x^\star, \ldots, x^\star)$*, where* $x^\star$ *solves* (G)*, when at least one of the following conditions is satisfied:*

(a) the coloring scheme uses two colors only (which implies that the network is bipartite);

(b) each function f_p *is strongly convex with modulus* μ_p *and*

$$0 < \rho < \min_{c=1,\ldots,C} \frac{2 \sum_{p \in \mathcal{C}_c} \mu_p}{3\,(C-1)\, \max_{p \in \mathcal{C}_c} D_p}\,. \tag{3.55}$$

Proof. We just need to show that (3.48), the problem to which we apply multi-block ADMM, satisfies the assumptions of Theorem 2.1. First, note that Assumptions 3.1 and 3.2 and the equivalence between (G) and (3.48) imply that problem (3.48) is solvable and that each function $\sum_{p\in\mathcal{C}_c} f_p(x_p)$ is closed and convex over $(\mathbb{R}^n)^C$. Next, we show that condition (a) (resp. (b)) implies condition (a) (resp. (b)) of Theorem 2.1.

(a) We first see that Assumption 3.3 implies that each $B_c^\top \otimes I_n$ has full column rank. Since the identity matrix I_n has always full rank, we just need to show that $B_c^\top$ has full column-rank. If, on the other hand, we prove that $B_c B_c^\top$ has full rank, then the result follows, because $\operatorname{rank}(B_c B_c^\top) = \operatorname{rank}(B_c^\top)$. As mentioned before, $B_c B_c^\top$ is a diagonal matrix, where the diagonal contains the degrees of the nodes belonging to the subnetwork composed by the nodes in $\mathcal{C}_c$. Since no node has degree 0 (cf. Assumption 3.3), $B_c B_c^\top$ has full rank. We thus have shown that, independently of the coloring scheme, each matrix $B_c^\top \otimes I_n$ has full column rank. Therefore, when the coloring scheme uses two colors, both requirements of point (a) in Theorem 2.1 are satisfied.

(b) When each function f_p is strongly convex with modulus μ_p and ρ satisfies (3.55), then each $\sum_{p\in\mathcal{C}_c} f_p$ is strongly convex with modulus $\sum_{p\in\mathcal{C}_c} \mu_p$ [31, Lem. 2.1.4] and conditions (2.27) and (3.55) are equivalent. To see this last point, just note that

$$\sigma_{\max}(A_c)^2 = \lambda_{\max}(A_c^\top A_c) = \lambda_{\max}(B_c B_c^\top \otimes I_n) = \lambda_{\max}(B_c B_c^\top) = \max_{p\in\mathcal{C}_c} D_p\,,$$

since, as we had seen before, each $B_c B_c^\top$ is a diagonal matrix with the degrees of the nodes with color c in the diagonal.

□

As stated before, it is believed that multi-block ADMM converges under condition (a) of Theorem 2.1 when $C > 2$. This requires that each $B_c^\top \otimes I_n$ has full column rank, which we just proved in part (a) of the proof above. Translated to Algorithm 3, this belief means that algorithm converges for generic (non-bipartite) networks when f_p is not necessarily strongly convex, i.e., that Theorem 3.6 holds even when neither condition (a) nor condition (b) are satisfied. Our simulations of Algorithm 3 provide some experimental evidence strengthening that belief, as we will soon see.

Note that the structure of Algorithms 1 and 2, which are based on the 2-block ADMM, is similar to the structure of Algorithm 3: in all of them, a parameter ρ has to be chosen, and each node performs the same kind of computations, i.e., compute an average of the estimates of the neighbors and compute the prox of its private function. While in Algorithms 1 and 2 all the nodes perform

TABLE 3.1: Network models.

Name	Parameters	Description
Erdős-Rényi [200]	p	Every pair of nodes $(i,j) \in \mathcal{E}$ is connected or not with probability p
Watts-Strogatz [201]	(n,p)	First, it creates a lattice where every node is connected to n nodes; then, it rewires every link with probability p. Rewiring link (i,j) means removing the link, and connecting node i or node j (chosen with equal probability) to another node in the network, chosen uniformly.
Barabasi-Albert [202]	——	It starts with one node. At each step, one node is added to the network by connecting it to 2 existing nodes: the probability to connect it to node p is proportional to D_p.
Geometric [203]	d	It drops P points, corresponding to the nodes of the network, randomly in a $[0,1]^2$ square; then, it connects nodes whose (Euclidean) distance is less than d.
Lattice	——	Creates a lattice of dimensions $m \times n$; m and n are chosen to make the lattice as square as possible.

all the tasks in parallel, the nodes in Algorithm 3 operate in a color-based way. Therefore, in environments where parallel communication is allowed, one iteration of Algorithm 3 takes longer than one iteration of Algorithms 1 and 2. In environments where parallel communication is impossible, e.g., in wireless networks, Algorithms 1 and 2 have to implement a MAC protocol and, for example, operate in the same color-based way as Algorithm 3. In either case, simulation shows that Algorithm 3 takes systematically less iterations to converge than Algorithms 1 and 2, for several different problems and several different networks. This means that it is more communication-efficient than the other algorithms, and hence more attractive in scenarios where the nodes are battery-operated.

3.4 Experimental results

In this section, we provide some experimental results that compare the performance of the proposed algorithm with prior distributed optimization algorithms. The performance of all the algorithms will be measured in terms of communication steps, defined next.

Communication steps. We say that a communication step (CS) has occurred whenever all the nodes have transmitted to their neighbors a new solution estimate, usually computed by evaluating a prox operator, as in step 5 of Algorithm 3. The number of CSs an algorithm uses to solve an optimization problem is intrinsic to the algorithm and does not take into account factors like MAC protocols, algorithm implementation, or computing platforms. Other performance measures, for example execution time, may give different results if we change any of these factors. Besides, the total number of communications can be easily obtained from the CSs by multiplying it by $2E$, i.e., by twice the number of edges in the network. Note that Algorithm 1 takes two CSs per iteration, while Algorithms 2 and 3 take only one.

TABLE 3.2: Network parameters, average degree, and number of colors.

Number	Model	Parameters	Average degree (top), Number of colors (bottom)							
			Number of nodes P							
			10	50	100	200	500	700	1000	2000
1	Erdős-Rényi	$1.1\log(P)/P$	3 3	6 5	5 5	6 5	12 7	14 8	18 9	8 6
2	Watts-Strogatz	$(4, 0.4)$	4 3	4 4	4 4	4 4	4 5	4 4	4 4	4 4
3	Barabasi-Albert	——	3 3	4 3	4 3	4 4	4 4	4 4	4 4	4 4
4	Geometric	$\sqrt{\log(P)/P}$	4 5	10 10	12 11	14 12	18 18	19 19	20 17	23 21
5	Lattice	——	3 2	3 2	4 2	4 2	4 2	4 2	4 2	4 2

Networks. We generated several networks in our experiments, ranging from networks with 10 nodes to networks with 2000 nodes. The models we used to generate them are described in Table 3.1. All models, except the lattice, are random, and yield networks with arbitrary topologies. Using these models, we created 40 different networks, as shown in Table 3.2. For each one of the models of Table 3.1, we generated 8 networks with different numbers of nodes, from $P = 10$ nodes, to $P = 2000$ nodes. All the networks were generated in Python [204] with the NetworkX library [205]. The parameters we used to generate the Erdős-Rényi and the geometric networks are known to generate connected networks with high probability. To color the networks, we used a built-in function in Sage [206]. The number of colors of each network and the average node degree are shown in Table 3.2. For example, the network with the largest average degree was the geometric network with 2000 nodes; the same network had the largest number of colors, 21. Note that all the lattice networks were colored with two colors, indicating that they are, in fact, bipartite. Note also that these are the only networks for which Algorithm 3 is proven to converge when the cost functions at each node are not strongly convex (cf. Theorem 3.6).

Choosing ρ. Almost all the algorithms we compare are based on augmented Lagrangian duality and, thus, are parametrized by a parameter ρ. We are unaware of any method that selects a good ρ before executing the algorithm; as discussed in Chapter 2, the existing heuristics for adapting ρ during the execution of the algorithm cannot be implemented in a distributed setting. Therefore, for each algorithm that depends on ρ, we execute the algorithm several times, one for a different value of ρ, and select the one that leads to the best performance. In our experiments, we used two strategies for selecting ρ. The simplest one just selects ρ out of a set of values, typically $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$. In the second strategy, for a given algorithm, we present the chosen value of ρ and give the precision value. We say that $\bar{\rho}$ was chosen with precision $\xi > 0$ for

a given algorithm whenever both $\rho = \bar{\rho} - \xi$ and $\rho = \bar{\rho} + \xi$ lead to more CSs than $\rho = \bar{\rho}$. This definition is motivated by the fact that the number of CSs in augmented Lagrangian algorithms seems to vary with ρ in a convex way.

Next, we present the results of our experiments for each of the applications of Section 3.2. The simplest of these applications is average consensus and, for this reason, we study average consensus in more detail.

3.4.1 Average consensus

We designed two sets of experiments for the average consensus problem. In one of them, we fix the network and run several distributed algorithms, comparing how the error evolves along the iterations (or better, along the CSs). In the other set of experiments, we observe only the total number of CSs that each algorithm takes to achieve a predefined relative error. While the first set of experiments is run on a single network and for many algorithms, the second set of experiments is run for all the networks of Table 3.2 and only for the most competitive algorithms. Next, we describe the how the experiments were designed, then we state which algorithms we compare, and finally we describe the results for both sets of experiments.

Experimental setup. In consensus, each node p holds a scalar θ_p, and the goal is to compute the average of all the θ_p's. We generated each θ_p independently from each other as a realization of a Gaussian distribution with mean 10 and standard deviation 100. Such a large standard deviation was chosen to ensure all the θ_p's differed significantly. We generated 8 sets of these numbers, each set for a network with a fixed number of nodes. This means that one set of θ_p's is used across networks with the same number of nodes, that is, the same set is used, for example, for a geometric network with 200 nodes and for a Barabasi-Albert network with 200 nodes.

In all the algorithms we compare, each node requires an initialization of its solution estimate. In all our experiments, the estimate of node p is initialized with θ_p. We only do this special initialization for the average consensus problem; the reason is to make a fair comparison between algorithms that were designed specifically for consensus and that require this exact initialization, and between general-purpose algorithms, which do not require any special initialization. This contrasts with the results in Figure 1.5, in Chapter 1, where some algorithms were initialized this way and others, including the algorithm we propose, were initialized with zeros. While those results are merely illustrative, they are not as fair as the ones we present next.

Algorithms for comparison. In our experiments, we compare the performance of Algorithm 3 not only with other algorithms solving the problem class (G), but also with algorithms that were designed only for average consensus and that cannot solve any other problem in that class. Namely, the algorithms in [10] and [11] are consensus algorithms and cannot be generalized

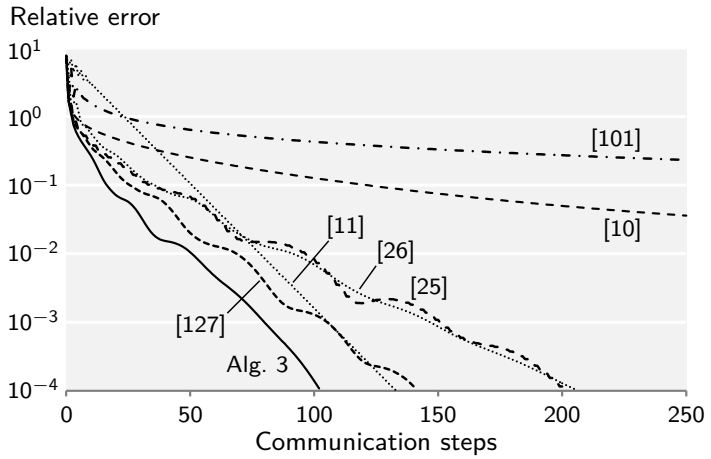
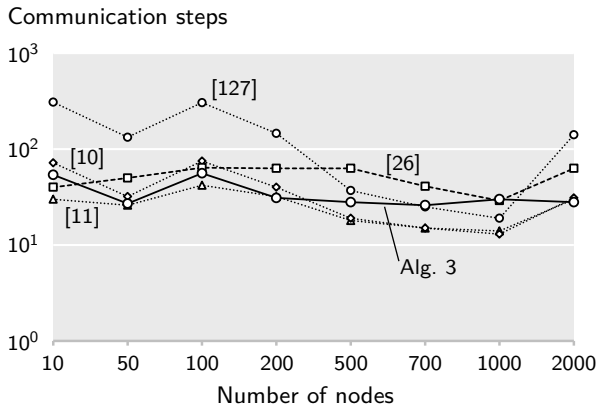


FIGURE 3.3: Comparison of several algorithms for the average consensus problem in a geometric network with $P = 2000$ nodes. The plot shows the relative error versus the number of CSs.

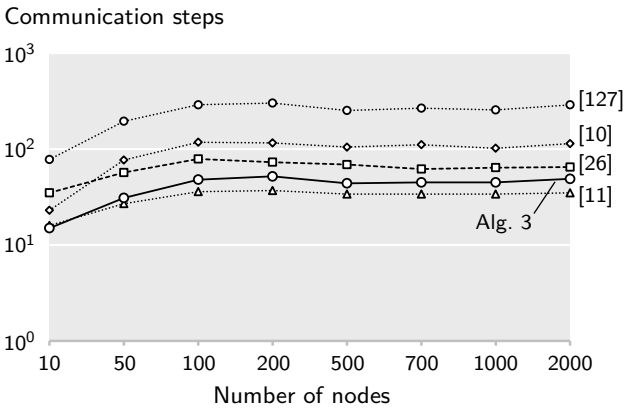
(at least, straightforwardly) to solve other problems written as (G). The algorithm in [11] is actually considered the fastest consensus algorithm, among the synchronous and the asynchronous ones [9]. Since each iteration takes one CS, it is also the most communication-efficient algorithm for consensus. We will see next that the algorithm we propose, when applied to consensus, performs as well as [11], and sometimes better. Note that our algorithm is general-purpose, in contrast with [11], which is specific to consensus.

Regarding general-purpose algorithms, we consider distributed algorithms based on the 2-block ADMM, namely, [25] (written as Algorithm 1), [26] (written as Algorithm 2), and [127]. All these algorithms (and also ours) require computing the prox operator of the function $f_p = (1/2)(x-\theta_p)^2$. This can be done in closed-form: $\text{prox}_{\tau f_p}(\eta) = (\tau\theta_p + \eta)/(1+\tau)$; see (2.34) for the definition of the prox operator. The algorithm in [127] is slightly different from the other ADMM-based algorithms since, instead of just one tuning parameter, it has two: the augmented Lagrangian ρ and a stepsize β. In our experiments, we set always $\beta = 0.9\mu$, just like the authors of [127] did in their experiments. We also consider the (sub)gradient-based method [101], which also solves the class (G). (Actually, the algorithm in [101] solves only unconstrained problems; to solve problems with constraints one has to consider the generalization in [207].) We implemented the algorithm in [101] with uniform weights, i.e., each node averages equally the estimates of its neighbors, and with stepsize $1/(k+1)$.

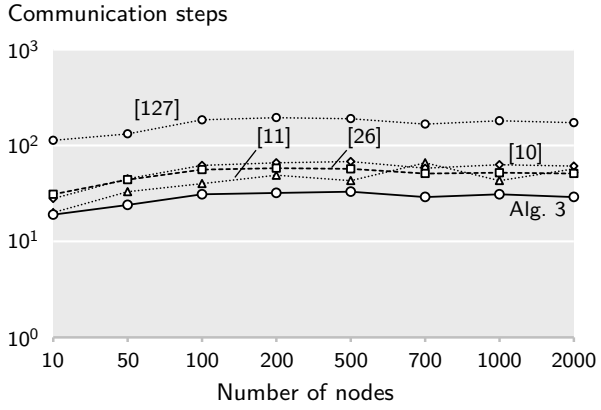
Results. The performance of all the above algorithms is compared on the geometric network with $P = 2000$ nodes, from Table 3.2. This is shown in Figure 3.3 and constitutes our first set of experiments. The plot in the figure shows the evolution of the relative error as a function of the CSs. The relative error is measured as $\|\bar{x}^k - \theta^\star 1_P\|/(\sqrt{P}|\theta^\star|)$, where $\bar{x}^k = (x_1^k, \ldots, x_P^k)$, x_p^k is



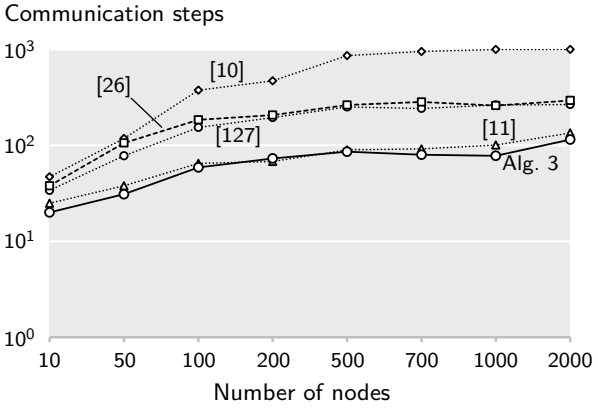
(a) Network 1: Erdős-Rényi



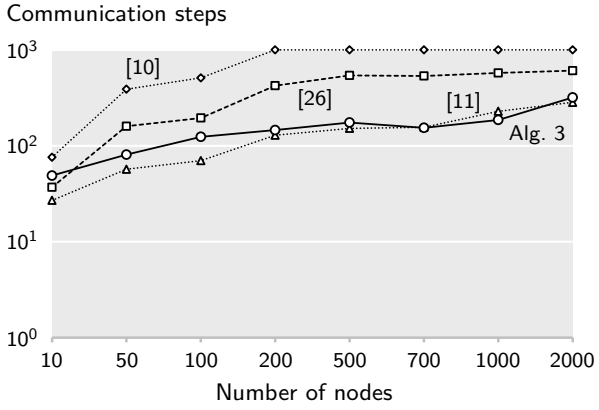
(b) Network 2: Watts-Strogatz



(c) Network 3: Barabasi-Albert



(d) Network 4: Geometric



(e) Network 5: Lattice

FIGURE 3.4: Results for the average consensus problem for all the networks of Table 3.1. The plots, organized by network type, compare Algorithm 3 with algorithms [26] (see Algorithm 2), [127], [10, 11]. The algorithm [127] does not appear in (e), because it always achieved the maximum number of iterations, except for the first network. Note that [10, 11] were designed specifically for consensus and cannot solve any other problem in the class (G).

the solution estimate of node p at iteration k, and $\theta^\star = (1/P)\sum_{p=1}^{P}\theta_p$ is the problem's solution. The augmented Lagrangian parameter ρ was 1.1 for Algorithm 3, 0.5 for [26], 0.6 for [25], and 0.4 for [127], and was computed with precision 0.1 for all the algorithms. In Figure 3.3, Algorithm 3 was the algorithm whose error decreased the fastest; in fact, it required uniformly less CSs than all the other algorithms to achieve any relative error between 10^{-1} and 10^{-4}. The algorithms with the second and third best performances were, respectively, the consensus algorithm [11] and the ADMM-based algorithm [127]. Next, the ADMM-based algorithms [25] and [26] had a very similar performance, requiring about 200 communication steps to achieve a relative error of 10^{-4}. Both the consensus algorithm [10] and the general-purpose algorithm [101] did not converge, i.e., achieve a 10^{-4} relative error in less than 250 CSs.

In our second set of experiments, shown in Figure 3.4, we discarded algorithms [101] and [25], since they exhibited performances inferior to the other algorithms. There are 5 plots in Figure 3.4, one per network type, i.e., row of Table 3.2. In contrast with the plot of Figure 3.3, the plots of Figure 3.4 show the number of CSs to achieve a relative error of 10^{-4} as a function of the network size. For example, in the Watts-Strogatz network with 200 nodes (Figure 3.4(b)), algorithm [127] took 302 CSs to converge, while [10] took 116, [26] took 73, Algorithm 3 took 52, and [11] took 37. The type of networks for which Algorithm 3 performed worst was, in fact, Watts-Strogatz type (Figure 3.4(b)) and Erdős-Rényi type (Figure 3.4(a)). For the remaining networks, Algorithm 3 was always among the best. For example, in Barabasi-Albert network types (Figure 3.4(c)), Algorithm 3 was always the algorithm requiring the least amount of CSs to converge. From theses experiments, we can conclude that Algorithm 3, a general-purpose distributed algorithm, ranks among the most communication-efficient algorithms for solving the average consensus problem.

3.4.2 Row partition: BP and BPDN

We now discuss our experiments on other application problems. In this subsection, we consider compressed sensing problems with a row partition (see Figure 3.1), namely, basis pursuit (BP) and basis pursuit denoising (BPDN). These problems, as well as their reformulation as (G), are discussed in Subsection 3.2.2. Next, we mention the experimental setup and how we implemented the computation of the prox operators. Then, we discuss the experimental results.

Experimental setup. In our experiments, we used all the networks with 50 nodes, i.e., all the networks in the second column of Table 3.2. The exact solution of BP (resp. BPDN) was computed in a centralized way with the Matlab toolbox spgl1 [208] (resp. GPSR [209]). Knowing the solutions of these problems, we were able to assess the relative error of each algorithm along its iterations. Let $x^\star$ denote the solution of either BP or BPDN. The relative error is measured as $\|x^k - x^\star\|/\|x^\star\|$, where x^k is the estimate of an arbitrary node in the network. The algorithms stopped whenever

they reached a relative error of 10^{-4}, or a maximum number of CSs. The maximum number of CSs was 1000 for BP and 2000 for BPDN. All the algorithms we compare are based on ADMM and, thus, have a tuning parameter ρ. In these experiments, ρ was always chosen as the best value from the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$. Regarding the data, i.e., the matrix $A \in \mathbb{R}^{m\times n}$ and the vector $b \in \mathbb{R}^m$, we used two different types of data, one for BP and other for BPDN. For BP, A had dimensions 500×2000 and each entry was generated randomly and independently from a Gaussian distribution with 0 mean and standard deviation $1/\sqrt{500} \simeq 0.045$; since there were 50 nodes, each node stored a matrix of size 10×2000. The vector b was generated from a sparse linear combination of the columns of A. For BPDN, we used a matrix from problem 902 of the Sparco toolbox [210]. That matrix has dimensions 200×1000 and, thus, each node stored a matrix of size 4×1000. The vector b was generated from a sparse linear combination of the columns of A, to which we added Gaussian noise. The noise parameter β in BPDN (see (3.5)) was set to 0.3.

Computation of the prox operator. In ADMM-based algorithms, at each iteration, each node has to compute the prox operator of its function. In the case of BP, the function at node p is given by $f_p(x) = (1/P)\|x\|_1 + \mathrm{i}_{A_p x = b_p}(x)$, as shown in (3.8). Computing the prox of f_p, in this case, is equivalent to finding the minimizer of:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + v^\top x + c\|x\|^2 \\ \text{subject to} & Ax = b\,, \end{array} \tag{3.56}$$

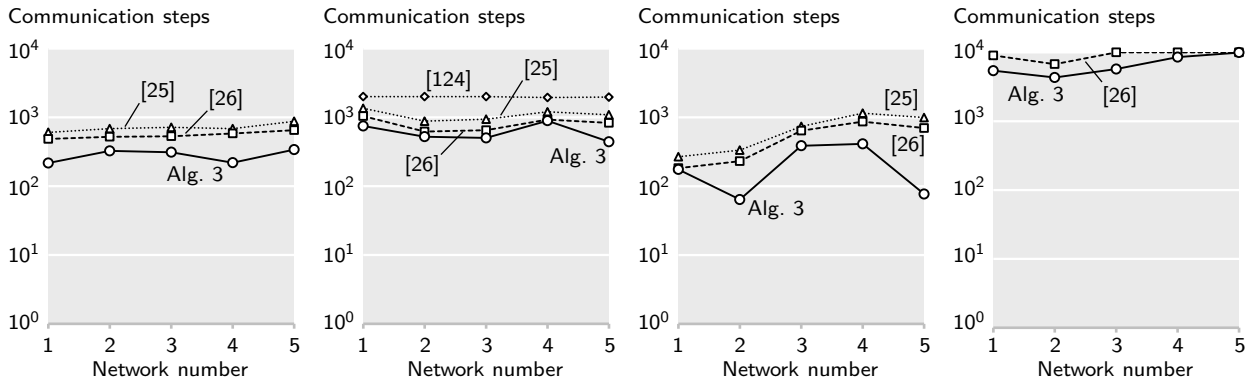
for some vector $v \in \mathbb{R}^n$ and some scalar $c > 0$. To simplify, we dropped the subscripts from the matrix A and the vector b. Since the objective of (3.56) is strictly convex, we can find a primal solution by solving its dual problem:

$$\underset{\lambda}{\text{maximize}} \quad b^\top \lambda + \textstyle\sum_{i=1}^n \inf_{x_i} \left(|x_i| + u_i(\lambda) x_i + c\, x_i^2 \right), \tag{3.57}$$

where $u(\lambda) = v - A^\top \lambda$. We solve (3.57) with the algorithm in [211], which is based on the Barzilai-Borwein method. In our implementation, we used warm-starts, that is, at each iteration and for a given node, the algorithm is initialized with the solution that the node found in the previous iteration.

Regarding BPDN, the function at node p is given by $f_p(x) = (1/2)\|A_p x - b_p\|^2 + (\beta/P)\|x\|_1$, as shown in (3.9). Computing the prox of function f_p, in this case, is actually equivalent to finding a minimizer of a function with the same format as f_p. An efficient method for doing that is GPSR [209], namely GPSR-BB, which uses the Barzilai-Borwein stepsize.

Results. The results for BP and BPDN are shown in Figures 3.5(a) and 3.5(b), respectively. In Figure 3.5(a), we compare Algorithm 3 against the ADMM-based methods [25] and [26] (written,



(a) BP (b) BPDN (c) Reversed lasso (d) SVM

FIGURE 3.5: Results of the simulations for (a) BP, (b) reversed lasso, (c) BPDN, and (d) SVM. The simulations were run on all the networks with 50 nodes.

as Algorithms 1 and 2, respectively). The behavior of the algorithms in this figure is very uniform: in all the networks, Algorithm 3 was the one requiring the least amount of CSs to converge, i.e., to achieve a relative error of 10^{-4}, the algorithm in [25] was always the one requiring the largest amount of CSs, and the algorithm in [26] was always in between.

The exact same behavior can be observed in Figure 3.5(b) for the BPDN, although the lines, and thus their performance, are closer together. The figure also shows the performance of [124, Alg.3], which is an ADMM-based method specifically designed to solve BPDN. That algorithm has the advantage of requiring simpler computations at each node but, as seen in the figure, at the cost of spending more CSs to converge. In fact, that algorithm achieved the maximum number of CSs, i.e., it failed to converge, in all but the last two networks.

3.4.3 Column partition: reversed lasso

In this subsection we give an example of a compressed sensing problem with a column partition. In particular, we consider the reversed lasso (3.6), which we showed how to recast as (G) in Subsection 3.2.2. As in the previous subsections, we first describe the experimental setup, then how we computed the prox operator at each node and, finally, we present the experimental results.

Experimental setup. The set of networks is the same as in the experiments for BP and BPDN, i.e., all the networks with 50 nodes. To compute the problem's solution $x^\star$ beforehand, we used the Matlab toolbox spgl1 [208]. We ran the algorithms either until they reached a maximum number of 1000 CSs or until they reached a relative error of 5×10^{-3}. The relative error has the same expression as before, $\|x^k - x^\star\|/\|x^\star\|$, but now x^k is the concatenation of all the nodes's estimates, i.e., $x^k = (x_1^k, x_2^k, \ldots, x_P^k)$; recall that in the column partition, the variable is partitioned

into blocks and each block is estimated by a single node. Recall also that, in order to recast the reversed lasso as (G), we compute the dual of a regularized version of the problem; see (3.33). The regularization parameter δ was set to 10^{-2} and the noise tolerance σ to 0.1. Again, the parameter ρ was selected from the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$. The problem data is the same as in BPDN, i.e., the matrix A was taken from problem 902 of the Sparco toolbox [210]. This means that A had dimensions 200×1000 and, given the column partition, each node stored a matrix of size 200×20.

Computation of the prox operator. In Subsection 3.2.2 we manipulated the reversed lasso in order to recast it as (G). More specifically, the dual problem of a regularized version of reversed lasso can be written as (3.37), where the function at node p is

$$f_p(\lambda) = h_p^\star(A_p^\top \lambda) + \frac{\sigma}{P}\|\lambda\| - \frac{1}{P} b^\top \lambda \,,$$

and $h_p^\star$ is the convex conjugate of $h_p(x_p) = \|x_p\|_1 + (\delta/2)\|x_p\|^2$. It can be shown that computing the prox of f_p is equivalent to finding the minimizer of the optimization problem:

$$\underset{\lambda}{\text{minimize}} \quad h_p^\star(A_p^\top \lambda) + \tfrac{\sigma}{P}\|\lambda\| - \tfrac{1}{P} b^\top \lambda + v^\top \lambda + c\|\lambda\|^2 \,, \tag{3.58}$$

for some vector $v \in \mathbb{R}^m$ and some scalar $c \in \mathbb{R}$. Introducing an epigraph variable t, (3.58) becomes equivalent to

$$\begin{array}{ll} \underset{\lambda,t}{\text{minimize}} & h_p^\star(A_p^\top \lambda) + \tfrac{\sigma}{P} t - \tfrac{1}{P} b^\top \lambda + v^\top \lambda + c\|\lambda\|^2 \\ \text{subject to} & \|\lambda\| \le t \,. \end{array} \tag{3.59}$$

Since h_p is strongly convex, its conjugate $h_p^\star$ is differentiable and its gradient is Lipschitz-continuous. In fact, the entire objective function of (3.59) is differentiable and its gradient is Lipschitz-continuous with constant $\sigma_{\max}^2(A_p)/\delta + 2c$, where $\sigma_{\max}(A_p)$ is the largest singular value of A_p. Moreover, given an arbitrary point (λ, t), its projection onto the Lorenz cone $\{(\lambda, t) \,:\, \|\lambda\| \le t\}$ is given in closed-form by [57, A.2.7]

$$\begin{cases} (\lambda, t) & , \text{ if } \; t \ge \|\lambda\| \\ (0, 0) & , \text{ if } \; t \le -\|\lambda\| \\ \frac{t+\|\lambda\|}{2}(\frac{\lambda}{\|\lambda\|}, 1) & , \text{ if } \; -\|x\| < t < \|x\| \,. \end{cases}$$

Therefore, (3.58) can be solved with projected gradient methods. We solve it with Nesterov's projected gradient method (2.10), also known as FISTA [58], whose convergence rate is $O(1/k^2)$.

Results. The results of the reversed lasso experiments are shown in Figure 3.5(c). There, Algorithm 3 is compared against the algorithms in [25] and in [26]. They exhibit the same behavior we had observed in Figures 3.5(a) and 3.5(b): Algorithm 3 required uniformly less CSs to converge.

Also, [26] required uniformly less CSs than [25] to converge.

3.4.4 SVM

Finally, we present our experimental results for training an SVM (3.2). Among all experiments that we performed, the ones for SVM required the largest number of CSs to converge, as can be seen by comparing all the plots in Figure 3.5. The results for the SVM experiments are shown in Figure 3.5(d). But before we analyze them, we describe the experimental setup and how we computed the respective prox operator.

Experimental setup. As in the other plots in the same figure, the experiments for the SVM problem (3.2) were executed on the networks with 50 nodes. Since problem (3.2) can be recast as a quadratic program, we obtained the problem's solution beforehand using the `quadprog` function of the Matlab optimization toolbox [212]. The algorithms ran until they achieved a maximum number of 10^4 CSs, or a relative error of 10^{-3}. The relative error in this case was measured exactly as in the compressed sensing problems with a row partition: $\|x^k - x^\star\|/\|x^\star\|$, where x^k is the estimate at an arbitrary node. And the augmented Lagrangian parameter ρ was selected exactly as in the previous experiments. Regarding the problem data, i.e., the sets of datapoints (x_k, y_k) in (3.2), we used data from [213], namely two overlapping sets of datapoints from the Iris dataset. In total, there were $m = 100$ points of size $n = 4$, which means that each node stored 2 datapoints. The parameter β in (3.2) was set to 1 in all the experiments.

Computation of the prox operator. We showed in Subsection 3.2.2 that in the SVM problem the function at each node is given by (3.3). It can be easily seen that computing the prox operator of (3.3) is equivalent to finding a minimizer of a quadratic program with inequality constraints. This problem has no closed-form solution, but it can be solved with standard quadratic program solvers, such as Matlab's `quadprog` function. We used this function in our implementation.

Results. As mentioned, the results of the experiments for SVM are shown in Figure 3.5(d). In this case, the algorithm in [25] achieved always the maximum number of CSs and, thus, is not represented in the plot. Both Algorithm 3 and the algorithm in [26] required always more than 1000 CSs to converge for all the networks. Again, Algorithm 3 required the least number of CSs to converge, never achieving the maximum number of 10^4 CSs. In contrast, the algorithm in [26] achieved the maximum number of CSs in all but the first two networks.

Chapter 4

Connected and Non-Connected Classes

In this chapter, we solve problem (P) with a generic variable, following ideas similar to the ones presented in the previous chapter for the global class. We first address the case of a connected variable, which is simpler, and then we see how to handle a non-connected variable. This chapter is based on the publications [214, 215, 216] and is organized as follows: in Section 4.1, we formally state the problem and outline our assumptions; then, in Section 4.2, we describe some application problems that can be written as (P) with a non-global variable. These include distributed model predictive control (D-MPC), network flow problems, and the reversed lasso with a row partition. In particular, we propose a new framework for D-MPC that considerably extends the modeling capability of the standard D-MPC; this, for example, will allow us to model scenarios where systems coupled through their dynamics do not necessarily communicate directly. Next, in Section 4.3, we derive our algorithm, first for a connected variable, and then for a non-connected variable. This will give us the most general algorithm in this thesis. Finally, in Section 4.4, we show how the performance of the proposed algorithm compares with prior algorithms for some of the problems introduced in Section 4.2.

4.1 Problem statement

As in the global class, here we also minimize the sum of P functions, where each function is known at one node only. However, each function here, rather than depending on all the components of the variable $x \in \mathbb{R}^n$, depends only on the ones indexed by the set $S_p \subseteq \{1, \ldots, n\}$. That is, we solve

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f_1(x_{S_1}) + f_2(x_{S_2}) + \cdots + f_P(x_{S_P})\,. \tag{P}$$

We make the following assumptions:

Assumption 4.1. *Each function $f_p : \mathbb{R}^{n_p} \to \mathbb{R} \cup \{+\infty\}$, not identically $+\infty$, is closed and convex over $\mathbb{R}^{n_p}$.*

Assumption 4.2. *Problem* (G) *is solvable, i.e., it has at least one solution $x^\star \in \mathbb{R}^n$.*

Assumptions 4.1 and (4.2) are essentially the same we made for the global class. The only difference is that now each function f_p is defined over $\mathbb{R}^{n_p}$, where $n_p = |S_p|$, and not over the entire domain of the variable x, $\mathbb{R}^n$. Note that the sum of the dimensions of the domains of each function, i.e., $n_1 + \cdots + n_P$, is always less than or equal to the corresponding sum in the case of a global variable, which is nP. In other words, $n_1 + \cdots + n_P \leq nP$. The following assumption makes the problem well-formulated by guaranteeing that, for each component x_l, there is always one node p whose function depends on x_l, i.e., $l \in S_p$:

Assumption 4.3. *There holds $\cup_{p=1}^{P} S_p = \{1, 2, \ldots, n\}$.*

This assumption was not required for the global class, because all functions there depended on all the components of the variable. Regarding the network, we make exactly the same assumptions we made for the global class:

Assumption 4.4. *The network is connected and does not vary with time.*

Assumption 4.5. *A coloring scheme $\mathcal{C}$ of the network is available; each node knows its own color and the color of its neighbors.*

The comments we made in Section 3.1 about these assumptions also apply here. Next, we describe some application problems that can be written as (P) with a non-global variable and under Assumptions 4.1-4.5.

4.2 Applications

There are many problems in signal processing, control engineering, and machine learning that can be written as (P). In the previous chapter, we described some that require a global variable. In this section, we focus on problems that require a variable that is non-global, for example, a star-shaped or a mixed variable. We start with distributed model predictive control (D-MPC), which appears in the literature as an instance of (P) with a star-shaped variable. One of the contributions of this thesis is a new framework for D-MPC that uses generic connected, and even non-connected, variables. This new framework allows modeling D-MPC scenarios where systems that are coupled through their dynamics need not to communicate directly. The second application

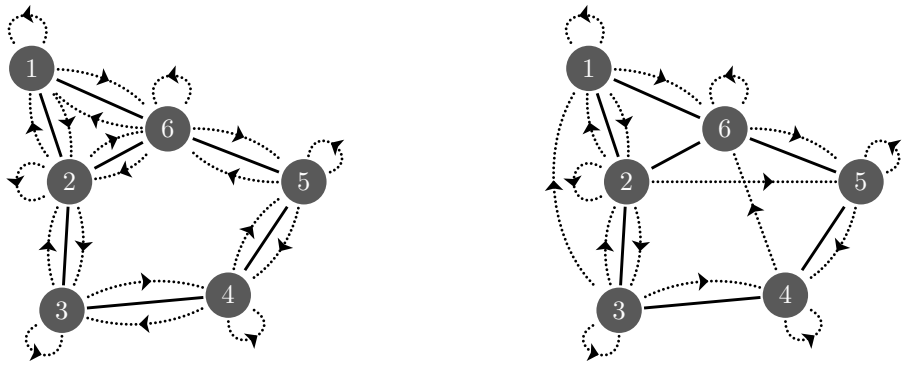
we will see the distributed compressed sensing problem reversed lasso with a row partition, which we formulate as (P) with a mixed variable. Then, we describe three applications that have been solved with distributed algorithms: network flow problems (star-shaped variable), network utility maximization (NUM) (star-shaped and mixed variable), and state estimation in power networks (star-shaped variable). The last application, state estimation in power networks, is described in [47], which also proposes an ADMM-based algorithm to solve it. This is the only algorithm we found in the literature that can be easily generalized to solve (P) for all types of variables. At the end of this section, we will describe the algorithm in [47] for a generic connected variable.

4.2.1 Distributed model predictive control

This subsection describes model predictive control (MPC), first from a centralized perspective, and then from a distributed one.

Centralized MPC. As mentioned in Chapter 2, model predictive control (MPC) is a popular strategy for controlling discrete-time systems. In MPC, a system is described at each time instant t by its state-space vector $x[t] \in \mathbb{R}^n$, whose value at time $t+1$ is determined by the state and control input at time t. Mathematically, $x[t+1] = \Theta^t(x[t], u[t])$, where $u[t] \in \mathbb{R}^m$ denotes the control input applied to the system at time t and $\Theta^t : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is an arbitrary, time-variant map modeling the system. Being a control strategy, the goal of MPC is to take the state vector of the system from an initial point $x[0]$ to some predefined "goal state." To be more concrete, let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a function that penalizes deviations from the goal state or, in other words, $\Phi(x)$ increases with the distance of x to the goal state. Almost always, there are several possible paths from $x[0]$ to the goal state and, typically, these paths have different energy consumptions, for example, the energy spent on the input signals $u[0], u[1], \ldots$. We model energy consumption at time t with the function $\Psi^t(x[t], u[t])$. Therefore, we want to choose the path from $x[0]$ to the goal state that uses the minimum amount of energy; this is actually the problem solved by MPC. However, in MPC, we make the key assumption that the system can measure its state at each time instant. This capability is used to mitigate model inaccuracies and disturbances to the system. It works as follows: instead of solving the problem at once, time is divided into slots of T units, where T is called the time-horizon. At each time-instant, the time variable t is set to zero and the state is measured, say, $x[0] = x^0$, where x^0 is the known measurement. Then, the following optimization problem is solved for a time-horizon T:

$$\begin{array}{ll} \underset{\bar{x},\bar{u}}{\text{minimize}} & \Phi(x[T]) + \sum_{t=0}^{T-1} \Psi^t(x[t], u[t]) \\ \text{subject to} & x[t+1] = \Theta^t(x[t], u[t]), \quad t = 0, \ldots, T-1 \\ & x[0] = x^0, \end{array} \tag{4.1}$$



(a) Connected star-shaped variable (b) Non-connected variable

FIGURE 4.1: Two D-MPC scenarios. Solid lines represent links in the communication network and dotted arrows represent system interactions. The optimization variable is star-shaped (and thus connected) in (a) and is non-connected in (b), because node 2 influences node 5 but not any neighbor of that node.

where $(\bar{x}, \bar{u}) := (\{x[t]\}_{t=0}^{T}, \{u[t]\}_{t=0}^{T-1})$ is the optimization variable and represents the set of states (resp. inputs) from time $t = 0$ to time T (resp. $T-1$). In the objective of (4.1), there is a tradeoff between achieving the goal state at time T, expressed by the term $\Phi(x[T])$, and minimizing the path energy, expressed by the term $\sum_{t=0}^{T-1} \Psi^t(x[t], u[t])$. While the first constraint in (4.1) enforces the state to satisfy the system dynamics, the second constraint encodes the measurement x^0. After solving problem (4.1), the first input $u[0]$ is applied to the system, the time t is again set to zero, and the process is repeated. This means that, at each time instant, only the first input is used, even though a set of inputs and states are computed for the entire horizon from $t = 0$ to $t = T$. MPC thus provides a conservative strategy to deal with model inaccuracies and system disturbances, which perhaps explains its effectiveness and, consequently, its popularity.

D-MPC. We now turn to distributed scenarios and focus on solving one instance of (4.1), i.e., for a fixed MPC iteration. Suppose that, instead of a single system, we now have a network of systems, where each system is described by its own state vector and has a local control input. Let $x_p[t] \in \mathbb{R}^{n_p}$ denote the state of system p at time t, and $u_p[t] \in \mathbb{R}^{m_p}$ denote its local input also at time t; we have $n_1 + \cdots + n_P = n$ and $m_1 + \cdots + m_P = m$. Each system is viewed as a node of a communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, whose edges determine which systems communicate directly. We assume that the state of system p evolves as

$$x_p[t+1] = \Theta_p^t(\{x_j[t], u_j[t]\}_{j \in \Omega_p}), \tag{4.2}$$

where $\Omega_p \subseteq \mathcal{V}$ is the set of nodes whose state and/or input influences x_p (we assume each node p influences itself, i.e., $\{p\} \subseteq \Omega_p$). In (4.2), we used the following notation: given a finite set $\Omega =$

$\{\omega_1, \omega_2, \ldots, \omega_L\}$ and a vector z_ω, indexed by a parameter $\omega \in \Omega$, the symbol $\{z_\omega\}_{\omega\in\Omega}$ denotes the L-tuple $(z_{\omega_1}, z_{\omega_2}, \ldots, z_{\omega_L})$. Many times, when Ω is represented as $\Omega = \{\omega \,:\, A(\omega) \text{ holds}\}$, we will represent $\{z_\omega\}_{\omega\in\Omega}$ simply as $\{z_\omega\}_{A(\omega) \text{ holds}}$. In contrast with what is usually assumed, Ω_p in (4.2) is not necessarily a subset of the neighbors of node p. This means that two systems that influence each other through their dynamics may be unable to communicate directly. This is illustrated in Figure 4.1(b) where, for example, the state/input of node 3 influences the state evolution of node 1 (dotted arrow), but there is no communication link (solid line) between them. Finally, we assume functions Φ and Ψ^t in (4.1) can be decomposed, respectively, as $\Phi(x[T]) = \sum_{p=1}^{P} \Phi_p(\{x_j[T]\}_{j\in\Omega_p})$ and $\Psi^t(x[t], u[t]) = \sum_{p=1}^{P} \Psi_p^t(\{x_j[t], u_j[t]\}_{j\in\Omega_p})$, where Φ_p and Ψ_p^t are both associated to node p. This means that non-communicating systems can have coupled goals or energy measures. Hence, in our distributed setting, the MPC problem (4.1) becomes

$$\begin{array}{ll} \underset{\bar{x},\bar{u}}{\text{minimize}} & \sum_{p=1}^{P}\Big[\Phi_p(\{x_j[T]\}_{j\in\Omega_p}) + \sum_{t=0}^{T-1} \Psi_p^t(\{x_j[t], u_j[t]\}_{j\in\Omega_p})\Big] \\ \text{subject to} & x_p[t+1] = \Theta_p^t(\{x_j[t], u_j[t]\}_{j\in\Omega_p})\,, \quad t = 0, \ldots, T-1\,, \quad p = 1, \ldots, P \\ & x_p[0] = x_p^0\,, \quad p = 1, \ldots, P\,, \end{array} \tag{4.3}$$

where x_p^0 is the initial measurement at node p. The optimization variable in this case is $(\bar{x}, \bar{u}) := (\{\bar{x}_p\}_{p=1}^P, \{\bar{u}_p\}_{p=1}^P)$, where $\bar{x}_p := \{x_p[t]\}_{t=0}^T$ and $\bar{u}_p := \{u_p[t]\}_{t=0}^{T-1}$ represent the collection for the time-horizon T of all the states and all the inputs at node p, respectively. Problem (4.3) can be written as (P) by making

$$f_p\Big(\{\bar{x}_j, \bar{u}_j\}_{j\in\Omega_p}\Big) = \Phi_p\Big(\{x_j[T]\}_{j\in\Omega_p}\Big) + \mathrm{i}_{\{x_p[0]=x_p^0\}}(\bar{x}_p) + \sum_{t=0}^{T-1}\left(\Psi_p^t(\{x_j[t], u_j[t]\}_{j\in\Omega_p}) + \mathrm{i}_{\Gamma_p^t}(\{\bar{x}_j, \bar{u}_j\}_{j\in\Omega_p})\right),$$

where $\mathrm{i}_{\Gamma_p^t}$ is the indicator function of the set $\Gamma_p^t := \Big\{\{\bar{x}_j, \bar{u}_j\}_{j\in\Omega_p} \,:\, x_p[t+1] = \Theta_p^t(\{x_j[t], u_j[t]\}_{j\in\Omega_p})\Big\}$.

Figure 4.1(a) illustrates the standard D-MPC scenario, where each system p is influenced only by itself and by its neighbors, i.e., $\Omega_p \subseteq \mathcal{N}_p \cup \{p\}$. According to the terminology introduced in Chapter 1, each component of the variable, in this case $(\bar{x}_p, \bar{u}_p)$, is star-shaped and thus the entire variable $(\bar{x}, \bar{u})$ is also star-shaped. Several instances of this particular case of (4.3) have been addressed, for example, by [145, 19, 146, 147], who propose heuristics that are not guaranteed to solve exactly (4.3), and by [144, 150, 151, 46, 152], who propose algorithms based on distributed optimization methods and thus, in principle, are guaranteed to solve (4.3).

The model we propose here is significantly more general, since it can handle scenarios where interacting nodes do not necessarily need to communicate, or even scenarios with a non-connected

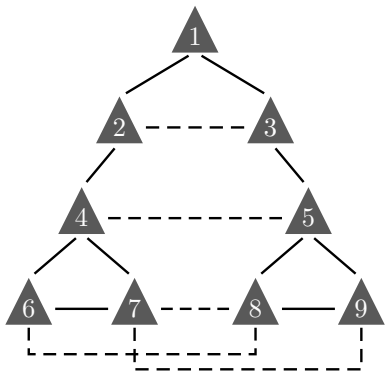


FIGURE 4.2: Example of a geometrical pattern used in formation for minimizing the effect of drag forces or for escorting a moving object. Solid lines indicate direct communication, while dashed lines indicate dynamic coupling, but not necessarily direct communication.

variable. Both cases are shown in Figure 4.1(b). For example, the subgraph induced by $(\bar{x}_3, \bar{u}_3)$ consists of the nodes $\{1, 2, 3, 4\}$ and is connected. (The reference for connectivity is always the communication network which, in the plots, is represented by solid lines.) Nodes 1 and 3, however, cannot communicate directly. This is an example of an induced subgraph that is not a star. On the other hand, the subgraph induced by $(\bar{x}_2, \bar{u}_2)$ consists of the nodes $\{1, 2, 3, 5\}$. This subgraph is not connected, which implies that the optimization variable is non-connected. A connected variable with induced subgraphs that are not starts, or even a non-connected variable, can be useful to model scenarios where communications links are expensive, hard to establish, or simply do not exist. We describe two such applications below.

Applications of our D-MPC model. Although D-MPC has been applied to solve many applications, we present here two applications where the scenario of Figure 4.1(b) might arise naturally, i.e., the variable is either non-connected, or is connected but not star-shaped. The first application is flight formation and the other is temperature regulation of buildings.

Figure 4.2 shows the setup of flight formation: there is a group of autonomous agents, such as unmanned airplanes, submarines, or robots, whose goal is to form a geometrical pattern while performing some task. This task could be simply flying and, at the same time, trying to minimize the effect of drag forces to reduce fuel consumption; or, for example, to escort a moving object, which might block some communications between the agents. We assume there is a communication network through which the agents communicate. In Figure 4.2, the links of this communication network are represented by the solid lines. Also, some agents influence the behavior of other agents with which they do not communicate directly; this is represented by the dashed lines in Figure 4.2. Flight formation is a widely studied topic and we refer to [217] for references and related work. In this problem, we extend the optimization model used in [217] to the MPC framework (4.1).

Namely, we write the dynamics of the pth agent as (4.2), where the relative position between two agents affects their dynamics due, for example, to drag forces. Regarding the objective, while Ψ_p^t models fuel consumption at time t, Φ_p models the geometrical pattern to be formed. Note that, in principle, Θ_p^t and Ψ_p^t depend only on the state/input of agent p and of its closest agents (i.e., its neighbors in the communication network); in Φ_p we can, in addition, include dependencies on agents that are not within communication reach. For example, suppose we specify agent 6 in Figure 4.2 to have a relative distance of δ_{46} and δ_{67} from its closest neighbors, agents 4 and 7, and also a relative distance of δ_{68} from agent 8, for symmetry reasons. Note that agents 6 and 8 do not communicate directly. In this case, $\Omega_6 = \{4, 6, 7, 8\}$, and Φ_6 would have a format similar to $\Phi_6(x_4, x_6, x_7, x_8) = \frac{1}{2}\|x_4 - x_6 - \delta_{46}\|^2 + \frac{1}{2}\|x_6 - x_7 - \delta_{67}\|^2 + \frac{1}{2}\|x_6 - x_8 - \delta_{68}\|^2$, where x_p represents the position of agent p; note that we dropped the time index T for notational simplicity. A similar reasoning can be applied to the remaining agents of Figure 4.2, where relative distances are specified for each edge (represented either with continuous or dashed lines).

We now describe another application of D-MPC where the variable can be connected, not necessarily star-shaped, or even non-connected. The application is temperature regulation of buildings and is described in the context of D-MPC in [218]. The algorithm proposed in [218], however, is heuristic and, thus, not guaranteed to solve the original problem. The motivation for using MPC in the control of room temperature stems from its ability to integrate in its model the prediction of future events, in this case, room occupation profiles. This feature is necessary in the regulation of room temperature, because temperature varies very slowly. If it did not, a simple PID controller would be enough. The work in [218] models room temperature, viewing it as a state x, which varies linearly with the heating power applied to the room, u. According to our notation in (4.1), the function Φ is identically zero, and $\Psi^t(x[t], u[t]) = \beta u[t] + \delta[t]|x[t] - r[t]|$, where $r[t]$ is the reference temperature for the room at time t, and $\delta[t] = 1$ if the room is predicted to be occupied at time t and $\delta[t] = 0$ otherwise. The parameter $\beta > 0$ sets the tradeoff between energy consumption and comfort. The power $u[t]$ is constrained to an interval: $0 \leq u[t] \leq u_{\max}$. This is the model for one room. However, [218] also models buildings, where the rooms are thermally coupled. The model it proposes can be written as (4.3), where the sets Ω_p's model coupling between adjacent rooms. Yet, it is assumed that adjacent rooms can communicate or, in other words, that interactions through coupling coincide with interactions through communication. If the communication technology is wired, it might be expensive to connect all the adjacent rooms with cables; if it is wireless, large concrete walls might prevent adjacent rooms from communicating directly. This is clearly an example where our proposed D-MPC model (4.3) could be useful, as the problem variable might have induced subgraphs that are not stars and might even be non-connected.

4.2.2 Reversed lasso with a row partition

We saw in Chapter 3 how to recast several compressed sensing problems as (G), that is, as (P) with a global variable. The only problem for which we were not able to do so was the reversed lasso (3.6) with a row partition. The goal of this subsection is to complete this missing part of the puzzle, by recasting that problem as (P) with a mixed variable.

Recall that the reversed lasso (3.6) is the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 \\ \text{subject to} & \|Ax-b\| \leq \sigma\,. \end{array} \tag{4.4}$$

Since we will use duality, we want to make sure that the primal objective is strictly convex, so that we can recover a primal solution after having solved the dual. We already computed a regularization of the reversed lasso, which we solved with a column partition, in (3.33) (page 53). So, for a small $\delta > 0$, (4.4) can be approximated by

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \frac{\delta}{2}\|x\|^2 + \frac{\delta}{4}\|Ax-b\|^2 \\ \text{subject to} & \|Ax-b\|^2 \leq \sigma^2\,, \end{array} \tag{4.5}$$

where we also squared both sides of the constraint. Consider now a row partition, as visualized in Figure 3.1, and rewrite (4.5) as

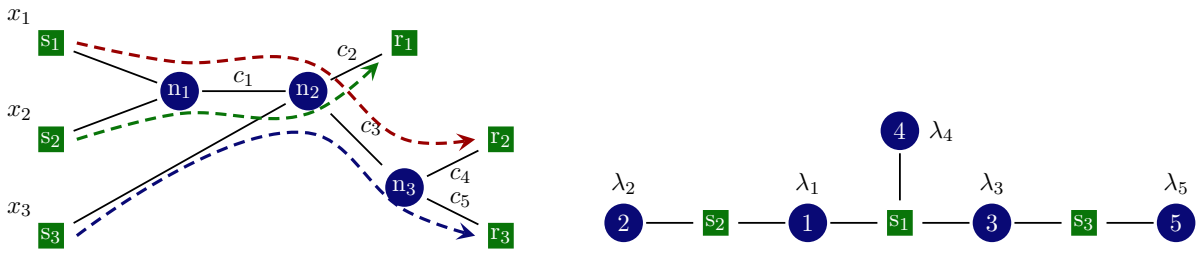
$$\begin{array}{ll} \underset{x}{\text{minimize}} & \sum_{p=1}^{P}\left(\frac{1}{P}\|x\|_1 + \frac{\delta}{2P}\|x\|^2 + \frac{\delta}{4}\|A_p x-b_p\|^2\right) \\ \text{subject to} & \sum_{p=1}^{P}\|A_p x-b_p\|^2 \leq \sigma^2\,. \end{array} \tag{4.6}$$

This problem has the following format

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f_1(x)+f_2(x)+\cdots+f_P(x) \\ \text{subject to} & h_1(x)+h_2(x)+\cdots+h_P(x) \leq r\,, \end{array} \tag{4.7}$$

with $f_p(x) = \frac{1}{P}\|x\|_1 + \frac{\delta}{2P}\|x\|^2 + \frac{\delta}{4}\|A_p x-b_p\|^2$, $h_p(x) = \|A_p x-b_p\|^2$, and $r=\sigma^2$. We will now see how to solve (4.7) in a network where each node p knows f_p, h_p, and r. We assume that each function f_p is strictly convex, as in (4.6). We start by cloning the variable x, rewriting (4.7) as

$$\begin{array}{ll} \underset{x_1,\ldots,x_P}{\text{minimize}} & f_1(x_1)+f_2(x_2)+\cdots+f_P(x_P) \\ \text{subject to} & h_1(x_1)+h_2(x_2)+\cdots+h_P(x_P) \leq r \\ & x_i = x_j\,, \quad (i,j)\in\mathcal{E}\,, \end{array} \tag{4.8}$$



(a) Original network (b) Bipartite graph obtained from (a)

FIGURE 4.3: (a) Example network with 3 source nodes s_1, s_2, and s_3, that use predetermined routes to send packets to three recipient nodes r_1, r_2, and r_3; (b) Bipartite graph obtained from (a): each link from (a) with a capacity associated is represented as a circular node in (b).

where x_p is the copy of x held at node p. Let μ be a dual variable associated to the first constraint of (4.8) and λ_{ij} the dual variable associated to the constraint $x_i = x_j$, for $(i,j) \in \mathcal{E}$. The dual problem of (4.8) is

$$\begin{array}{ll} \underset{\mu,\{\lambda_{ij}\}_{(i,j)\in\mathcal{E}}}{\text{minimize}} & g_1(\mu,\{\lambda_{1j}\}_{j\in\mathcal{N}_1}) + g_2(\mu,\{\lambda_{2j}\}_{j\in\mathcal{N}_2}) + \cdots + g_P(\mu,\{\lambda_{Pj}\}_{j\in\mathcal{N}_P}) \\ \text{subject to} & \mu \geq 0\,, \end{array} \tag{4.9}$$

where, for each p,

$$g_p(\mu,\{\lambda_{pj}\}_{j\in\mathcal{N}_p}) := \sup_{x_p} \Big(\sum_{j\in\mathcal{N}_p} \text{sign}(p-j)\lambda_{pj}\Big)^\top x_p - \Big(f_p(x_p) + \mu^\top(h_p(x_p) - \frac{1}{P}r)\Big)\,. \tag{4.10}$$

We slightly abused notation in (4.9), since each λ_{ij} is only defined for $i < j$. We extended that definition: $\lambda_{ij} = -\lambda_{ji}$ when $i > j$. We thus can see that (4.9) has the same format as (M) or, in other words, it has a mixed variable. The dual variable μ is a global component, since it appears in the function of all the nodes, and all the components of $\lambda = (\ldots,\lambda_{ij},\ldots)$ are non-global. Since we assume that each f_p is strictly convex, the pth block of the primal variable of (4.7), i.e., $x_p^\star$ will be available at the pth node as the solution of the optimization problem in (4.10), for $\mu = \mu^\star$ and $\{\lambda_{pj}\}_{j\in\mathcal{N}_p} = \{\lambda_{pj}^\star\}_{j\in\mathcal{N}_p}$, where the starred vectors solve the dual problem (4.9).

4.2.3 Network utility maximization

Network utility maximization (NUM) is usually used for modeling congestion control in networks. The setup of congestion control is a network with some source nodes sending information, encoded in packets, to other nodes of the network, called recipient nodes. Independently of the network, its links have always finite capacity and, therefore, there is a limit on the number of packets that can circulate in the network without congesting it. The goal of congestion control is to avoid

congesting the network; this is done by implementing a protocol between the source nodes and the nodes through which they send their packets, called intermediate nodes. The communication between these nodes can occur implicitly or explicitly. The algorithm we propose for solving (P) will require an explicit communication between the source and the intermediate nodes.

Star-shaped variable model. Given a network, let $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{N}$ represent the source nodes, the recipient nodes, and the intermediate nodes, respectively. Figure 4.3(a) shows an example of such network with 3 nodes of each kind, i.e., $|\mathcal{S}| = |\mathcal{R}| = |\mathcal{N}| = 3$. The source nodes are squares on the left side, the recipient nodes are the squares on the right side, and the intermediate nodes are the circles. We assume each source sends packets only to one recipient node. Also, the routes through which each source sends its packets are predetermined (see the arrows in Figure 4.3(a)). Each link l in the network has a finite capacity $c_l > 0$, and the total number of links will be denoted with L. In Figure 4.3(a), to simplify, we represent the capacity of only some links. Congestion control can be modeled with a problem called *network utility maximization* (NUM):

$$\begin{array}{ll} \underset{\{x_s\}_{s\in\mathcal{S}}}{\text{maximize}} & \sum_{s\in\mathcal{S}} U_s(x_s) \\ \text{subject to} & \sum_{s\in\mathcal{S}(l)} x_s \le c_l\,, \quad l = 1,\dots,L\,, \end{array} \tag{4.11}$$

where x_s represents the sending rate of source s and $U_s(x_s)$ its utility, or "satisfaction." The constraints in (4.11) are simply the link capacity constraints: $\mathcal{S}(l)$ represents the set of sources that use link l and thus $\sum_{s\in\mathcal{S}(l)} x_s$ represents the rate of packets flowing in link l, which has to be smaller than the link capacity c_l. The goal in (4.11) is to maximize the aggregate utilities of the sources, while satisfying the link capacity constraints. It is generally assumed that each utility is increasing and strictly concave. For example, TCP Vegas, FAST, and Scalable TCP have been modeled as (4.11) with $U_s(x_s) = w_s \log x_s$, for some $w_s > 0$ [132, 134]. This makes the objective of (4.11) strictly concave, and hence its dual problem can be solved instead:

$$\begin{array}{ll} \underset{\lambda=(\lambda_1,\dots,\lambda_L)}{\text{minimize}} & \sum_{l\in\mathcal{L}} c_l\lambda_l + \sum_{s\in\mathcal{S}} \bar{U}_s\Big(\sum_{l\in\mathcal{L}(s)} \lambda_l\Big) \\ \text{subject to} & \lambda_l \ge 0\,, \quad l = 1,\dots,L\,, \end{array} \tag{4.12}$$

where $\mathcal{L}(s)$ is the set of links source s uses to route its packets and $\bar{U}_s(t) := \sup_{x_s}\big(U_s(x_s) - tx_s\big)$ has always a unique solution $x_s(t)$, due to the strict concavity of U_s. In the case of TCP Vegas, FAST, and Scalable TCP, $\bar{U}_s(t) = w_s \log w_s - w_s \log t - w_s$, and $x_s(t) = w_s/t$. After a solution $\lambda^\star$ to (4.12) has been found, the optimal value for the rate of source s can be found as $x_s^\star = x_s(\sum_{l\in\mathcal{L}(s)} \lambda_l^\star)$, which is $w_s/\sum_{l\in\mathcal{L}(s)} \lambda_l^\star$ in the case of TCP Vegas, FAST, and Scalable TCP.

The "physical communications" occur in a network that has a format similar to the one represented in Figure 4.3(a). However, a congestion control protocol establishes direct communications

between the source nodes and the intermediate nodes that manage the respective links. Therefore, the communication network it considers is actually the one represented in Figure 4.3(b). This network is constructed as follows: each link l, which we assume is unidirectional for the sake of simplicity, has a node associated (in Figure 4.3(b), a circle node), and each source s has also a node associated (in Figure 4.3(b), a square node); the recipient nodes are not considered in this new network. If link l is used in the route assigned to source s, the nodes representing link l and source s are connected to each other. Figure 4.3(b) shows the network obtained from Figure 4.3(a) by considering only the links marked with capacities. The way this network is constructed makes it automatically bipartite. In the model considered here, the intermediate node having link l as output manages λ_l. For example, node n_2 in Figure 4.3 manages both λ_2 and λ_3. Note that each communication occurring in the network of Figure 4.3(b) corresponds to an arbitrary number of communications in the original network of Figure 4.3(a). Regarding problem (4.12), it can be written as (P) with the function at node p given by

$$f_p(\lambda_1, \ldots, \lambda_L) = \begin{cases} c_p\lambda_p + \mathrm{i}_{\mathbb{R}^+}(\lambda_p)\,, & \text{if } p \text{ is an intermediate node} \\ \bar{U}_s\left(\sum_{l\in\mathcal{L}(p)} \lambda_l\right), & \text{if } p \text{ is a source node} \end{cases},$$

where $\mathrm{i}_{\mathbb{R}^+}(\cdot)$ is the indicator of the set of the nonnegative real numbers, and the variable is $\lambda = (\lambda_1, \ldots, \lambda_L)$. The variable in this case is star-shaped. The algorithm we propose for (P) can then be used to inspire a new congestion control protocol. However, it has a disadvantage with respect to gradient-based algorithms: while gradient-based algorithms can work with implicit communication, due to their linearity, the algorithm we propose requires explicit communication between each each source and all the intermediate nodes along its route.

Mixed variable model. The NUM problem was introduced as (4.11) to model congestion control in networks. In (4.11), the utility function U_p of source/node p depends only on its sending rate x_p, i.e., $U_p(x_p)$. However, in cooperative or competitive scenarios it might be useful to consider coupled objectives, e.g., $U_p(\{x_l\}_{l\in S_p})$, where S_p is the set of nodes whose rates influence the utility of source p. Such model was considered in [33] (see also [52]). For example, in digital subscriber line (DSL) spectrum management, or in wireless power control, the signal-to-interference ratio at one user depends on the transmit powers of other users, making the scenario competitive. A cooperative scenario would be rate allocation in clusters: the higher the rate allocated to one cluster, the higher the rate allocated to each node inside that cluster. In particular, [33] considered the following variation of (4.11):

$$\begin{array}{ll} \underset{x_1,\ldots,x_P}{\text{maximize}} & \sum_{p=1}^{P} U_p(\{x_l\}_{l\in S_p}) \\ \text{subject to} & \sum_{p=1}^{P} g_p(x_p) \leq c\,, \end{array} \tag{4.13}$$

where each g_p is a convex function and c is a globally known vector. We slightly changed the notation with respect to (4.11): we now denote each source by p and the total number of sources is P. In [33] it is also assumed that source p can communicate with all the sources that interfere with its utility, and vice-versa. The work in [33] proposes a gradient-based algorithm that solves a dual problem of (4.13). To arrive at that dual problem, we first perform a splitting (or cloning) of x_p among all the nodes whose utilities depend on x_p:

$$\begin{array}{ll} \underset{\{\bar{x}_l\}_{l=1}^P}{\text{maximize}} & \sum_{p=1}^P U_p(\{x_l^{(p)}\}_{l\in S_p}) \\ \text{subject to} & \sum_{p=1}^P g_p(x_p^{(p)}) \leq c \\ & x_l^{(i)} = x_l^{(j)}\,, \quad l \in S_i \cap S_j\,,\ (i,j) \in \mathcal{E}\,, \end{array} \tag{4.14}$$

where $x_l^{(p)}$ is the copy of the variable x_l held by node p, and $\mathcal{E}$ is the set of edges in the communication network. The variable in (4.14) is $\{\bar{x}_l\}_{l=1}^P$, where $\bar{x}_l := \{x_l^{(p)}\}_{p\in\mathcal{V}_l}$, and $\mathcal{V}_l$ is the set of nodes whose utilities depend on x_l. Associating a dual variable μ to the first constraint in (4.14) and λ_l^{ij} to each constraint of the second set of constraints, the dual problem of (4.14) is

$$\underset{\mu,\{\lambda_l^{ij}\}}{\text{minimize}} \;\; h_1(\mu, \{\bar{\lambda}^{1j}\}_{j\in\mathcal{N}_1}) + h_2(\mu, \{\bar{\lambda}^{2j}\}_{j\in\mathcal{N}_2}) + \cdots + h_P(\mu, \{\bar{\lambda}^{Pj}\}_{j\in\mathcal{N}_P})\,, \tag{4.15}$$

where the function h_p is associated to source p and is given by

$$h_p(\mu, \{\bar{\lambda}^{pj}\}_{j\in\mathcal{N}_p}) = \sup_{x^{(p)}} \; U_p(\{x_l^{(p)}\}_{l\in S_p}) + \mu^\top g_p(x_p^{(p)}) - \frac{1}{P}\mu^\top c \\ + \sum_{j\in\mathcal{N}_p} \sum_{l\in S_p\cap S_j} \operatorname{sign}(j-p)(\lambda_l^{pj})^\top x_l^{(p)} + \mathrm{i}_{\mathbb{R}^+}(\mu)\,.$$

We used the notation $\bar{\lambda}^{ij} = \{\lambda_l^{ij}\}_{l\in S_i\cap S_j}$. Note that to arrive at (4.15) we used the identity

$$\sum_{(i,j)\in\mathcal{E}} \sum_{l\in S_i\cap S_j} (\lambda_l^{ij})^\top (x_l^{(i)} - x_l^{(j)}) = \sum_{p=1}^P \sum_{j\in\mathcal{N}_p} \sum_{l\in S_p\cap S_j} \operatorname{sign}(j-p)(\lambda_l^{pj})^\top x_l^{(p)}$$

and, with it, we extended the notation λ_l^{ij} for $i > j$ as $\lambda_l^{ij} := -\lambda_l^{ji}$. The variable in (4.15) has the global components μ, appearing in all the functions h_p, and non-global components $\{\lambda_l^{ij}\}$. Thus, (4.15) is a particular instance of (P) with a mixed variable.

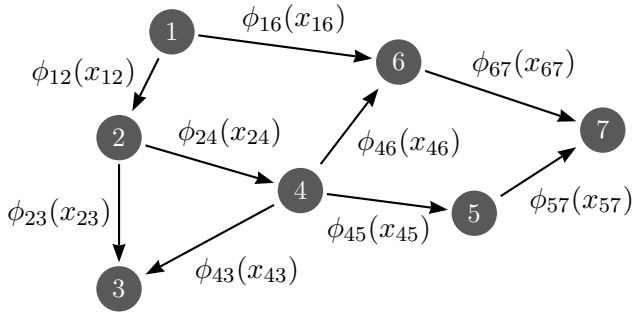


FIGURE 4.4: Example of a network flow problem. Each edge has associated both a variable x_{ij} and function of that variable, $\phi_{ij}(x_{ij})$. The goal is to minimize the sum of all the functions, while satisfying conservation of flow constraints.

4.2.4 Network flow problems

A network flow problem is formulated on a network with arcs $\mathcal{A}$ (or directed edges), where an arc from node i to node j, i.e., $(i,j) \in \mathcal{A}$, indicates a flow in that direction. Figure 4.4 shows an example which allows, for example, a flow from node 1 to node 6, but not from node 6 to node 1. To quantify the flow in an arc $(i,j) \in \mathcal{A}$, we use a non-negative variable x_{ij}. Also, each arc $(i,j) \in \mathcal{A}$ has associated a cost function $\phi_{ij}(x_{ij})$, depending only on x_{ij}, that typically increases with x_{ij}. The goal in network flow problems is to minimize the sum of all these cost functions, while constraining the flows to satisfy conservation laws; namely, the inflows at a given node have to equal the outflows. These inflows/outflows are either caused by neighboring nodes, or are injected/extracted externally at the node itself. A node to which flow is injected (resp. extracted) is called source (resp. sink). For example, in Figure 4.4, if either x_{12} or x_{16} is positive, node 1 can only be a source, since all of its edges point outwards. Nodes 3 and 7, in contrast, can only be sinks, if the flow in their incident arcs is nonzero. Other nodes in that network, for example node 4, can be sources, sinks, or neither. A way to represent a network with flows is via the node-arc incidence matrix B, where the column associated to an arc from node i to node j has a -1 in the ith entry, a 1 in the jth entry, and zeros elsewhere. We assume the components of the variable x and the columns of B are in lexicographic order. For example, $x = (x_{12}, x_{16}, x_{23}, x_{24}, x_{43}, x_{45}, x_{46}, x_{57}, x_{67})$ would be the variable in Figure 4.4. The laws of conservation of flow are expressed as $Bx = d$, where $d \in \mathbb{R}^P$ is the vector of external inputs/outputs. The entries of d sum up to zero and $d_p < 0$ (resp. $d_p > 0$) if node p is a source (resp. sink). When node p is neither a source nor a sink, $d_p = 0$. The problem we solve is

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \sum_{(i,j)\in\mathcal{A}} \phi_{ij}(x_{ij}) \\ \text{subject to} & Bx = d \\ & x \geq 0\,, \end{array} \tag{4.16}$$

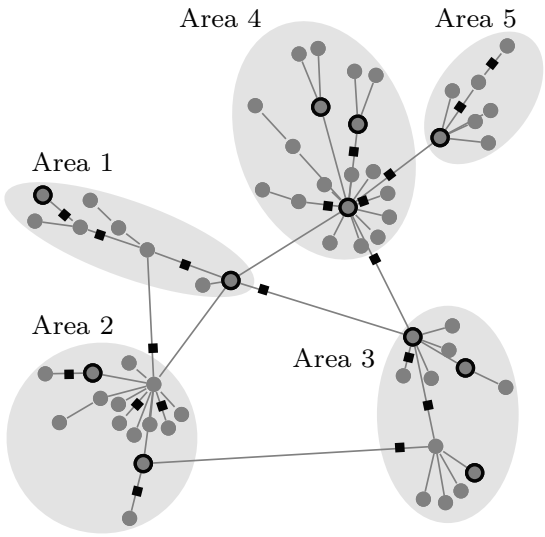


FIGURE 4.5: Illustration of five connected areas in a power network. Nodes represent buses, i.e, generators or loads, and they are connected through transmission lines. Circled nodes indicate voltage measurements and squares in the transmission lines indicate current measurements.

which can be written as (P) by setting

$$f_p\Big(\{x_{pj}\}_{(p,j)\in\mathcal{A}}, \{x_{jp}\}_{(j,p)\in\mathcal{A}}\Big) = \frac{1}{2}\sum_{(p,j)\in\mathcal{A}} \phi_{pj}(x_{pj}) + \frac{1}{2}\sum_{(j,p)\in\mathcal{A}} \phi_{jp}(x_{jp}) + \mathrm{i}_{\{b_p^\top x = d_p\}}\Big(\{x_{pj}\}_{(p,j)\in\mathcal{A}}, \{x_{jp}\}_{(j,p)\in\mathcal{A}}\Big),$$

where $b_p^\top$ is the pth row of B. In words, f_p consists of the sum of the functions associated to all arcs involving node p, plus the indicator function of the set $\{x : b_p^\top x = d_p\}$, which enforces the conservation of flow at node p and only involves the variables $\{x_{pj}\}_{(p,j)\in\mathcal{A}}$ and $\{x_{jp}\}_{(j,p)\in\mathcal{A}}$.

Regarding the communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we assume it consists of the underlying undirected network. This means that nodes i and j can exchange messages directly, i.e., $(i,j) \in \mathcal{E}$ for $i < j$, if there is an arc between these nodes, i.e., $(i,j) \in \mathcal{A}$ or $(j,i) \in \mathcal{A}$. Therefore, in contrast with the flows, messages do not necessarily need to be exchanged satisfying the direction of the arcs. In fact, messages and flows might represent different physical quantities: think, for example, in a network of water pipes controlled by actuators at each pipe junction; while the pipes might enforce a direction in the flow of water (by using, for example, special valves), there is no reason to impose the same constraint on the electrical signals exchanged by the actuators. In problem (4.16), the subgraph induced by x_{ij}, $(i,j) \in \mathcal{A}$, consists only of nodes i and j and an edge connecting them. This makes the variable in (4.16) connected and star-shaped.

4.2.5 State estimation in the power grid

The power grid is the network that connects the energy producers to the energy consumers. Both its large-scale dimensions and the large number of parameters it has make it an important application of distributed optimization. For example, state estimation in the power grid [219] can be posed as a particular instance of (P) with a star-shaped variable, as was done by Kekatos and Giannakis in [47]. In this subsection, we briefly describe the problem from their point of view and derive the algorithm they propose, but adapted to solve (P) for a generic connected variable. As we had mentioned in Chapter 2, the algorithm proposed in [47] is actually the only algorithm we found that can solve (P) for connected variables that are neither global nor stars. Later, in subsection 4.3.2, we will generalize it to solve (P) with a non-connected variable. For other problems and applications of distributed optimization in the power grid, see, for example, [220, 48, 221, 222].

State estimation. To explain state estimation in the power grid, consider the network of Figure 4.5, whose nodes are divided into 5 disjoint areas. The nodes represent either generators or loads or, in the terminology of power systems, buses. Buses are connected with transmission lines, which are represented as the edges of the network and through which current flows. Each area in the network, although controlled by its own operator, is also connected to other areas for robustness and reliability. It is essential for the proper functioning of the power network to know or, at least, to estimate the state of the system; this includes knowing power flows, voltage and current magnitudes at the buses, and generator outputs. Figure 4.5 illustrates a typical scenario where circled nodes indicate buses at which voltage measurements are taken, and edges with squares indicate lines where current measurements are taken. Based on these measurements and on a model for the system, the goal of state estimation is to determine what are the voltages and currents at the other buses and lines of the network. Let us denote the state of the entire network, i.e., the set of all the voltages and all the currents, with $x \in \mathbb{R}^n$. The state of a given area p is a set $S_p \in \{1, \dots, n\}$ of $n_p = |S_p|$ components of x, i.e., x_{S_p} denotes the state of area p. Let us represent the set of measurements taken at this area with $y_p \in \mathbb{R}^{m_p}$. These measurements are related with x_{S_p} through $y_p = h_p(x_{S_p}) + w_p$, where $h_p : \mathbb{R}^{n_p} \to \mathbb{R}^{m_p}$ models area p and is typically a nonlinear function, and $w_p \in \mathbb{R}^{m_p}$ models measurement noise and model inaccuracies. The areas that are connected with transmission lines will share some state variables, i.e., $S_i \cap S_j$. The problem of state estimation in power systems can then be formulated as

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \sum_{p=1}^{P} \left\| y_p - h_p(x_{S_p}) \right\|^2 . \tag{4.17}$$

Since each function h_p is nonlinear, problem (4.17) is nonconvex. Yet, as mentioned in [47], either using Gauss-Newton methods to solve (4.17) directly or using a DC-approximation model for the system, one usually ends up with a linearized version of the system, i.e., $y_p = H_p x_{S_p} + w_p$, where $H_p \in \mathbb{R}^{m_p \times n_p}$ is the Jacobian of h_p at some nominal operating point. So, instead of solving the nonconvex problem (4.17), we can solve

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \sum_{p=1}^{P} \left\| y_p - H_p x_{S_p} \right\|^2 , \tag{4.18}$$

which is convex. If we view each area in Figure 4.5 as a node of a network, then (4.18) (and also (4.17)) have the format of (P), where each f_p is given by $f_p(x_{S_p}) = (1/2)\|y_p - H_p x_{S_p}\|^2$. In this case, the variable is star-shaped because each set S_p indexes components of the state of area p and, possibly, of areas adjacent (i.e., neighbors) of area p.

The algorithm in [47]. The algorithm proposed in [47] solves (4.18) in a distributed way and assumes that neighboring areas communicate, i.e., that they are able to exchange estimates of their common state variables. However, that algorithm can be easily generalized to solve (P) under a generic connected variable, i.e., all induced subgraphs are connected. The algorithm is presented as Algorithm 4 and is derived in Appendix C.

Algorithm 4 [47]

Initialization: Choose $\rho \in \mathbb{R}$; for all $p \in \mathcal{V}$ and $l \in S_p$, set $\gamma_l^{(p),0} = x_l^{(p),0} = 0$; set $k = 0$

1: **repeat**
2: **for all** $p \in \mathcal{V}$ [in parallel] **do**
3: Compute $v_l^{(p),k} = \gamma_l^{(p),k} - \frac{\rho}{2}\Big(D_{p,l}\, x_l^{(p),k} + \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} x_l^{(j),k} \Big)$, for all $l \in S_p$
4: Compute $x_{S_p}^{(p),k+1} = \underset{x_{S_p}^{(p)} = \{x_l^{(p)}\}_{l \in S_p}}{\arg\min} \ f_p(x_{S_p}^{(p)}) + \sum_{l \in S_p} {v_l^{(p),k}}^{\top} x_l^{(p)} + \frac{\rho}{2} \sum_{l \in S_p} D_{p,l} \big(x_l^{(p)}\big)^2$
5: For each component $l \in S_p$, exchange $x_l^{(p),k+1}$ with neighbors $\mathcal{N}_p \cap \mathcal{V}_l$
6: Update the dual variables $\gamma_l^{(p),k+1} = \gamma_l^{(p),k} + \frac{\rho}{2} \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} (x_l^{(p),k+1} - x_l^{(j),k+1})$, for all $l \in S_p$
7: $k \leftarrow k + 1$
8: **end for**
9: **until** some stopping criterion is met

Algorithm 4 solves (P) by first reformulating it as

$$\begin{array}{ll} \underset{\{\bar{x}_l\}_{l=1}^n}{\text{minimize}} & f_1(x_{S_1}^{(1)}) + f_1(x_{S_2}^{(2)}) + \cdots + f_1(x_{S_P}^{(P)}) \\ \text{subject to} & x_l^{(p)} = x_l^{(j)} \,, \quad l \in S_p \cap S_j \,, \quad j \in \mathcal{N}_p \,, \quad p = 1, \ldots, P \,, \end{array} \tag{4.19}$$

where we created a copy of the component x_l in all the nodes whose functions depend on x_l, i.e., on all $p \in \mathcal{V}_l$, where $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$ is the subgraph induced by x_l. The copy at node p is $x_l^{(p)}$. All the copies held by node p are denoted with $x_{S_p}^{(p)} := \{x_l^{(p)}\}_{l \in S_p}$. The constraints in (4.19) enforce copies of the component x_l to be equal for neighboring nodes that depend on it. That is, if nodes i and j are neighbors, $(i,j) \in \mathcal{E}$, and both depend on x_l, $l \in S_i$ and $l \in S_j$, then $x_l^{(i)} = x_l^{(j)}$ will be on the constraints of (4.19). Since we assume a connected variable, problems (P) and (4.19) are equivalent. We also denote by $\bar{x}_l$ the set of all copies of the component x_l, i.e., $\bar{x}_l = \{x_l^{(p)}\}_{p \in \mathcal{V}_l}$. Similarly to what was done for Algorithm 2, [47] introduces a variable per network edge and writes (4.19) equivalently as

$$\begin{array}{ll} \underset{\{\bar{x}_l\}_{l=1}^n, \{\bar{z}_l\}_{l=1}^n}{\text{minimize}} & f_1(x_{S_1}^{(1)}) + f_1(x_{S_2}^{(2)}) + \cdots + f_1(x_{S_P}^{(P)}) \\ \text{subject to} & x_l^{(p)} = z_l^{\{p,j\}}\,, \quad l \in S_p \cap S_j\,, \quad j \in \mathcal{N}_p\,, \quad p = 1, \ldots, P\,, \end{array} \tag{4.20}$$

where $z_l^{\{i,j\}} = z_l^{\{j,i\}}$ is associated to the lth common component between nodes i and j, for $(i,j) \in \mathcal{E}_l$. We used $\bar{z}_l$ to denote the set of variables $z_l^{\{i,j\}}$ associated to the component x_l, i.e., $\bar{z}_l = \{z_l^{\{i,j\}}\}_{(i,j) \in \mathcal{E}_l}$. Problem (4.20) has two sets of variables, $\{\bar{x}_l\}_{l=1}^n$ and $\{\bar{z}_l\}_{l=1}^n$, and linear constraints. Therefore, the 2-block ADMM (2.18)-(2.20) can be applied and yields Algorithm 4, as shown in Appendix C. Algorithm 4 has a structure very similar to Algorithm 2; indeed, it is derived using the same principles, but adapted to the problem (P). In particular, all nodes perform the same tasks in parallel. These tasks consist of solving an optimization problem in step 4 and sending components of the respective solution to the neighbors that have common components, in step 5. We used $D_{p,l}$ to denote the degree of node p in the subgraph induced by component x_l, $\mathcal{G}_l$. Steps 3, 4, and 5 in Algorithm 4 correspond to step 3 of Algorithm 2: now, however, the prox notation is not as convenient as it was for the global class algorithms. Also, if in Algorithm 2 each node broadcasts all the components of its new update, in Algorithm 4 each node p needs only to transmit to its neighbor $j \in \mathcal{N}_p$ their common components $x_{S_p \cap S_j}$. After these exchanges occur, node p can update its set of dual variables γ_l, $l \in \mathcal{S}_p$, as in step 6.

The algorithm we propose for (P) relates to the algorithm we proposed for the global class (G) in the same way that Algorithm 4 relates to Algorithm 2. We will derive it in the next section, first for a connected variable, and then for a general variable, connected or not. Similarly to the algorithms for the global class, our algorithm outperforms Algorithm 4 in terms of the number of communications, as will be observed in Section 4.4.

4.3 Algorithm derivation

In this section, we derive our algorithm for (P). First, we consider a connected variable, i.e., every induced subgraph is connected, and then we propose a way to address a non-connected variable, i.e., when there is at least one induced subgraph that is non-connected.

4.3.1 Connected variable

The main idea to derive an algorithm for (P) when the variable is connected is the same idea we used to derive Algorithm 3 for the global class: we manipulate (P) to make the multi-block ADMM (2.22)-(2.26) applicable. The difference is in the way we manipulate the problem, more specifically, in how we create copies of the variables. Recall our notation on the coloring scheme: $\mathcal{C}_c \subset \mathcal{V}$ denotes the nodes that have color c and $\mathcal{C}(p)$ denotes the color of node p; also, $C_c = |\mathcal{C}_c|$ is the number of nodes with color c. As in the derivation of the global class algorithm, we will assume, without loss of generality, that the nodes are numbered according to their colors: the first C_1 nodes have color 1, $\mathcal{C}_1 = \{1, \ldots, C_1\}$, the next C_2 nodes have color 2, $\mathcal{C}_2 = \{C_1 + 1, \ldots, C_1 + C_2\}$, and so on.

Problem manipulation. Recall that $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$ denotes the subgraph induced by the component x_l. In this subsection, we assume each $\mathcal{G}_l$ is connected. Similarly to [47], we create a copy of the component x_l only in the nodes that are interested in it, which are precisely the nodes in $\mathcal{G}_l$; let $x_l^{(p)}$ be the copy at node p. Since a given node p depends on the components x_{S_p} of the variable x, it will have $|S_p|$ copies; let $x_{S_p}^{(p)} := \{x_l^{(p)}\}_{l \in S_p}$ be the set of all these copies, at node p. We now rewrite (P) in a way slightly different than (4.19):

$$\begin{array}{ll} \underset{\{\bar{x}_l\}_{l=1}^n}{\text{minimize}} & f_1(x_{S_1}^{(1)}) + f_2(x_{S_2}^{(2)}) + \cdots + f_P(x_{S_P}^{(P)}) \\ \text{subject to} & x_l^{(i)} = x_l^{(j)}\,, \quad (i,j) \in \mathcal{E}_l\,, \;\; l = 1, \ldots, n\,, \end{array} \tag{4.21}$$

where the optimization variable is $\{\bar{x}_l\}_{l=1}^L$ and it represents the set of all copies. We used $\bar{x}_l$ to denote all copies of the component x_l, which are located only in the nodes of $\mathcal{G}_l$: $\bar{x}_l := \{x_l^{(p)}\}_{p \in \mathcal{V}_l}$. Although problems (4.19) and (4.21) are both equivalent to (P), (4.19) has twice the constraints of (4.21). While in (4.19) each constraint appears twice (to make the introduction of the z's in (4.20) possible), our reformulation (4.21) uses the minimal number of constraints. Recall that our notation $(i,j) \in \mathcal{E}_l$ (or $\mathcal{E}$) implies that $i < j$, and therefore there are no repeated equations in (4.21). Finally, note that our assumption that the variable is connected is what makes problems (P) and (4.21) equivalent, since each induced subgraph is connected. When the variable is non-connected, this equivalence no longer holds.

Let A_l denote the transpose of the node-arc incidence matrix of the subgraph $\mathcal{G}_l$. Then, the constraint $x_l^{(i)} = x_l^{(j)}$, $(i,j) \in \mathcal{E}_l$ can be written as $A_l \bar{x}_l = 0$. We now use the coloring scheme (cf. Assumption 4.5) to partition each variable $\bar{x}_l$ as $\bar{x}_l = (\bar{x}_l^1, \ldots, \bar{x}_l^C)$, where

$$\bar{x}_l^c = \begin{cases} \{x_l^{(p)}\}_{p \in \mathcal{V}_l \cap \mathcal{C}_c}, & \text{if } \mathcal{V}_l \cap \mathcal{C}_c \neq \emptyset \\ \emptyset, & \text{if } \mathcal{V}_l \cap \mathcal{C}_c = \emptyset \end{cases}.$$

Recall that $\mathcal{C}_c$ is the set of nodes that have color c. In words, $\bar{x}_l^c$ represents the set of copies of x_l held by the nodes that have color c. If no node with color c depends on x_l, then $\bar{x}_l^c$ is empty. Using a similar notation for the columns of the matrix A_l, we write $A_l \bar{x}_l$ as $\bar{A}_l^1 \bar{x}_l^1 + \cdots + \bar{A}_l^C \bar{x}_l^C$, for all l. Therefore, (4.21) is equivalent to

$$\begin{array}{ll} \underset{\bar{x}^1, \ldots, \bar{x}^C}{\text{minimize}} & \sum_{p \in \mathcal{C}_1} f_p(x_{S_p}^{(p)}) + \cdots + \sum_{p \in \mathcal{C}_C} f_p(x_{S_p}^{(p)}) \\ \text{subject to} & \bar{A}^1 \bar{x}^1 + \cdots + \bar{A}^C \bar{x}^C = 0\,, \end{array} \tag{4.22}$$

where $\bar{x}^c = \{\bar{x}_l^c\}_{l=1}^n$, and $\bar{A}^c$ is the diagonal concatenation of the matrices $\bar{A}_1^c$, $\bar{A}_2^c$, $\ldots$, $\bar{A}_n^c$, i.e., $\bar{A}^c = \mathrm{diag}(\bar{A}_1^c, \bar{A}_2^c, \ldots, \bar{A}_n^c)$. For better visualization, we wrote the constraint in (4.22) as

$$\underbrace{\begin{bmatrix} \bar{A}_1^1 & & & \\ & \bar{A}_2^1 & & \\ & & \ddots & \\ & & & \bar{A}_n^1 \end{bmatrix}}_{\bar{A}^1} \underbrace{\begin{bmatrix} \bar{x}_1^1 \\ \bar{x}_2^1 \\ \vdots \\ \bar{x}_n^1 \end{bmatrix}}_{\bar{x}^1} + \underbrace{\begin{bmatrix} \bar{A}_1^2 & & & \\ & \bar{A}_2^2 & & \\ & & \ddots & \\ & & & \bar{A}_n^2 \end{bmatrix}}_{\bar{A}^2} \underbrace{\begin{bmatrix} \bar{x}_1^2 \\ \bar{x}_2^2 \\ \vdots \\ \bar{x}_n^2 \end{bmatrix}}_{\bar{x}^2} + \cdots + \underbrace{\begin{bmatrix} \bar{A}_1^C & & & \\ & \bar{A}_2^C & & \\ & & \ddots & \\ & & & \bar{A}_n^C \end{bmatrix}}_{\bar{A}^C} \underbrace{\begin{bmatrix} \bar{x}_1^C \\ \bar{x}_2^C \\ \vdots \\ \bar{x}_n^C \end{bmatrix}}_{\bar{x}^C} = 0\,. \tag{4.23}$$

Note that the cth term in the objective of (4.22) depends only on $\bar{x}^c$, the set of copies associated with nodes with color c. Thus, (4.22) has the format of (2.21), the problem solved by the multi-block ADMM, and thus the iterations (2.22)-(2.26) can be applied.

Applying multi-block ADMM. To apply the multi-block ADMM iterations (2.22)-(2.26) to (4.22), we first need to write the augmented Lagrangian. Let λ_l^{ij} be the dual variable associated to the constraint $x_l^{(i)} = x_l^{(j)}$, for some $l \in \{1, \ldots, n\}$ and $(i,j) \in \mathcal{E}_l$ (cf. (4.21)). The augmented Lagrangian of (4.22) is then

$$L_\rho(\bar{x}^1, \ldots, \bar{x}^C; \lambda) = \sum_{c=1}^{C} \sum_{p \in \mathcal{C}_c} f_p(x_{S_p}^{(p)}) + \sum_{c=1}^{C} \lambda^\top \bar{A}^c \bar{x}^c + \frac{\rho}{2} \Big\| \sum_{c=1}^{C} \bar{A}^c \bar{x}^c \Big\|^2, \tag{4.24}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$ is the dual variable, whose lth block is $\lambda_l := \{\lambda_l^{ij}\}_{(i,j) \in \mathcal{E}_l}$. The multi-block ADMM consists of a sequence of subproblems, obtained by minimizing L_ρ with respect to each

block $\bar{x}^c$, and then updating each dual variable with

$$\lambda_l^{ij,k+1} = \lambda_l^{ij,k} + \rho(x_l^{(i),k+1} - x_l^{(j),k+1}), \tag{4.25}$$

for every $(i,j) \in \mathcal{E}_l$ and every $l = 1, \ldots, n$. In (4.25), k denotes the iteration number and $x_l^{(p),k+1}$ is the estimate of the component x_l, by node p, after iteration k. We now analyze the subproblem each node solves to find those estimates. In particular, we will see that minimizing (4.24) with respect to $\bar{x}^c$ yields $|\mathcal{C}_c|$ problems that can be solved in parallel, i.e., all nodes with color c "work" in parallel. For example, the copies of the nodes with color 1 are updated according to (2.22):

$$\bar{x}^{1,k+1} = \arg\min_{\bar{x}^1} \sum_{p\in\mathcal{C}_1} f_p(x_{S_p}^{(p)}) + {\lambda^k}^\top \bar{A}^1 \bar{x}^1 + \frac{\rho}{2}\left\| \bar{A}^1 \bar{x}^1 + \sum_{c=2}^{C} \bar{A}^c \bar{x}^{c,k} \right\|^2 \tag{4.26}$$

$$= \arg\min_{\bar{x}^1} \sum_{p\in\mathcal{C}_1} \left(f_p(x_{S_p}^{(p)}) + \sum_{l\in S_p} \sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l} \left(\lambda_l^{pj,k} - \rho\, x_l^{(j),k}\right)^\top x_l^{(p)} + \frac{\rho}{2} \sum_{l\in S_p} D_{p,l} \left(x_l^{(p)}\right)^2 \right), \tag{4.27}$$

whose equivalence is established in Lemma 4.6 below. As in Algorithm 4, $D_{p,l}$ is the degree of node p in the subgraph $\mathcal{G}_l$, i.e., the number of neighbors of node p that also depend on x_l. Of course, $D_{p,l}$ is only defined when $l \in S_p$. Note that all the dual variables λ_l^{pj} are well-defined because of our assumption that the nodes are numbered according to their colors; namely, any neighbor j of a node $p \in \mathcal{C}_1$ will have a color larger than 1, and hence $p < j$, making λ_l^{pj} well-defined. Recall our convention that $(i,j) \in \mathcal{E}$ implies $i < j$. Before we establish the equivalence between (4.26) and (4.27), note that (4.27) actually consists of $|\mathcal{C}_1|$ problems that can be solved in parallel. This is because nodes with the same color are not neighbors and, thus, none of the components of the optimization variable $\bar{x}^1$, which corresponds to all the copies of the nodes with color 1, appears as $x_l^{(j),k}$ in the second term of (4.27). This means that all nodes $p \in \mathcal{C}_1$ solve, in parallel,

$$x_{S_p}^{(p),k+1} = \arg\min_{x_{S_p}^{(p)} = \{x_l^{(p)}\}_{l\in S_p}} f_p(x_{S_p}^{(p)}) + \sum_{l\in S_p} \sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l} (\lambda_l^{pj,k} - \rho\, x_l^{(j),k})^\top x_l^{(p)} + \frac{\rho}{2} \sum_{l\in S_p} D_{p,l} \left(x_l^{(p)}\right)^2. \tag{4.28}$$

Node p can only solve (4.28) if it knows $x_l^{(j),k}$ and $\lambda_l^{pj,k}$, for $j \in \mathcal{N}_p \cap \mathcal{V}_l$ and $l \in S_p$. This is possible if, in the previous iteration, it received the respective copies of x_l from its neighbors. This is also enough for knowing $\lambda_l^{pj,k}$, although we will see later that no node needs to know each $\lambda_l^{pj,k}$ individually. We finally show how to obtain (4.27) from (4.26).

Lemma 4.6. (4.26) *and* (4.27) *are equivalent.*

Proof.

To go from (4.26) to (4.27), we first develop the last two terms of (4.26), respectively,

$$\lambda^{k\top} \bar{A}^1 \bar{x}^1 \tag{4.29}$$

and

$$\frac{\rho}{2}\left\| \bar{A}^1 \bar{x}^1 + \sum_{c=2}^{C} \bar{A}^c \bar{x}^{c,k} \right\|^2 . \tag{4.30}$$

We first address (4.29). Given the structure of $\bar{A}^1$, as seen in (4.23), we can write (4.29) as $\sum_{l=1}^{n}((\bar{A}_l^1)^\top \lambda_l^k)^\top \bar{x}_l^1$. Recall that $(\bar{A}_l^1)^\top$, if it exists (i.e., if there is a node with color 1 that depends on component x_l), consists of the block of rows of the node-arc incidence matrix of $\mathcal{G}_l$ corresponding to the nodes with color 1. Therefore, if there exists $p \in \mathcal{C}_1 \cap \mathcal{V}_l$, the vector $(\bar{A}_l^1)^\top \lambda_l^k$ will have an entry $\sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} \operatorname{sign}(j-p)\lambda_l^{pj,k}$. The sign function appears here because the column of the node-arc incidence matrix corresponding to $x_l^{(i)} - x_l^{(j)} = 0$, for a pair $(i,j) \in \mathcal{E}_l$, contains 1 in the ith entry and -1 in the jth entry, where $i < j$. In the previous expression, we used an extension of the definition of λ_l^{ij}, which was only defined for $i < j$ (due to our convention that for any edge $(i,j) \in \mathcal{E}$ we have always $i < j$). Assume λ_l^{ij} is initialized with zero; switching i and j in (4.25), we obtain $\lambda_l^{ji,k} = -\lambda_l^{ij,k}$, which holds for all iterations k. To be consistent with the previous equation, we define λ_l^{ij} as $\lambda_l^{ij} := -\lambda_l^{ji}$ whenever $i > j$. Therefore, (4.29) develops as

$$\begin{aligned}
\lambda^{k\top} \bar{A}^1 \bar{x}^1 &= \sum_{l=1}^{n} ((\bar{A}_l^1)^\top \lambda_l^k)^\top \bar{x}_l^1 \\
&= \sum_{l=1}^{n} \sum_{p \in \mathcal{C}_1} \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} \operatorname{sign}(j-p) \left(\lambda_l^{pj,k}\right)^\top x_l^{(p)} \\
&= \sum_{p \in \mathcal{C}_1} \sum_{l=1}^{n} \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} \operatorname{sign}(j-p) \left(\lambda_l^{pj,k}\right)^\top x_l^{(p)} .
\end{aligned} \tag{4.31}$$

Regarding (4.30), it can be written as

$$\frac{\rho}{2}\left\| \bar{A}^1 \bar{x}^1 + \sum_{c=2}^{C} \bar{A}^c \bar{x}^{c,k} \right\|^2 = \frac{\rho}{2}\left\| \bar{A}^1 \bar{x}^1 \right\|^2 + \rho (\bar{A}^1 \bar{x}^1)^\top \sum_{c=2}^{C} \bar{A}^c \bar{x}^{c,k} + \frac{\rho}{2}\left\| \sum_{c=2}^{C} \bar{A}^c \bar{x}^{c,k} \right\|^2 . \tag{4.32}$$

Since the last term does not depend on $\bar{x}^1$, it can be dropped from the optimization problem. We now use the structure of $\bar{A}^1$ to rewrite the first term of (4.32):

$$\frac{\rho}{2}\left\| \bar{A}^1 \bar{x}^1 \right\|^2 = \frac{\rho}{2} \sum_{l=1}^{n} (\bar{x}_l^1)^\top (\bar{A}_l^1)^\top \bar{A}_l^1 \bar{x}_l^1 \tag{4.33}$$

$$= \frac{\rho}{2} \sum_{l=1}^{n} \sum_{p \in \mathcal{C}_1} D_{p,l} \left(x_l^{(p)} \right)^2 \tag{4.34}$$

$$= \frac{\rho}{2} \sum_{p \in \mathcal{C}_1} \sum_{l \in S_p} D_{p,l} \left(x_l^{(p)} \right)^2 . \tag{4.35}$$

From (4.33) to (4.34) we used the structure of $\bar{A}_l^1$. Namely, if it exists, $(\bar{A}_l^1)^\top \bar{A}_l^1$ is a diagonal matrix, where each diagonal entry is extracted from the diagonal of $A_l^\top A_l$, the Laplacian matrix for $\mathcal{G}_l$. Since each entry in the diagonal of a Laplacian matrix contains the degrees of the respective nodes, the diagonal of $(\bar{A}_l^1)^\top \bar{A}_l^1$ contains $D_{p,l}$ for all $p \in \mathcal{C}_1$. The reason why $(\bar{A}_l^1)^\top \bar{A}_l^1$ is diagonal is because nodes with the same color are never neighbors. As in (4.32), we exchanged the order of the summations from (4.34) to (4.35).

Finally, we develop the second term of (4.32):

$$\rho (\bar{A}^1 \bar{x}^1)^\top \sum_{c=2}^{C} \bar{A}^c \bar{x}^{c,k} = \rho \sum_{c=2}^{C} \sum_{l=1}^{n} (\bar{x}_l^1)^\top (\bar{A}_l^1)^\top (\bar{A}_l^c)\, \bar{x}_l^{c,k} \tag{4.36}$$

$$= -\rho \sum_{c=2}^{C} \sum_{l=1}^{n} \sum_{p \in \mathcal{C}_1} \sum_{j \in \mathcal{N}_p \cap \mathcal{C}_c \cap \mathcal{V}_l} {x_l^{(p)}}^\top x_l^{(j),k} \tag{4.37}$$

$$= -\rho \sum_{p \in \mathcal{C}_1} \sum_{l \in S_p} {x_l^{(p)}}^\top \sum_{c=2}^{C} \sum_{j \in \mathcal{N}_p \cap \mathcal{C}_c \cap \mathcal{V}_l} x_l^{(j),k} \tag{4.38}$$

$$= -\rho \sum_{p \in \mathcal{C}_1} \sum_{l \in S_p} \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} {x_l^{(p)}}^\top x_l^{(j),k} . \tag{4.39}$$

In (4.36) we just used the structure of $\bar{A}^1$ and $\bar{A}^c$, as visualized in (4.23). From (4.36) to (4.37) we used the fact that $(\bar{A}_l^1)^\top \bar{A}_l^c$ is a submatrix of $A_l^\top A_l$, the Laplacian of $\mathcal{G}_l$, containing some of its off-diagonal elements. More concretely, $(\bar{A}_l^1)^\top \bar{A}_l^c$ contains the entries of $A_l^\top A_l$ corresponding to all the nodes $i \in \mathcal{C}_1 \cap \mathcal{V}_l$ and $j \in \mathcal{C}_c \cap \mathcal{V}_l$. And, for such nodes, the corresponding entry in $A_l^\top A_l$ is -1 if i and j are neighbors, and 0 otherwise. From (4.38) to (4.39) we just used the fact that the set $\{\mathcal{C}_c\}_{c=2}^{C}$ is nothing but a partition of the set of neighbors of any node with color 1. Using (4.31), (4.32), (4.35), and (4.39) in (4.26), we get (4.27). □

The optimization problem (4.27) decomposes into $|\mathcal{C}_1|$ decoupled optimization problems, each one solved by a node with color 1. For node p, the problem is (4.28). For the other colors, the same reasoning and equations apply, just with one small difference: in the second term of (4.28) we have $x_l^{(j),k+1}$ from the neighbors with a smaller color and $x_l^{(j),k}$ from the nodes with a larger color.

Algorithm 5 shows the resulting algorithm. As in the global class algorithm (Algorithm 3), the coloring scheme functions as a schedule: the nodes with color 1 work first, the nodes with color 2

Algorithm 5 Algorithm for a connected variable

Initialization: Choose $\rho \in \mathbb{R}$; for all $p \in \mathcal{V}$ and $l \in S_p$, set $\gamma_l^{(p),0} = x_l^{(p),0} = 0$; set $k = 1$

1: **repeat**
2: **for** $c = 1, \ldots, C$ **do**
3: **for all** $p \in \mathcal{C}_c$ [in parallel] **do**
4: Compute $v_l^{(p),k} = \gamma_l^{(p),k} - \rho \sum_{\substack{j \in \mathcal{N}_p \cap \mathcal{V}_l \\ C(j) < c}} x_l^{(j),k+1} - \rho \sum_{\substack{j \in \mathcal{N}_p \cap \mathcal{V}_l \\ C(j) > c}} x_l^{(j),k}$, for all $l \in S_p$
5: Compute $x_{S_p}^{(p),k+1} = \underset{x_{S_p}^{(p)} = \{x_l^{(p)}\}_{l \in S_p}}{\arg\min} f_p(x_{S_p}^{(p)}) + \sum_{l \in S_p} {v_l^{(p),k}}^\top x_l^{(p)} + \frac{\rho}{2} \sum_{l \in S_p} D_{p,l}\big(x_l^{(p)}\big)^2$
6: For each component $l \in S_p$, exchange $x_l^{(p),k+1}$ with neighbors $\mathcal{N}_p \cap \mathcal{V}_l$
7: **end for**
8: **end for**
9: **for all** $p \in \mathcal{V}$ and $l \in S_p$ [in parallel] **do**
$\gamma_l^{(p),k+1} = \gamma_l^{(p),k} + \rho \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} (x_l^{(p),k+1} - x_l^{(j),k+1})$
10: **end for**
11: $k \leftarrow k + 1$
12: **until** some stopping criterion is met

work next, and so on. Each "work" consists of computing $v_l^{(p),k}$ for all $l \in S_p$, as in step 4, solving the optimization problem in step 5, and then send the new component estimates to the neighbors that also depend on those components, as in step 6. After a given node p has received the new estimates from all its neighbors, it can update each dual variable $\gamma_l^{(p)}$ as in step 9. Note that the edge-wise dual variables λ_l^{ij} were replaced by the node-wise dual variables $\gamma_l^{(p)}$. The reason is because the optimization solved by node p (see (4.28)) depends on $\gamma_l^{(p),k} := \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} \operatorname{sign}(j - p) \lambda_l^{pj,k}$ and not on the individual λ_l^{ij}'s. The update of step 9 is obtained by replacing

$$\lambda_l^{ij,k+1} = \lambda_l^{ij,k} + \rho \operatorname{sign}(j - i)(x_l^{(i),k+1} - x_l^{(j),k+1}) \tag{4.40}$$

in the definition of $\gamma_l^{(p),k}$. Note that (4.40) differs from (4.25) in the extra "sign." This is because we extended the definition of the dual variable λ_l^{ij} for $i > j$ (see the proof of Lemma 4.6).

Note that if we make the variable global, i.e., if $S_p = \{1, \ldots, n\}$ for all p, Algorithm 5 becomes the global class algorithm, that is, Algorithm 3. This means that Algorithm 5 is a generalization of Algorithm 3 since, in fact, it cannot be obtained from it. The comments about the coordination of the nodes we made for Algorithm 3 also apply to its generalization, Algorithm 5. Namely, if each node knows its own color and the color of its neighbors, as specified in Assumption 4.5, then the algorithm becomes automatically distributed, because each node can work immediately after it has received estimates from its neighbors with smaller colors. See Figure 3.2 from Chapter 3 for an illustration. Regarding the convergence of Algorithm 5, we have:

Theorem 4.7.

Let Assumptions 4.1-4.5 hold and let the variable of

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f_1(x_{S_1}) + f_2(x_{S_2}) + \cdots + f_P(x_{S_P})\,, \tag{P}$$

be connected. Then, Algorithm 5 produces a sequence $(x^k_{S_1}, \ldots, x^k_{S_P})$ *convergent to* $(x^\star_{S_1}, \ldots, x^\star_{S_P})$*, where* $x^\star$ *solves* (P)*, when at least one of the following conditions is satisfied:*

(a) the coloring scheme uses two colors only (which implies that the network is bipartite);

(b) each function f_p *is strongly convex with modulus* μ_p *and*

$$0 < \rho < \min_{c=1,\ldots,C} \frac{2\sum_{p\in\mathcal{C}_c}\mu_p}{3\,(C-1)\,\max_{p\in\mathcal{C}_c,\, l\in S_p} D_{p,l}}\,. \tag{4.41}$$

Proof.

As in the proof of Theorem 3.6, we have to show that (4.22), which is the problem to which we apply the multi-block ADMM, satisfies the conditions of Theorem 2.1. In fact, Assumptions 4.1, 4.2, and 4.3, together with the equivalence between (P) and (4.22) (for a connected variable), imply that each function $\sum_{p\in\mathcal{C}_c} f_p(x^{(p)}_{S_p})$ in (4.22) is closed and convex over the full space. Next we see that condition (a) (resp. (b)) implies condition (a) (resp. (b)) of Theorem 2.1.

(a) We first see that Assumption 4.4 together with the fact that the variable is connected implies that each $\bar{A}^c$ has full column rank. Let c be any color in $\{1, 2, \ldots, C\}$. By definition, $\bar{A}^c = \operatorname{diag}(\bar{A}^c_1, \bar{A}^c_2, \ldots, \bar{A}^c_n)$; therefore, we have to prove that each $\bar{A}^c_l$ has full column rank, for $l = 1, 2, \ldots, n$. Let then c and l be fixed. We are going to prove that $(\bar{A}^c_l)^\top \bar{A}^c_l$, a square matrix, has full rank, and therefore $\bar{A}^c_l$ has full column rank. Since $\bar{A}_l = \begin{bmatrix}\bar{A}^c_1 & \bar{A}^c_2 & \cdots & \bar{A}^c_n\end{bmatrix}$, $(\bar{A}^c_l)^\top \bar{A}^c_l$ corresponds to the lth block in the diagonal of the matrix $A_l^\top A_l$, the Laplacian matrix of the induced subgraph $\mathcal{G}_l$. Recall that each induced subgraph $\mathcal{G}_l$ is connected, because the variable is connected. Consequently, each node in $\mathcal{G}_l$ has at least one neighbor also in $\mathcal{G}_l$ and hence each entry in the diagonal of $A_l^\top A_l$ is greater than zero.[1] The same happens to the entries in the diagonal of $(\bar{A}^c_l)^\top \bar{A}^c_l$. In fact, these are the only nonzero entries of $(\bar{A}^c_l)^\top \bar{A}^c_l$, since this matrix is diagonal. The reason is because $(\bar{A}^c_l)^\top \bar{A}^c_l$ corresponds to the Laplacian entries of nodes that have the same color, which are never neighbors. Therefore, $(\bar{A}^c_l)^\top \bar{A}^c_l$ has full rank. This shows that, independently of the coloring scheme, each matrix $\bar{A}_c$ has full

[1] Implicitly, we are assuming that there is no component x_l that appears in only one node, say node p; this would lead to a Laplacian matrix $A_l^\top A_l$ equal to 0. This can be easily addressed by redefining f_p, the function at node p, to $\tilde{f}_p(\cdot) = \inf_{x_l} f_p(\ldots, x_l, \ldots)$.

column rank. As a consequence, when the network is bipartite and the coloring scheme has two colors, point (a) of Theorem 2.1 holds.

(b) When each function f_p is strongly convex with modulus μ_p and ρ satisfies (4.41), then $\sum_{p\in\mathcal{C}_c} f_p$ is strongly convex with modulus $\sum_{p\in\mathcal{C}_c}\mu_p$ [31, Lem. 2.1.4] and conditions (2.27) and (4.41) are equivalent. To see this, note that

$$\sigma_{\max}(\bar{A}^c)^2 = \lambda_{\max}((\bar{A}^c)^\top \bar{A}^c) = \max_{l=1,\dots,n} \lambda_{\max}((\bar{A}_l^c)^\top \bar{A}_l^c) = \max_{l=1,\dots,n,\, p\in\mathcal{C}_c} D_{p,l} = \max_{p\in\mathcal{C}_c,\, l\in S_p} D_{p,l}\,,$$

since each $(\bar{A}_l^c)^\top \bar{A}_l^c$ is a diagonal matrix whose entries are the degrees of the nodes with color c that depend on component x_l.

□

4.3.2 Non-connected variable

In this subsection we drop the assumption that the variable is connected. This means that there exists at least one component x_l for which the induced subgraph $\mathcal{G}_l$ is non-connected. In this case, problem (4.21) is no longer equivalent to problem (P), because its constraints fail to enforce equality between all the copies of x_l. We propose a trick to make these problems equivalent, based on the following assumption:

Assumption 4.8. *When the variable is non-connected, the communication network and all the sets S_p are known before the execution of the algorithm.*

The reason we require both the communication network and the sets S_p to be known beforehand is to allow some preprocessing: first, we identify the non-connected components of the variable, and then, we select which nodes should retransmit them. Note that this assumption only requires knowing beforehand the components each node depends on, but not the functions f_p. In other words, this preprocessing can be done before any data arrives.

Let x_l be a non-connected component, i.e., the induced subgraph $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$ is non-connected. As we have seen, the constraint $x_l^{(i)} = x_l^{(j)}$, $(i,j) \in \mathcal{E}_l$, in (4.21) is not enough to enforce equality of all the copies of x_l. We propose enlarging the subgraph $\mathcal{G}_l$ by selecting other nodes in the network that will retransmit estimates of x_l. In other words, we will add nodes (and edges) to $\mathcal{G}_l$ in order to make that induced subgraph connected. Since our goal is to minimize the overall number of communications, we should add the least number of edges to this subgraph. It turns out that this is exactly the problem of finding an optimal *Steiner tree* in the communication network.

Steiner tree problem. To describe the Steiner tree problem, consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in our case, the communication network, and let $\mathcal{R} \subset \mathcal{V}$ be a set of *required nodes*, in our

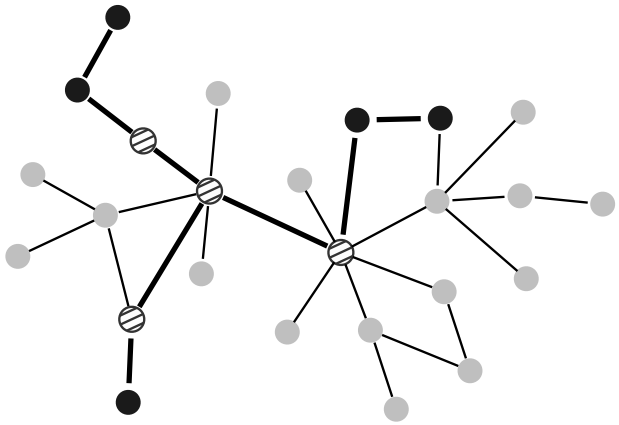


FIGURE 4.6: Example of an optimal Steiner tree. The required nodes $\mathcal{R}$ are black, and the Steiner nodes $\mathcal{S}$ are striped. The Steiner tree edges are represented thicker.

case, the nodes $\mathcal{V}_l$ of a non-connected induced subgraph $\mathcal{G}_l$. Figure 4.6 shows an example where $\mathcal{G}$ is the entire network, and $\mathcal{R}$ are the black nodes. A Steiner tree in $\mathcal{G}$ is any tree in that contains the required nodes $\mathcal{R}$; in other words, it is an acyclic connected subgraph $(\mathcal{T}, \mathcal{F}) \subseteq \mathcal{G}$ such that $\mathcal{R} \subseteq \mathcal{T}$ and $\mathcal{F} \subseteq \mathcal{E}$. The *Steiner nodes*, which will be represented with $\mathcal{S}$, are the nodes in that tree that are not required, i.e., $\mathcal{S} := \mathcal{T} \backslash \mathcal{R}$. For example, in Figure 4.6, the Steiner nodes $\mathcal{S}$ are striped and the Steiner tree edges $\mathcal{F}$ are thicker. Note that the set of black and striped nodes and the thicker edges form a subgraph that is a tree.

Now we can state the Steiner tree problem: *given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a set of required nodes $\mathcal{R} \subset \mathcal{V}$, and a set of costs c_{ij} for each edge of the network $(i,j) \in \mathcal{E}$, find a Steiner tree whose edges have a minimal cost.* In our case, since we want to minimize the total number of communications, all edges are viewed equal, that is, they all have the same cost, for example, $c_{ij} = 1$. The set of required nodes in our case are the nodes in the subgraph induced by a non-connected component x_l, i.e., $\mathcal{R} = \mathcal{V}_l$. Of course, we have to solve a Steiner tree problem for each non-connected component. Unfortunately, solving Steiner tree problems is NP-hard [223]. However, many approximation algorithms are available, some of which have approximation guarantees. For example, the Steiner tree problem can be formulated as the following optimization problem [224]

$$\begin{aligned} \underset{\{z_{ij}\}_{(i,j)\in\mathcal{E}}}{\text{minimize}} \quad & \sum_{(i,j)\in\mathcal{E}} c_{ij} z_{ij} \\ \text{subject to} \quad & \sum_{\substack{i\in\mathcal{U} \\ j\notin\mathcal{U}}} z_{ij} \geq 1\,, \quad \forall_{\mathcal{U}} \,:\, 0 < |\mathcal{U} \cap \mathcal{R}| < |\mathcal{R}| \\ & z_{ij} \in \{0,1\}\,, \quad (i,j) \in \mathcal{E}\,. \end{aligned} \tag{4.42}$$

In the first constraint of (4.42), $\mathcal{U}$ represents any subset of nodes that separates at least two required

nodes, i.e., $\mathcal{U}$ contains at least one node in $\mathcal{R}$, but not all of them. The optimization variable of problem (4.42) is $z \in \mathbb{R}^E$ and each z_{ij} is associated to edge $(i,j) \in \mathcal{E}$. If the optimal value is $z^\star_{ij} = 1$, then edge (i,j) is in the selected Steiner tree. Note that the last constraint of (4.42) imposes each component of z to be either 0 or 1. Let us denote the objective of problem (4.42) by $h(z) := \sum_{(i,j)\in\mathcal{E}} c_{ij} z_{ij}$. We say that an algorithm for (4.42) has an *approximation ratio of* α if it produces a feasible point $\bar{z}$ such that $h(\bar{z}) \leq \alpha h(z^\star)$, for any problem instance. The primal-dual algorithm for combinatorial problems [224, 225], for example, has an approximation ratio of 2. To the best of our knowledge, the algorithm for computing Steiner trees with the smallest approximation ratio, namely $1 + \ln 3/2 \simeq 1.55$, was proposed in [226].

Application to our problem. Based on Assumption 4.8 and on the concept of Steiner tree problem, we now propose a modification to Algorithm 5 to make it applicable to a non-connected variable. This modification applies to Algorithm 4 exactly the same way. According to Assumption 4.8, both the communication network and the sets S_p are known before the execution of the algorithm. This allows solving a Steiner tree problem for each non-connected component, as a preprocessing step, which can be done in a distributed or a centralized way; for distributed algorithms computing Steiner trees see, for example, [227, 228]. More concretely, for every non-connected component x_l with induced subgraph $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$, we can compute a Steiner tree $(\mathcal{T}_l, \mathcal{F}_l) \subseteq \mathcal{G}$ using $\mathcal{V}_l$ as the set of required nodes. Let $\mathcal{S}_l := \mathcal{T}_l \backslash \mathcal{V}_l$ denote the Steiner nodes in that tree. The functions associated to these Steiner nodes do not depend on x_l, i.e., $l \notin S_p$ for all $p \in \mathcal{S}_l$. But we artificially force them to depend on it by defining a new induced graph as $\mathcal{G}'_l = (\mathcal{V}'_l, \mathcal{E}'_l)$, with $\mathcal{V}'_l := \mathcal{T}_l$ and $\mathcal{E}'_l := \mathcal{E}_l \cup \mathcal{F}_l$. Then, we can create copies of x_l in all nodes in $\mathcal{V}'_l$, and write (P) equivalently as

$$\begin{array}{ll} \underset{\{\bar{x}_l\}_{l=1}^n}{\text{minimize}} & f_1(x_{S_1}^{(1)}) + f_2(x_{S_2}^{(2)}) + \cdots + f_P(x_{S_P}^{(P)}) \\ \text{subject to} & x_l^{(i)} = x_l^{(j)}, \quad (i,j) \in \mathcal{E}'_l\,,\ l = 1, \ldots, n\,, \end{array} \tag{4.43}$$

where $\{\bar{x}_l\}_{l=1}^L$ is the optimization variable, and $\bar{x}_l := \{x_l^{(p)}\}_{p\in\mathcal{V}'_l}$ denotes the set of all copies of x_l. If node p is a Steiner node for any component of the variable, it will hold "extra" copies, but its function f_p remains unchanged. In particular, it has the copies $x^{(p)}_{S_p \cup S'_p}$, where S'_p is the set of components of which node p is a Steiner node, but its function f_p depends only on $x^{(p)}_{S_p} := \{x_l^{(p)}\}_{l\in S_p}$. Of course, if a component x_l is connected, we set $\mathcal{G}'_l = \mathcal{G}_l$, and if node p is not Steiner for any component, we set $S'_p = \emptyset$. If we replace problem (4.21) by the modified problem (4.43) and repeat the derivation that followed problem (4.21), we get Algorithm 6.

Algorithm 6 is essentially an adapted version of Algorithm 5, with a preprocessing step, which can be computed in a centralized or in a distributed way. The preprocessing step relies on Assumption 4.8 by assuming that both the communication network and the dependency sets S_p are

Algorithm 6 Algorithm for a generic variable, connected or non-connected

Preprocessing:

1: Set $S'_p = \emptyset$ for all $p \in \mathcal{V}$, and $\mathcal{V}'_l = \mathcal{V}_l$ for all $l = \{1, \dots, n\}$
2: **for all** non-connected components x_l, $l \in \{1, \dots, n\}$ **do**
3: Compute a Steiner tree $(\mathcal{T}_l, \mathcal{F}_l)$, setting $\mathcal{V}_l$ as the set of required nodes
4: Set $\mathcal{V}'_l = \mathcal{T}_l$ and $\mathcal{S}_l := \mathcal{T}_l \backslash \mathcal{V}_l$ (Steiner nodes)
5: For all $p \in \mathcal{S}_l$, $S'_p = S'_p \cup \{x_l\}$
6: **end for**

Main algorithm:

Initialization: Choose $\rho \in \mathbb{R}$; for all $p \in \mathcal{V}$ and $l \in S_p \cup S'_p$, set $\gamma_l^{(p),0} = x_l^{(p),0} = 0$; set $k = 0$

7: **repeat**
8: **for** $c = 1, \dots, C$ **do**
9: **for all** $p \in \mathcal{C}_c$ [in parallel] **do**
10: Compute $v_l^{(p),k} = \gamma_l^{(p),k} - \rho \sum_{\substack{j \in \mathcal{N}_p \cap \mathcal{V}'_l \\ C(j) < c}} x_l^{(j),k+1} - \rho \sum_{\substack{j \in \mathcal{N}_p \cap \mathcal{V}'_l \\ C(j) > c}} x_l^{(j),k}$, for all $l \in S_p \cup S'_p$
11: Compute $x_{S_p \cup S'_p}^{(p),k+1} = \underset{x_{S_p \cup S'_p}^{(p)}}{\arg\min}\ f_p(x_{S_p}^{(p)}) + \sum_{l \in S_p \cup S'_p} \left({v_l^{(p),k}}^\top x_l^{(p)} + \frac{\rho}{2} D'_{p,l} \left(x_l^{(p)} \right)^2 \right)$
12: For each component $l \in S_p \cup S'_p$, exchange $x_l^{(p),k+1}$ to neighbors $\mathcal{N}_p \cap \mathcal{V}'_l$
13: **end for**
14: **end for**
15: **for all** $p \in \mathcal{V}$ and $l \in S_p \cup S'_p$ [in parallel] **do**
$\gamma_l^{(p),k+1} = \gamma_l^{(p),k} + \rho \sum_{j \in \mathcal{N}_p \cap \mathcal{V}'_l} (x_l^{(p),k+1} - x_l^{(j),k+1})$
16: **end for**
17: $k \leftarrow k + 1$
18: **until** some stopping criterion is met

known. Note that the specific functions f_p are not required for this preprocessing step. Regarding the main algorithm, it is similar to Algorithm 5 except that each node, in addition to estimating the components its function originally depends on, it also estimates the components for which it is a Steiner node. The computation for these additional components can, however, be found in closed-form: if node p is a Steiner node for component x_l, it updates it as $x_l^{(p),k+1} = -(1/(\rho\, D_{p,l}))v_l^{(p),k}$ in step 11. In Algorithm 6, $D'_{p,l}$ is defined as the degree of node p in the subgraph $\mathcal{G}'_l$. The steps we took to generalize Algorithm 5 to a non-connected variable can be easily applied the same way to Algorithm 4, the algorithm proposed by [47].

4.4 Experimental results

In this section, we assess experimentally the performance of the proposed algorithms, namely Algorithm 5 and Algorithm 6, with respect to prior distributed algorithms. We focus on two applications: networks flow problems and D-MPC. While network flow problems are formulated as (P)

with a star-shaped variable, D-MPC has more flexibility, since it can be formulated with any type of variable (see Subsection 4.2.1). As mentioned before, most of the prior distributed optimization algorithms solve (P) only when the variable is global or star-shaped. The only exception is the algorithm proposed by [47], which we presented as Algorithm 4. Indeed, that algorithm can solve (P) with any connected variable and, if using the adaptation we proposed in the previous section, it can also solve it with a non-connected variable.

Communication steps. The performance metric we use in our experiments is the number of communication steps (CSs). The concept is the same we introduced in Chapter 3 for the global class: after all the nodes have updated their estimates of the components they depend on and broadcast them to their neighbors, we say that a CS has occurred. The only difference with respect to the CS concept in Chapter 3 is in the size of the messages exchanged between nodes: here, two neighbors $(i, j) \in \mathcal{E}$ only exchange the common components their functions depend on, i.e., $x_{S_i \cap S_j}$, rather than the entire vector x. This applies to all the algorithms we compare in this chapter. The only exception is Algorithm 3, the algorithm we proposed for the global class, which we show here for comparison purposes. In fact, we will see that, even ignoring the difference in the size of the exchanged messages, Algorithm 3 takes more CSs to converge than any of the algorithms solving (P) with a non-global variable. This effectively illustrates how important it is to explore the structure of the problem in order to design communication-efficient algorithms.

4.4.1 Network flow problems

We start with the experiments on network flow problems. First, we describe the model we used in our experiments, then the experimental setup and the algorithms we compare, and finally we present our results.

Model. Recall that a network flow problem has the format of (4.16). Its objective consists of the sum of the costs $\phi_{ij}(x_{ij})$ associated to all the arcs of the directed network. The constraint $Bx = d$ enforces the laws of conservation of flow, whereas the constraint $x \geq 0$ forbids negative flows on each arc. We considered two scenarios for problem (4.16):

$$\text{Scenario 1:} \qquad \phi_{ij}(x_{ij}) = \frac{1}{2}(x_{ij} - a_{ij})^2\,, \text{ and the constraint } x \geq 0 \text{ is dropped,}$$
$$\text{Scenario 2:} \qquad \phi_{ij}(x_{ij}) = \frac{x_{ij}}{c_{ij} - x_{ij}} + \mathrm{i}_{x_{ij} \leq c_{ij}}(x_{ij})\,.$$

In scenario 1, the cost function associated to each arc $(i, j) \in \mathcal{A}$ is quadratic, $\phi_{ij}(x_{ij}) = \frac{1}{2}(x_{ij} - a_{ij})^2$, where a_{ij} is positive. Also, we drop the nonnegativity constraint $x \geq 0$ in order to make the

algorithm in [162] applicable. Scenario 1 is thus very simple: it solves

$$\begin{array}{ll} \underset{x=\{x_{ij}\}_{(i,j)\in\mathcal{A}}}{\text{minimize}} & \sum_{(i,j)\in\mathcal{A}} \frac{1}{2}(x_{ij}-a_{ij})^2 \\ \text{subject to} & Bx = d\,. \end{array} \tag{4.44}$$

Regarding scenario 2, besides the cost function being more complicated, $\phi_{ij}(x_{ij}) = x_{ij}/(c_{ij}-x_{ij})$, where $c_{ij} > 0$, is the maximum capacity of arc (i,j), it also has the constraints $0 \leq x_{ij} \leq c_{ij}$, for each arc. That is, scenario 2 solves

$$\begin{array}{ll} \underset{x=\{x_{ij}\}_{(i,j)\in\mathcal{A}}}{\text{minimize}} & \sum_{(i,j)\in\mathcal{A}} \frac{x_{ij}}{c_{ij}-x_{ij}} \\ \text{subject to} & Bx = d \\ & 0 \leq x_{ij} \leq c_{ij}\,. \end{array} \tag{4.45}$$

Problem (4.45) models aggregate system delays in multicommodity flow problems [20, Ch.4].

The problem each node has to solve at each iteration, for example, at step 5 of Algorithm 5, has a closed-form solution in scenario 1, but not in scenario 2. In scenario 2, node p has to solve a problem with the following format:

$$\begin{array}{ll} \underset{y=(y_1,\ldots,y_{D_p})}{\text{minimize}} & \sum_{i=1}^{D_p}(\frac{y_i}{c_i-y_i} + v_i y_i + a_i y_i^2) \\ \text{subject to} & b_p^\top y = d_p \\ & 0 \leq y \leq c\,, \end{array} \tag{4.46}$$

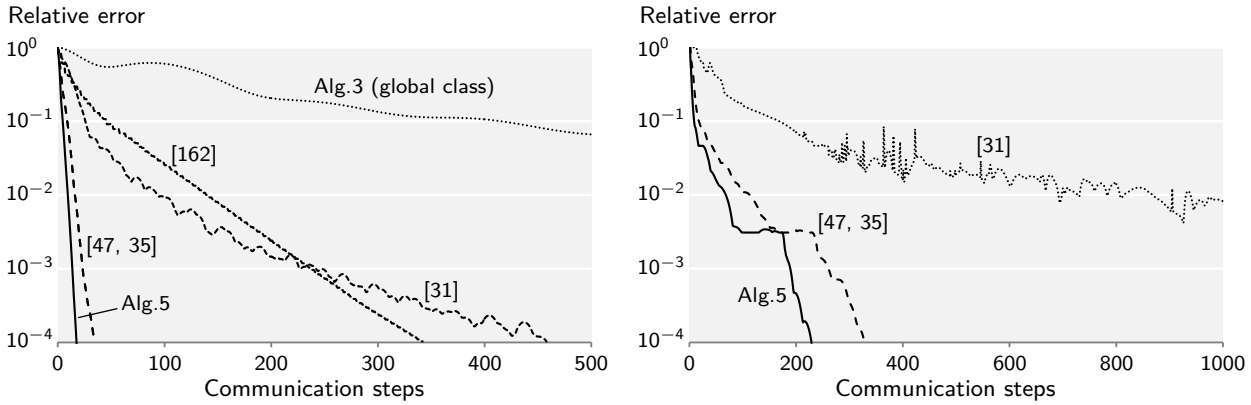
where each y_i corresponds to x_{pj} if $(p,j) \in \mathcal{A}$, or to x_{jp} if $(j,p) \in \mathcal{A}$. Since projecting a point onto the set of constraints of (4.46) can be done in closed-form [229], any projected gradient method is easy to apply. In our implementation, we chose [230], a gradient projection method with a Barzilai-Borwein step.

Experimental setup. In both instances of the network flow problem we solve, we use a network with $P = 2000$ nodes and $E = 3996$ edges, generated randomly in Network X [205] according to the Barabasi-Albert model [202]; see Table 3.1 of Chapter 3 for a brief description. As in the network flow problem illustrated in Figure 4.4, considered that between any pair of nodes there is at most one arc. As a consequence, the size of the problem variable, x, is equal to the number of edges $|\mathcal{E}|$, in this case, 3996. The diameter of the network was 8, it had an average node degree of 3.996, and it was colored with 3 colors in Sage [206]. We then assigned a direction to each edge of this network: for each edge (i,j), we assigned the directions $i \to j$ and $i \leftarrow j$ with equal probability, thus creating a set of arcs $\mathcal{A}$ from the set of edges $\mathcal{E}$. To each edge, we also assigned a number drawn randomly from the set $\{10, 20, 30, 40, 50, 100\}$. The probabilities were 0.2 for the

first four elements and 0.1 for 50 and 100. These numbers played the role of the a_{ij}'s in scenario 1 and the role of the capacities c_{ij} in scenario 2. To generate the vector d or, in other words, to determine which nodes are sources or sinks, we proceeded as follows. For each $k = 1, \ldots, 100$, we picked a source s_k randomly (uniformly) out of the set of 2000 nodes and then picked a sink r_k randomly (uniformly) out of the set of reachable nodes of s_k. For example, if we were considering the network of Figure 4.4 and picked $s_k = 4$ as a source node, the set of its reachable nodes would be $\{3, 5, 6, 7\}$. Then, we added to the entries s_k and r_k of d the values $-f_k/100$ and $f_k/100$, respectively, where f_k is a number drawn randomly exactly as c_{ij} (or a_{ij}). This corresponds to injecting a flow of quantity $f_k/100$ at node s_k and extracting the same quantity at node r_k. After repeating this process 100 times, for $k = 1, \ldots, K$, we obtained vector d.

Before executing the distributed algorithms and to assess their error, we computed the solutions of (4.44), from scenario 1, and (4.45), from scenario 2, in a centralized way. In scenario 1, the solution can be computed in closed-form, because the problem is quadratic with linear constraints. In scenario 2, we used CVXOPT [231] to obtain a solution of (4.45).

Algorithms for comparison. The network flow problems (4.44) and (4.45) are formulated as (P) with a star-shaped variable (see also (4.16)). As discussed before, in this case, the ADMM-based algorithm [35, §7.2] becomes distributed. In fact, for network flow problems it becomes exactly algorithm [47] (Algorithm 4); this is not surprising, since both are based on the same underlying algorithm, the 2-block ADMM. Also, a star-shaped variable makes gradient methods directly applicable. We then also consider Nesterov's fast gradient method [31], more precisely, the algorithm (2.10). Finally, we consider the distributed Newton method [162], which was designed specifically for network flow problems. All these methods, including ours, have tuning parameters: ρ for the ADMM-based algorithms, a Lipschitz constant L for Nesterov's algorithm, and a stepsize α for the distributed Newton algorithm. Note that Nesterov's algorithm requires the objective function to be differentiable and have a Lipschitz-continuous gradient. While this is true for (4.44), in scenario 1, it is not for (4.45), in scenario 2. Namely, the gradient of the objective of (4.45) is not Lipschitz-continuous in all the domain, although it is near the solution. Therefore, in scenario 2, we have to estimate a Lipschitz constant the same way we estimate the parameters of the other algorithms. To do that, we use the concept of *precision*, defined in Chapter 3: for example, $\bar{\rho}$ *has precision* γ for an ADMM-based algorithm if both $\bar{\rho} - \gamma$ and $\bar{\rho} + \gamma$ lead to worse results, i.e., to more CSs. Regarding the number of CSs each of these algorithms takes per iteration, all the ADMM-based ones (Algorithms 4 [47], 5, and [35, §7.2]) and Nesterov's algorithm [31] take one CS per iteration. Our implementation of the distributed Newton method [161], in turn, takes 3 CSs per iteration, since we used a fixed stepsize α and set the parameter N, the order of the approximation of Newton's direction, to 2. We will also show the performance of Algorithm 3, our proposed



(a) Scenario 1 (b) Scenario 2

FIGURE 4.7: Results of our experiments for the network flow problems. The problem solved in (a) is (4.44), a simple quadratic program. The problem solved in (b) is (4.45), which models aggregate delays in multicommodity flow problems. In both cases, the network has $P = 2000$ nodes and $E = 3996$ edges, and was generated randomly according to the Barabasi-Albert model.

algorithm for the global class, in scenario 1. That algorithm makes all the nodes compute the full solution $x^\star$, which has dimensions 3996 in this case. Hence, each message exchanged in one CS of Algorithm 3 is 3996 times larger than the messages exchanged by the other algorithms.

Results. The results of our experiments for scenarios 1 and 2 are shown, respectively, in Figures 4.7(a) and 4.7(b). These show the relative error on the primal variable $\|x^k - x^\star\|_\infty / \|x^\star\|_\infty$, where x^k is the concatenation of the estimates at all nodes, versus the number of CSs. It can be seen in Figure 4.7(a) that Algorithm 5 in scenario 1 was the one requiring the least amount of CSs to achieve any relative error between 1 and 10^{-4}. It was closely followed by the ADMM-based algorithms [47] and [35, §7.2], whose lines coincide because they become the same algorithm when applied to network flows. Nesterov's method [31] and the Newton-based method [162] had a performance very similar to each other, but worse than the ADMM-based algorithms. In the same plot we can also see that Algorithm 3, which solves the global class, had the worst performance; furthermore, each message exchange by that algorithm is 3996 times larger than a message exchanged by the other algorithms. This clearly shows that if we want to derive communication-efficient algorithms, we have to explore the structure of (P). Regarding the parameters for each algorithm in these experiments, we used $\rho = 2$ for all the ADMM-based algorithms (precision 1), a Lipschitz constant $L = 70$ for [31] (precision 5), and a stepsize $\alpha = 0.4$ for [162] (precision 0.1).

The results for scenario 2, i.e., for problem (4.45), are shown in Figure 4.7(b). We were not able to make the algorithm in [162] converge for this scenario (actually, that algorithm is not guaranteed to converge for problem (4.45)). Overall, scenario 2 looks more challenging to solve, since all algorithms took more CSs to achieve the same relative error. Again, Algorithm 5 was

the algorithm with the best performance. This time we could not find any choice for L that made Nesterov's algorithm [31] achieve the relative error of 10^{-4} in less than 1000 CSs. The best result, obtained for $L = 15000$, is shown in Figure 4.7(b). The augmented Lagrangian parameter ρ was 0.08 for Algorithm 5 and 0.12 for algorithms [47, 35], both computed with precision 0.02.

4.4.2 D-MPC

We now describe our experiments for distributed model predictive control (D-MPC). We start by describing the particular MPC model we used, and then the experimental setup.

Model. We proposed our D-MPC model (4.3) in Subsection 4.2.1, which is reproduced here for convenience:

$$\begin{array}{ll} \underset{\bar{x},\bar{u}}{\text{minimize}} & \sum_{p=1}^{P}\left[\Phi_p(\{x_j[T]\}_{j\in\Omega_p}) + \sum_{t=0}^{T-1}\Psi_p^t(\{x_j[t],u_j[t]\}_{j\in\Omega_p})\right] \\ \text{subject to} & x_p[t+1] = \Theta_p^t(\{x_j[t],u_j[t]\}_{j\in\Omega_p})\,, \quad t=0,\ldots,T-1\,, \quad p=1,\ldots,P \\ & x_p[0] = x_p^0\,, \quad p=1,\ldots,P\,. \end{array} \tag{4.47}$$

Problem (4.47) is associated to a network with P dynamic systems where each dynamic system is viewed as a node of that network. The pth system is described at each time instant t by the state vector $x_p[t] \in \mathbb{R}^{n_p}$ and has a control input $u_p[t] \in \mathbb{R}^{m_p}$. The D-MPC model (4.47) generalizes prior D-MPC models in the sense that it allows the state of any system be influenced by the state or input of any other system in the network, and not only by its neighbors; see also Figure 4.1 for a visual comparison between these two scenarios. Therefore, the optimization variable in (4.47) is arbitrary and not necessarily star-shaped. In our experiments, we consider a simple instance of (4.47) that preserves this feature. Namely, we assume linear coupling through the inputs, i.e., $x_p[t+1] = A_p x_p[t] + \sum_{j\in\Omega_p} B_{pj}u_j[t]$, where $A_p \in \mathbb{R}^{n_p\times n_p}$ and each $B_{pj} \in \mathbb{R}^{n_p\times m_j}$ are arbitrary matrices (in fact, randomly generated), known only at node p. The set $\Omega_p \subseteq \mathcal{V}$ is the set of nodes whose control input influences the state of node p, x_p. We assume that the control input at node p influences always its own state, i.e., $\{p\} \subset \Omega_p$, for all $p \in \mathcal{V}$. We also assume there is no coupling through the objective functions. In particular, we consider $\Phi_p(\{x_j[T]\}_{j\in\Omega_p}) = x_p[T]^\top \bar{Q}_p^f x_p[T]$ and $\Psi_p^t(\{x_j[t]\}_{j\in\Omega_p}) = x_p[t]^\top \bar{Q}_p x_p[t] + u_p[t]^\top \bar{R}_p$, where $\bar{Q}_p$ and $\bar{Q}_p^f$ are positive semidefinite matrices, and $\bar{R}_p$ is positive definite. With this choice, problem (4.47) becomes

$$\begin{array}{ll} \underset{\substack{x_1,\ldots,x_P \\ u_1,\ldots,u_P}}{\text{minimize}} & \sum_{p=1}^{P} u_p^\top R_p u_p + x_p^\top Q_p x_p \\ \text{subject to} & x_p = C_p\{u_j\}_{j\in\mathcal{S}_p} + D_p^0\,,\ p=1,\ldots,P\,, \end{array} \tag{4.48}$$

where, $x_p = (x_p[0], \ldots, x_p[T])$, $u_p = (u_p[0], \ldots, u_p[T-1])$, for each p, and

$$Q_p = \begin{bmatrix} I_T \otimes \bar{Q}_p & 0 \\ 0 & \bar{Q}_p^f \end{bmatrix}, \qquad R_p = I_T \otimes \bar{R}_p\,,$$

$$C_p = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ B_p & 0 & \cdots & 0 \\ A_{pp}B_p & B_p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{pp}^{T-1}B_p & A_{pp}^{T-2}B_p & \cdots & B_p \end{bmatrix}, \qquad D_p^0 = \begin{bmatrix} I \\ A_{pp} \\ A_{pp}^2 \\ \vdots \\ A_{pp}^T \end{bmatrix} x_p^0\,.$$

We defined the matrix B_p (in the entries of C_p) as the horizontal concatenation of the matrices B_{pj}, for all $j \in \Omega_p$. Note that the variables x_p and u_p in (4.48) now contain the states and inputs for the entire horizon. For this reason, we changed from the notation $j \in \Omega_p$ to the notation $j \in S_p$; while Ω_p is a subset of the set of nodes $\mathcal{V}$, S_p is a subset of components of the optimization variable, i.e., $S_p \in \{1, \ldots, (T+1)\sum_{p=1}^P n_p + T\sum_{p=1}^P m_p\}$. One reason we chose this simple linear model is that all the state variables x_p in (4.48) can be eliminated; indeed, (4.48) can be written equivalently as

$$\underset{u_1,\ldots,u_P}{\text{minimize}} \;\; \sum_{p=1}^{P} \{u_j\}_{j\in S_p}^\top E_p \{u_j\}_{j\in S_p} + w_p^\top \{u_j\}_{j\in S_p}\,, \tag{4.49}$$

where $w_p = 2C_p^\top Q_p D_p^0$ and each E_p is obtained by summing R_p with $C_p^\top Q_p C_p$ in the correct entries. Note that (4.49) is an unconstrained quadratic program. Therefore, in a centralized scenario, where all matrices E_p and all vectors w_p are known at the same location, the solution of (4.49) is simply the solution of a linear system. For the same reason, the solution of the problem each node has to solve at each iteration, for example in step 5 of Algorithm 5, can be found by solving a linear system.

TABLE 4.1: Networks used in the D-MPC experiments.

Name	Source	# Nodes	# Edges	Diam.	# Colors	Av. Deg.	Description
A	[202]	100	196	6	3	3.92	Barabasi-Albert (parameter 2)
B	[201]	4941	6594	46	6	2.67	US Western states power grid

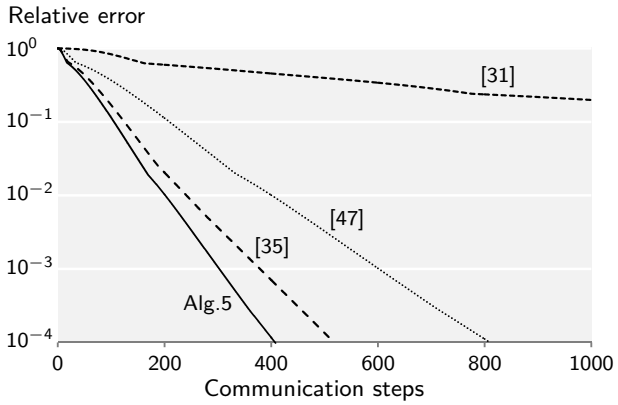
Experimental setup. We solved problem (4.49) in the two networks of Table 4.1. Network A has 100 nodes, 196 edges, and was generated randomly according to the Barabasi-Albert model [202], as briefly described in Table 3.1 of Chapter 3. A parameter of 2 means that every time a node is added to the network it connects to other 2 nodes. Network B is considerably larger, with 4941 nodes and 6594 edges and represents the topology of the power grid of the US Western

states [201]. Table 4.1 also shows the diameter of each network, the average degree of each node, and the number of colors they were colored with. We used a built-in function in Sage [206] to color these networks.

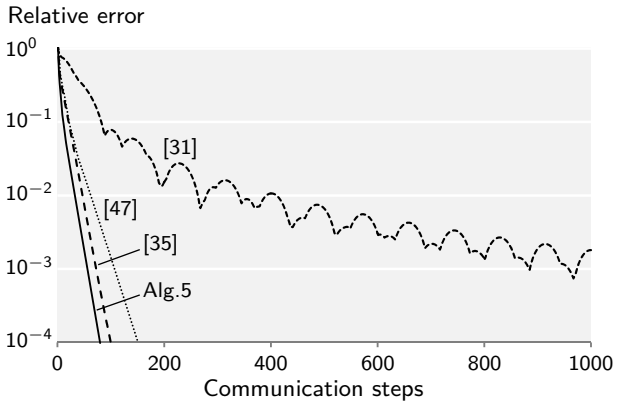
In all our experiments we considered a time horizon T of dimension 5, the state x_p of each node p always had dimensions $n_p = 3$, and the control input u_p was always scalar, $m_p = 1$, for all p. Given that the size of the variable in (4.49) is m_pTP, network A implied a variable of size 500 and network B implied a variable of size 24705. While each dynamical system in network A could be unstable, each dynamical system in network B was always guaranteed stable. More specifically, for both networks, we generated the entries of the dynamics matrix A_p of each system p from the normal distribution (independently); however, for network B, after generating each A_p, we always "shrunk" its eigenvalues to the interval $[-1, 1]$, making the corresponding system stable. Regarding the the input-state matrices B_{pj}, each of its entries were also drawn from the normal distribution.

We now describe how we generated the system couplings, i.e., the sets $\Omega_p \in \mathcal{V}$; see also the dotted arrows in Figure 4.1. We generated three types of couplings, and thus of variables. We generated star-shaped variables, where the state of system p is influenced by the inputs of all its neighbors, that is, $\Omega_p = \mathcal{N}_p$, for all p. This case is illustrated in Figure 4.1(a) and was considered so that we could compare Algorithms 4 and 5 with other prior D-MPC algorithms. We also generated instances of the system couplings to make the variable connected (not necessarily star-shaped), and non-connected. To generate a connected variable we proceeded as follows: given a node p, we make it depend on u_p (recall our assumption that $\{p\} \subset \Omega_p$). Then, we initialize a set $\mathcal{F}_p$, which we will call the "fringe," with the neighbors of node p, $\mathcal{F}_p = \mathcal{N}_p$. Next, we select randomly (uniformly) a node q from the fringe, $q \in \mathcal{F}_p$, and make its state depend on u_p, i.e., $p \in \Omega_q$. And we add its set of neighbors to the fringe and remove node q from it, since it already depends on u_p: $\mathcal{F}_p = \Big(\mathcal{F}_p \backslash \{q\}\Big) \cup \mathcal{N}_q$. This process is repeated 3 times for each node p, and is done for all the nodes in the network. To generate a non-connected variable, the process is exactly the same, including the concept of fringe. The difference is that, at each iteration, any node in the entire network could be selected, not just the nodes in the fringe; however, the nodes in the fringe had twice the probability of being selected with respect to the remaining nodes in the network. We generated a non-connected variable only for network A, running the described algorithm for each one of its 500 components (the size of the variable for this network is $m_pTP = 1 \times 5 \times 100 = 500$). We obtained 400 components for which the respective induced subgraphs were non-connected. According to the preprocessing step of Algorithm 6, we have to compute a Steiner tree for each of these 400 components. To do that, we used a built-in function in Sage [206]. In the end, 44 nodes in the network (out of 100) were Steiner nodes for at least one component.

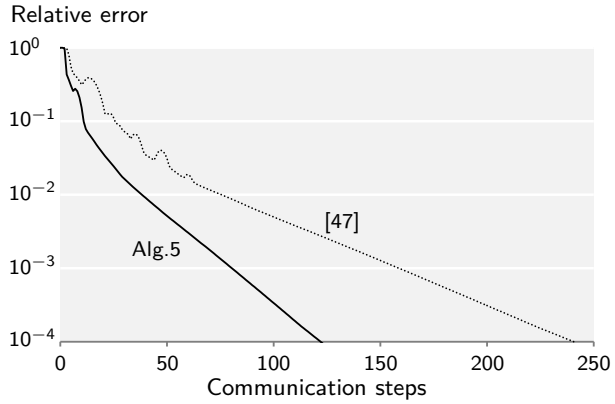
Results. The results of our experiments are shown in Figure 4.8 for connected variables, and



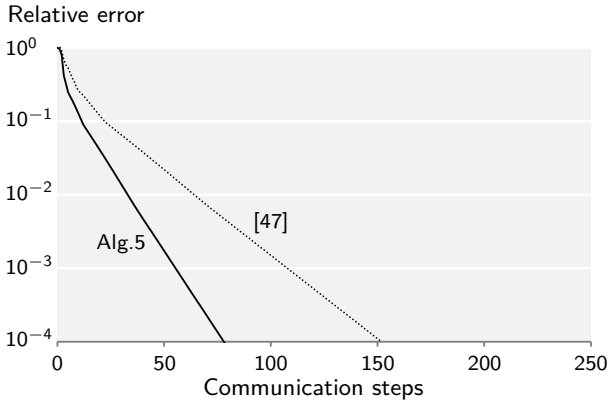
(a) Network A with a star-shaped variable



(b) Network B with a star-shaped variable



(c) Network A with a generic connected variable



(d) Network B with a generic connected variable

FIGURE 4.8: Results for D-MPC with a connected variable. On the left, (a) and (c) show the results for network A, and, on the right, (b) and (d) show the results for network B. The optimization variable is star-shaped on the top plots, (a) and (b), and is non-star-shaped (and non-global) on the bottom plots, (c) and (d).

in Figure 4.9 for a non-connected variable. Each plot shows how the relative error evolves as the number of CSs increases. The relative error is measured the same way as in the network flow experiments: $\|x^k - x^\star\|_\infty / \|x^\star\|_\infty$, where x^k is the concatenation of all the nodes' control input estimates. The results for networks A and B, both with a star-shaped variable, are shown in Figures 4.8(a) and 4.8(b), respectively. The relative behavior of all the compared algorithms is the same: the proposed Algorithm 5 required uniformly less CSs to achive any relative error between 1 and 10^{-4}; the ADMM-based algorithms [35, §7.2] and [47] (shown as Algorithm 4) followed, with [35, §7.2] more efficient than [47]. Finally, Nesterov's algorithm [31] failed to converge in both cases. A curious fact is that all algorithms required more CSs to converge in the network of Figure 4.8(a), which has 100 nodes, than in the network of Figure 4.8(b), which is considerably larger, with nearly 5000 nodes. In fact, what influenced the performance of all the algorithms was the stability

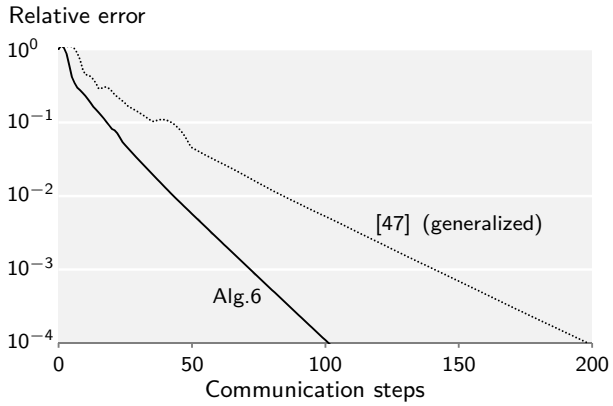


FIGURE 4.9: Results for D-MPC with a non-connected variable. All the dynamic systems were designed stable in this case, and the network was A.

of the systems: while each system in Figure 4.8(b) was guaranteed to be stable, no system in Figure 4.8(a) was guaranteed stable. The difficulty of each problem instances can be measured by the size of the Lipschitz constant of the gradient of the objective function of (4.49): 1.63×10^6 for Figure 4.8(a) and 3395 for Figure 4.8(b). Note that this Lipschitz constant can be computed in closed-form. Regarding the augmented Lagrangian parameter ρ, its was computed, with precision 5, for Figure 4.8(a) as 120 for [35, §7.2] and as 135 for the other algorithms. For Figure 4.8(b), it was computed as 25 for Algorithm 5 and [35, §7.2] and as 30 for [47], also with precision 5.

Figures 4.8(c) and 4.8(d) show the results for generic, non-star-shaped variables for networks A and B, respectively. Since the ADMM-based algorithm [35, §7.2] and Nesterov's algorithm [31] are distributed only for star-shaped variables, they do not appear in these plots. Only the proposed Algorithm 5 and the algorithm in [47] (see Algorithm 4) can handle generic connected variables. In both plots, Algorithm 5 required uniformly less CSs than [47] to achieve any relative error between 1 and 10^{-4}. Again, both algorithms required more CSs to converge in the smaller network A than in the larger network B. The reason, as we saw for the other plots, is because each system in network A can be unstable, while all systems in network B are stable. The value of ρ was the same for both algorithms: 40 for network A in Figure 4.8(c) (precision 5), and 23 for network B (precision 1).

Finally, we present the results for a non-connected variable in Figure 4.9. Neither Algorithm 5 nor the algorithm in [47] are applicable in this case. However, they can be adapted to non-connected variables, as described in Subsection 4.3.2. The generalization of Algorithm 5 yields Algorithm 6, and exactly the same generalization can be applied to the algorithm in [47]. Figure 4.9 shows that the behavior we had seen for the non-generalized versions of the algorithms in the previous experiments translates into the generalized versions: Algorithm 6 requires uniformly less CSs than the generalized version of [47] to achieve any relative error between 1 and 10^{-4}. Note that, although we used network A in these experiments, we guaranteed that all the systems were stable.

Chapter 5

Conclusions and Future Work

We restate our main problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f_1(x_{S_1}) + f_2(x_{S_2}) + \cdots + f_P(x_{S_P})\,. \tag{P}$$

and recall the main goals of this thesis, as presented before in Chapter 1:

> We aim to design, analyze, and implement algorithms that solve optimization problems of the form (P) on networks. The algorithms should be
>
> **Distributed:** no node has complete knowledge about the problem data and no central node is allowed; also, each node communicates only with its neighbors;
>
> **Communication-efficient:** the number of communications they use is minimized;
>
> **Network-independent:** the algorithms run on networks with arbitrary topology and their output is independent of the network.

First, we summarize our contributions to achieve this goal and discuss current limitations; then, we describe potential future work.

5.1 Major contributions

We group the contributions of the thesis into the following categories:

- **Classification scheme.** The optimization problem (P) is quite generic because each function may depend on an arbitrary subset of components of the optimization variable. This makes the design of a distributed algorithm a challenging task. We solve this problem with a

classification scheme that allowed us to first identify particular instances of (P) that are easier to solve in a distributed way. After that, we generalized the algorithms to solve larger classes and eventually all problems of the form (P). Besides helping us develop our algorithms, our classification scheme is also useful to categorize applications and to organize prior work on distributed optimization.

- **Algorithms.** Based on the proposed classification scheme, we developed a set of algorithms that solve subclasses of distributed optimization problems of the form (P). Each algorithm was built from a previous one, by modifying it to increase generality. Our most general algorithm solves (P) in full generality. Our algorithms satisfy all the requirements we had set forth: they are distributed, network-independent and, most significantly, they are communication-efficient. Under certain conditions, they are proven to converge to the same solution as a centralized algorithm and, as shown through several experiments, they usually outperform prior distributed optimization algorithms; namely, they use systematically less communications to achieve a prescribed solution accuracy. A surprising fact is that, despite their generality, they sometimes even outperform distributed algorithms that were designed for specific applications.

- **Applications.** We applied our algorithms to several known distributed problems, and also proposed new applications for them, such as several instances of compressed sensing (or sparse approximation) problems. Namely, we solve the three most important optimization problems in compressed sensing in both the cases where the sensing matrix is partitioned vertically (by rows) and horizontally (by columns). We also propose a new, more general framework for distributed model predictive control (D-MPC). This framework models scenarios where, for example, two dynamical systems that are coupled through their dynamics do not communicate directly. Thus, it is useful in scenarios where establishing communications between systems is expensive.

- **Implementation and benchmarking.** Since there are no tight lower bounds on how many communications are needed to solve (P) in a distributed setting, the performance assessment of our algorithms had to be done by comparing them to other prior distributed algorithms. This involved implementing both our algorithms and the algorithms for which no implementation was publicly available. We performed several experiments on different types of networks and for different applications where all the algorithms were compared. The size of both the data and the networks varied considerably. For example, the smallest network had only 10 nodes, while the largest one had around 5000 nodes. As mentioned before, these experiments enabled us to confirm the communication-efficiency of our algorithms.

5.2 Current limitations

Despite the excellent communication-efficiency of our algorithms, they still have several limitations:

- **Selection of ρ.** The algorithms we proposed are based on an augmented Lagrangian method called multi-block alternating direction method of multipliers (ADMM). Augmented Lagrangian methods are generally parametrized by a scalar parameter, which we denote with ρ, and their performance is strongly dependent on that parameter. Currently, there is no known method for selecting ρ before the execution of the algorithm. And, although there are some heuristics to adapt ρ while the algorithm is running, implementing those heuristics in distributed algorithms destroys their distributivity, since it requires aggregating information that is spread over the entire network. Therefore, the performance of the algorithms we proposed are conditionally dependent on a good choice for the parameter ρ. While in some situations it is possible to select beforehand a good ρ using training data, this is still a current limitation.

- **Convergence results.** As mentioned, our algorithms are based on the centralized multi-block ADMM algorithm. There is a proof of the convergence of this algorithm only in the case where all the cost functions are strongly convex. Yet, it has been observed experimentally, including in this thesis, that the multi-block ADMM converges for generic closed convex functions. Proving its convergence for this case is, however, still a well-known open problem. The lack of theoretical results for the multi-block ADMM transfers directly to our algorithms. In particular, we could only prove their convergence for generic closed convex functions when the network is bipartite. When it is not, our algorithms are only (theoretically) guaranteed to converge when the functions associated to each node are strongly convex.

- **Coloring scheme.** All our algorithms use the concept of network coloring and require a coloring scheme to be available before their execution. This coloring scheme is used by our algorithms to synchronize the order of operation of the nodes. In many platforms, most notably, in wireless networks, the nodes already have to operate with such a synchronization scheme in order to avoid packet collisions. In those cases, our algorithms integrate naturally with these low-level protocols. There are, however, some platforms that use other types of protocols or that even all fully parallel communication. In those cases, the coloring scheme required by our algorithms is clearly a limitation.

5.3 Future work

We see three main future research directions, as described next:

- **Algorithm analysis.** We mentioned as a limitation of our algorithms the lack of convergence results. This is closely related to the lack of convergence of the multi-block ADMM, a currently well-known open problem. Therefore, results on this direction would have a significant impact on the distributed algorithms we proposed. Also in this category is the task of developing an heuristic to adapt the augmented Lagrangian parameter ρ during the execution of the algorithm, and in a distributed way.

- **New distributed algorithms.** Another possible research direction is the development of new distributed optimization algorithms. The current most efficient algorithms are based on ADMM, which can be viewed as an application of a monotone operator splitting method to an optimization problem. Therefore, exploring monotone operator theory and devising new splitting methods may yield new and more efficient distributed optimization algorithms. A topic that became more relevant with the advent of the "big data" is privacy. In our view, it would be interesting to study privacy guarantees offered by distributed algorithms in the processing of distributed data.

- **New applications.** Although there are many applications for distributed optimization, including the ones presented in this thesis, the majority of them involve convex problems. Yet, many optimization problems formulated on networks are inherently nonconvex, for example, network coloring or the computation of Steiner trees. An interesting area to explore is the design of distributed approximation schemes for these types of nonconvex problems.

Appendix A

ADMM-based Algorithms For The Global Class: Derivation

In this appendix, we derive algorithms 1 and 2, from Chapter 2. Although these algorithms were proposed in [25] and [26], respectively, they were derived there for particular instances of the global class (G). Here, we generalize them to solve the entire class. Before their derivation, we need some identities for quantities defined on the edges of a network.

A.1 Network identities

Recall that we took the convention in Section 1.3.1 that if $(i,j) \in \mathcal{E}$, then $i < j$. The following lemma will be useful for exchanging between "edge notation" and "node notation."

Lemma A.1.

(a) Let a_{ij} be any quantity associated with the edge $(i,j) \in \mathcal{E}$. Then,

$$\sum_{(i,j)\in\mathcal{E}} a_{ij} = \sum_{p=1}^{P} \Bigg(\sum_{\substack{j\in\mathcal{N}_p \\ p<j}} a_{pj} + \sum_{\substack{j\in\mathcal{N}_p \\ j<p}} a_{jp} \Bigg) . \tag{A.1}$$

Furthermore, if $a_{ij} = a_{ji}$ for all $(i,j) \in \mathcal{E}$, (A.1) *becomes*

$$\sum_{(i,j)\in\mathcal{E}} a_{ij} = \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_p} a_{pj} \, . \tag{A.2}$$

(b) Let a_{ij} and a_{ji} be associated with edge $(i,j) \in \mathcal{E}$. Then,

$$\sum_{p=1}^{P} \sum_{j \in \mathcal{N}_p} a_{pj} = \sum_{p=1}^{P} \sum_{j \in \mathcal{N}_p} a_{jp} \,. \tag{A.3}$$

Proof.

(a) We have

$$\begin{aligned} \sum_{(i,j)\in\mathcal{E}} a_{ij} &= \sum_{(p,j)\in\mathcal{E}} a_{pj} + \sum_{(i,p)\in\mathcal{E}} a_{ip} + \sum_{\substack{(i,j)\in\mathcal{E} \\ i,j\neq p}} a_{ij} \\ &= \sum_{\substack{j\in\mathcal{N}_p \\ p<j}} a_{pj} + \sum_{\substack{j\in\mathcal{N}_p \\ j<p}} a_{jp} + \sum_{\substack{(i,j)\in\mathcal{E} \\ i,j\neq p}} a_{ij} \,, \end{aligned}$$

and repeating iteratively for all P nodes,

$$= \sum_{p=1}^{P} \Big(\sum_{\substack{j\in\mathcal{N}_p \\ p<j}} a_{pj} + \sum_{\substack{j\in\mathcal{N}_p \\ j<p}} a_{jp} \Big) \,.$$

When $a_{pj} = a_{jp}$, then

$$\sum_{(i,j)\in\mathcal{E}} a_{ij} = \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_p} a_{pj} \,.$$

(b) There holds

$$\begin{aligned} \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_p} a_{pj} &= \sum_{i\in\mathcal{N}_1} a_{i1} + \sum_{p=1}^{P} \sum_{\substack{j\in\mathcal{N}_p \\ j\neq 1}} a_{pj} \\ &= \sum_{i\in\mathcal{N}_1} a_{i1} + \sum_{i\in\mathcal{N}_2} a_{i2} + \sum_{p=1}^{P} \sum_{\substack{j\in\mathcal{N}_p \\ j\neq 1,2}} a_{pj} \,, \end{aligned}$$

and repeating for all nodes,

$$= \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_p} a_{jp} \,.$$

□

A.2 Derivation of Algorithm 1

We reproduce here problem (2.33), which was obtained as a reformulation of (G):

$$\begin{array}{ll} \underset{\bar{x},\bar{z}}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} & x_p = z_j\,, \quad j \in \mathcal{N}_p^+\,, \quad p = 1, \ldots, P\,. \end{array}$$

Recall that each node p has two copies, $x_p \in \mathbb{R}^n$ and $z_p \in \mathbb{R}^n$, of the original problem variable $x \in \mathbb{R}^n$. The collection of the x_p's and of the z_p's are $\bar{x} = (x_1, \ldots, x_P)$ and $\bar{z} = (z_1, \ldots, z_P)$, respectively. We can apply the 2-block ADMM (2.18)-(2.20) to this problem, seeing $\bar{x}$ and $\bar{z}$ as the two block variables. The augmented Lagrangian is

$$L_\rho(z, x; \lambda) = \sum_{p=1}^{P} f_p(x_p) + \sum_{p=1}^{P} \sum_{j \in \mathcal{N}_p^+} \lambda_{pj}^\top (x_p - z_j) + \frac{\rho}{2} \sum_{p=1}^{P} \sum_{j \in \mathcal{N}_p^+} \|x_p - z_j\|^2\,, \tag{A.4}$$

where λ_{pj} is the dual variable associated to the constraint $x_p - z_j = 0$, and $\lambda = (\ldots, \lambda_{ij}, \ldots)$ is the collection of dual variables. We consider $\bar{z}$ as the first block variable, and $\bar{x}$ as the second block variable.

Minimization in $\bar{z}$. Fixing $\bar{x}$ and λ at $\bar{x}^k$ and λ^k, respectively, $\bar{z}$ is updated as

$$\bar{z}^{k+1} = \arg\min_{\bar{z}} \sum_{p=1}^{P} \sum_{j \in \mathcal{N}_p^+} {\lambda_{jp}^k}^\top (x_j^k - z_p) + \frac{\rho}{2} \sum_{p=1}^{P} \sum_{j \in \mathcal{N}_p^+} \|x_j^k - z_p\|^2\,, \tag{A.5}$$

where we used the identity (A.3). Note that we also dropped the first term in (A.4), since it does not depend on $\bar{z}$. Now, (A.5) decouples into P problems that can be solved in parallel. The problem associated to node p is

$$\begin{aligned} z_p^{k+1} &= \arg\min_{z_p} \sum_{j \in \mathcal{N}_p^+} {\lambda_{jp}^k}^\top (x_j^k - z_p) + \frac{\rho}{2} \sum_{j \in \mathcal{N}_p^+} \|z_p - x_j^k\|^2 \\ &= \arg\min_{z_p} -\Big(\sum_{j \in \mathcal{N}_p^+} \lambda_{jp}^k + \rho \sum_{j \in \mathcal{N}_p^+} x_j^k \Big)^\top z_p + \frac{\rho(D_p + 1)}{2} \|z_p\|^2\,, \end{aligned}$$

which has the closed-form solution

$$\begin{aligned} z_p^{k+1} &= \frac{1}{\rho(D_p+1)} \Big(\sum_{j \in \mathcal{N}_p^+} \lambda_{jp}^k + \rho \sum_{j \in \mathcal{N}_p^+} x_j^k \Big) \\ &= \tau_p \sum_{j \in \mathcal{N}_p^+} \lambda_{jp}^k + \frac{1}{D_p + 1} \sum_{j \in \mathcal{N}_p^+} x_j^k\,, \end{aligned} \tag{A.6}$$

where $\tau_p = 1/(\rho(D_p+1))$.

Minimization in $\bar{x}$. Fixing $\bar{z}$ and λ at $\bar{z}^{k+1}$ and λ^k, respectively, $\bar{x}$ is updated as

$$\bar{x}^{k+1} = \arg\min_{\bar{x}} \sum_{p=1}^{P} f_p(x_p) + \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_p^+} {\lambda_{pj}^k}^\top (x_p - z_j^{k+1}) + \frac{\rho}{2} \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_p^+} \|x_p - z_j^{k+1}\|^2,$$

which decouples into P optimization problems that can be solved in parallel. The problem associated to node p is

$$\begin{aligned} x_p^{k+1} &= \arg\min_{x_p} f_p(x_p) + \sum_{j\in\mathcal{N}_p^+} {\lambda_{pj}^k}^\top (x_p - z_j^{k+1}) + \frac{\rho}{2} \sum_{j\in\mathcal{N}_p^+} \|x_p - z_j^{k+1}\|^2 \\ &= \arg\min_{x_p} f_p(x_p) + \Big(\sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^k - \rho \sum_{j\in\mathcal{N}_p^+} z_j^{k+1} \Big)^\top x_p + \frac{\rho(D_p+1)}{2} \|x_p\|^2, \end{aligned}$$

and after completing the square,

$$\begin{aligned} &= \arg\min_{x_p} f_p(x_p) + \frac{1}{2\tau_p} \Big\| x_p + \tau_p \Big(\sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^k - \rho \sum_{j\in\mathcal{N}_p^+} z_j^{k+1} \Big) \Big\|^2 \\ &= \operatorname{prox}_{\tau_p f_p} \Big(\frac{1}{D_p+1} \sum_{j\in\mathcal{N}_p^+} z_j^{k+1} - \tau_p \sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^k \Big), \end{aligned} \tag{A.7}$$

where the operator prox is defined in (2.34).

Update of the dual variables. According to ADMM (cf. (2.20)), each dual variable λ_{pj}, for $j \in \mathcal{N}_p^+$ and $p = 1, \ldots, P$, is updated as $\lambda_{pj}^{k+1} = \lambda_{pj}^k + \rho(x_p^{k+1} - z_j^{k+1})$. Node p, however, does not need to know each individual λ_{ij}. In fact, (A.6) and (A.7) only depend on the sums $\mu_p^k := \sum_{j\in\mathcal{N}_p^+} \lambda_{jp}^k$ and $\eta_p^k := \sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^k$, respectively. And these sums (or better, these new dual variables μ_p and η_p) can be updated as

$$\mu_p^{k+1} = \sum_{j\in\mathcal{N}_p^+} \lambda_{jp}^{k+1} = \underbrace{\sum_{j\in\mathcal{N}_p^+} \lambda_{jp}^k}_{\mu_p^k} + \rho \sum_{j\in\mathcal{N}_p^+} (x_j^{k+1} - z_p^{k+1}) = \mu_p^k + \frac{1}{\tau_p} \Big(\frac{1}{D_p+1} \sum_{j\in\mathcal{N}_p^+} x_j^{k+1} - z_p^{k+1} \Big)$$

$$\eta_p^{k+1} = \sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^{k+1} = \underbrace{\sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^k}_{\eta_p^k} + \rho \sum_{j\in\mathcal{N}_p^+} (x_p^{k+1} - z_j^{k+1}) = \eta_p^k + \frac{1}{\tau_p} \Big(x_p^{k+1} - \frac{1}{D_p+1} \sum_{j\in\mathcal{N}_p^+} z_j^{k+1} \Big).$$

These updates constitute step 5 of Algorithm 1. If we replace $\sum_{j\in\mathcal{N}_p^+} \lambda_{jp}^k$ in (A.6) and $\sum_{j\in\mathcal{N}_p^+} \lambda_{pj}^k$ in (A.7) by μ_p^k and η_p^k, respectively, we get steps 3 and 4.

A.3 Derivation of Algorithm 2

The reformulation [26] makes of (G) is (2.35), which we reproduce here:

$$\begin{array}{ll} \underset{\bar{x},\bar{z}}{\text{minimize}} & f_1(x_1) + f_2(x_2) + \cdots + f_P(x_P) \\ \text{subject to} & x_i = z_{ij}\,, \quad (i,j) \in \mathcal{E} \\ & x_j = z_{ij}\,, \quad (i,j) \in \mathcal{E}\,. \end{array}$$

Associating the dual variables λ_{ij} to the first set of constraints and η_{ij} to the second one, the augmented Lagrangian is

$$L_\rho(\bar{x},\bar{z};\lambda,\eta) = \sum_{p=1}^{P} f_p(x_p) + \sum_{(i,j)\in\mathcal{E}} \left(\lambda_{ij}^\top (x_i - z_{ij}) + \eta_{ij}^\top (x_j - z_{ij}) + \frac{\rho}{2}\|x_i - z_{ij}\|^2 + \frac{\rho}{2}\|x_j - z_{ij}\|^2 \right), \tag{A.8}$$

where λ (resp. η) is the collection of the dual variables λ_{ij} (resp. η_{ij}). The 2-block ADMM (2.18)-(2.20) applied to this problem translates into

$$\bar{x}^{k+1} = \arg\min_{\bar{x}} L_\rho(\bar{x},\bar{z}^k;\lambda^k,\eta^k) \tag{A.9}$$

$$\bar{z}^{k+1} = \arg\min_{\bar{z}} L_\rho(\bar{x}^{k+1},\bar{z};\lambda^k,\eta^k) \tag{A.10}$$

$$\lambda_{ij}^{k+1} = \lambda_{ij}^k + \rho(x_i^{k+1} - z_{ij}^{k+1})\,, \quad (i,j) \in \mathcal{E} \tag{A.11}$$

$$\eta_{ij}^{k+1} = \eta_{ij}^k + \rho(x_j^{k+1} - z_{ij}^{k+1})\,, \quad (i,j) \in \mathcal{E}\,. \tag{A.12}$$

We first analyze the minimization with respect to $\bar{z}$, (A.10); then, we analyze the minimization with respect to $\bar{x}$, (A.9); and, finally, we will see how to simplify the updates of the dual variables (A.11) and (A.12).

Minimization in $\bar{z}$. Since the augmented Lagrangian is quadratic in $\bar{z}$, problem (A.10) has a closed form solution. To compute it component-wise, just select $(i,j) \in \mathcal{E}$, and

$$\begin{aligned} \frac{\partial}{\partial z_{ij}} L_\rho(\bar{x}^{k+1},\bar{z};\lambda^k,\eta^k)\Big|_{z_{ij}=z_{ij}^{k+1}} = 0 \quad &\iff \quad -(\lambda_{ij}^k + \eta_{ij}^k) - \rho(x_i^{k+1} - z_{ij}^{k+1}) - \rho(x_j^{k+1} - z_{ij}^{k+1}) = 0 \\ &\iff \quad z_{ij}^{k+1} = \frac{\lambda_{ij}^k + \eta_{ij}^k}{2\rho} + \frac{x_i^{k+1} + x_j^{k+1}}{2}\,. \end{aligned} \tag{A.13}$$

Replacing (A.13) in (A.11) and (A.12), we get, respectively,

$$\lambda_{ij}^{k+1} = \lambda_{ij}^k + \rho\left(x_i^{k+1} - \frac{\lambda_{ij}^k + \eta_{ij}^k}{2\rho} - \frac{x_i^{k+1} + x_j^{k+1}}{2} \right) = \frac{\lambda_{ij}^k - \eta_{ij}^k}{2} + \rho\,\frac{x_i^{k+1} - x_j^{k+1}}{2} \tag{A.14}$$

$$\eta_{ij}^{k+1} = \eta_{ij}^{k} + \rho\left(x_j^{k+1} - \frac{\lambda_{ij}^k + \eta_{ij}^k}{2\rho} - \frac{x_i^{k+1} + x_j^{k+1}}{2}\right) = \frac{\eta_{ij}^k - \lambda_{ij}^k}{2} + \rho\,\frac{x_j^{k+1} - x_i^{k+1}}{2}\,. \tag{A.15}$$

Note that if we sum up (A.14) and (A.15), we get

$$\lambda_{ij}^{k+1} + \eta_{ij}^{k+1} = 0\,, \tag{A.16}$$

which holds for all $k \geq 1$. Let us assume that it also holds for $k = 0$, i.e., λ_{ij}^0 and η_{ij}^0 are initialized with symmetric values. Then, the first term in (A.13) is zero, and updating z_{ij} simplifies to

$$z_{ij}^{k+1} = \frac{x_i^{k+1} + x_j^{k+1}}{2}\,. \tag{A.17}$$

Similarly, the updates of the dual variables, (A.14) and (A.15) simplify, respectively, to

$$\lambda_{ij}^{k+1} = \lambda_{ij}^k + \rho\,\frac{x_i^{k+1} - x_j^{k+1}}{2} \tag{A.18}$$

$$\eta_{ij}^{k+1} = \eta_{ij}^k + \rho\,\frac{x_j^{k+1} - x_i^{k+1}}{2}\,, \tag{A.19}$$

Since we assume that writing $(i,j) \in \mathcal{E}$ means that $i < j$, the sets of dual variables λ_{ij} and η_{ij} are only defined for $i < j$. Let us extend their definition in a meaningful way, i.e., such that (A.18) and (A.19) make sense. Then, for $i > j$, we define λ_{ij}^k and η_{ij}^k, respectively, as

$$\lambda_{ij}^k = \eta_{ji}^k\,, \qquad i > j \tag{A.20}$$

$$\eta_{ij}^k = \lambda_{ji}^k\,, \qquad i > j\,. \tag{A.21}$$

Next, we use the identity (A.16), which holds for all k, and the simplified updates (A.17), (A.18), and (A.19) to find a simple expression for the minimization in $\bar{x}$, (A.9).

Minimization in $\bar{\boldsymbol{x}}$. If we set $z_{ij} = z_{ij}^k$, $\lambda_{ij} = \lambda_{ij}^k$, and $\eta_{ij} = \eta_{ij}^k$ in the augmented Lagrangian (A.8), the second term becomes

$$\sum_{(i,j)\in\mathcal{E}} \left(\lambda_{ij}^{k\,\top}(x_i - z_{ij}^k) + \eta_{ij}^{k\,\top}(x_j - z_{ij}^k) + \frac{\rho}{2}\|x_i - z_{ij}^k\|^2 + \frac{\rho}{2}\|x_j - z_{ij}^k\|^2\right) \tag{A.22}$$

$$= \sum_{(i,j)\in\mathcal{E}} \left(\lambda_{ij}^{k\,\top} x_i + \eta_{ij}^{k\,\top} x_j - \underbrace{(\lambda_{ij}^k + \eta_{ij}^k)}_{=0}{}^{\top} z_{ij}^k + \frac{\rho}{2}\|x_i - z_{ij}^k\|^2 + \frac{\rho}{2}\|x_j - z_{ij}^k\|^2\right), \tag{A.23}$$

$$= \sum_{(i,j)\in\mathcal{E}} \underbrace{\left(\lambda_{ij}^{k\,\top} x_i + \lambda_{ji}^{k\,\top} x_j\right)}_{:=a_{ij}} + \frac{\rho}{2} \sum_{(i,j)\in\mathcal{E}} \underbrace{\left(\|x_i - z_{ij}^k\|^2 + \|x_j - z_{ij}^k\|^2\right)}_{:=b_{ij}}. \tag{A.24}$$

From (A.22) to (A.23), we just rearranged the first two terms in the sum and used identity (A.16). From (A.23) to (A.24), we used definition (A.20). Now note that $a_{ij} = a_{ji}$ and also that $b_{ij} = b_{ji}$ (since, by (A.17), $z_{ij}^k = z_{ji}^k$). By identity (A.2) in Lemma A.1, we can write (A.24) as

$$\begin{aligned}
&\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\left({\lambda_{pj}^k}^\top x_p + {\lambda_{jp}^k}^\top x_j\right) + \frac{\rho}{2}\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\left(\|x_p - z_{pj}^k\|^2 + \|x_j - z_{pj}^k\|^2\right)\\
=&\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}{\lambda_{pj}^k}^\top x_p + \sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}{\lambda_{jp}^k}^\top x_j + \frac{\rho}{2}\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\|x_p - z_{pj}^k\|^2 + \frac{\rho}{2}\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\|x_j - z_{pj}^k\|^2 \qquad \text{(A.25)}\\
=&2\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}{\lambda_{pj}^k}^\top x_p + \rho\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\|x_p - z_{pj}^k\|^2. \qquad \text{(A.26)}
\end{aligned}$$

From (A.25) to (A.26), we used identity (A.3) from Lemma A.1 in the second and fourth terms, and also that $z_{ij}^k = z_{ji}^k$. We can now write (A.26) as

$$\begin{aligned}
\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\left(2\,{\lambda_{pj}^k}^\top x_p + \rho\|x_p - z_{pj}^k\|^2\right) &= \sum_{p=1}^{P}\sum_{j\in\mathcal{N}_p}\left(\left(2\,\lambda_{pj}^k - 2\rho\, z_{pj}^k\right)^\top x_p + \rho\|x_p\|^2 + \rho\|z_{pj}^k\|^2\right)\\
&= \sum_{p=1}^{P}\Bigg(\Big(\underbrace{2\sum_{j\in\mathcal{N}_p}\lambda_{pj}^k}_{:=\mu_p^k} - 2\rho\sum_{j\in\mathcal{N}_p}z_{pj}^k\Big)^\top x_p + \rho D_p\Big(\|x_p\|^2 + \|z_{pj}^k\|^2\Big)\Bigg) \qquad \text{(A.27)}\\
&= \sum_{p=1}^{P}\Bigg(\Big(\mu_p^k - 2\rho\sum_{j\in\mathcal{N}_p}z_{pj}^k\Big)^\top x_p + \rho D_p\Big(\|x_p\|^2 + \|z_{pj}^k\|^2\Big)\Bigg). \qquad \text{(A.28)}
\end{aligned}$$

Therefore, updating $\bar{x}$ as in (A.9) amounts to

$$\bar{x}^{k+1} = \arg\min_{\bar{x}}\ \sum_{p=1}^{P}\left(f_p(x_p) + \Big(\mu_p^k - 2\rho\sum_{j\in\mathcal{N}_p}z_{pj}^k\Big)^\top x_p + \rho D_p\|x_p\|^2\right), \qquad \text{(A.29)}$$

where we dropped $\|z_{pj}^k\|^2$ in the last term in (A.28), since it is independent of the problem variable $\bar{x}$. Problem (A.29) yields P independent optimization problems, each depending only on an x_p, which can be executed in parallel. The problem associated to node p is

$$x_p^{k+1} = \arg\min_{x_p}\ f_p(x_p) + \Big(\mu_p^k - 2\rho\sum_{j\in\mathcal{N}_p}z_{pj}^k\Big)^\top x_p + \rho D_p\|x_p\|^2$$

$$= \arg\min_{x_p} \; f_p(x_p) + \rho D_P \left\| x_p - \Big(\frac{1}{D_p} \sum_{j \in \mathcal{N}_p} z_{pj}^k - \frac{1}{2\rho D_p} \mu_p^k \Big) \right\|^2 \tag{A.30}$$

$$= \mathrm{prox}_{\tau_p f_p} \Big(\frac{1}{D_p} \sum_{j \in \mathcal{N}_p} z_{pj}^k - \tau_p \mu_p^k \Big) \tag{A.31}$$

$$= \mathrm{prox}_{\tau_p f_p} \Big(\frac{1}{2D_p} \sum_{j \in \mathcal{N}_p} (x_p^k + x_j^k) - \tau_p \mu_p^k \Big) . \tag{A.32}$$

From (A.30) to (A.31), we used the definition of the prox operator (2.34) and $\tau_p = 1/(2\rho D_p)$. From (A.31) to (A.32), we replaced z_{pj}^k as in (A.17).

Update of the dual variables. Each node p does not need to know each individual λ_{ij} associated to its incident edges. In fact, as shown in (A.27), it only need to know $\mu_p^k = \sum_{j \in \mathcal{N}_p} \lambda_{pj}^k$. According to (A.18), this variable is updated as

$$\mu_p^{k+1} = 2 \sum_{j \in \mathcal{N}_p} \lambda_{pj}^{k+1} = \underbrace{2 \sum_{j \in \mathcal{N}_p} \lambda_{pj}^k}_{\mu_p^k} + \rho \sum_{j \in \mathcal{N}_p} (x_p^{k+1} - x_j^{k+1}) = \mu_p^k + \frac{1}{2\tau_p} \Big(x_p^{k+1} - \frac{1}{D_p} \sum_{j \in \mathcal{N}_p} x_j^{k+1} \Big) . \tag{A.33}$$

We thus see that (A.33) corresponds to step 4 in Algorithm 2, while (A.32) corresponds to step 3.

Appendix B

Some Conjugate Functions

In this appendix, we compute some conjugate functions that appear throughout the thesis.

ℓ_1-norm plus quadratic regularization. In Subsection 3.2.2, we reformulate BP (3.4) as a problem in the global class (G). That reformulation uses duality and, in (3.19), we use the convex conjugate of the function $h(x) = \|x\|_1 + (c/2)\|x\|^2$, where the term $\|x\|^2$ plays the role of a regularization function. We now show that the convex conjugate of h has a closed-form expression. Suppose $x \in \mathbb{R}^n$. We have

$$\begin{aligned} h^\star(\lambda) &= \sup_x \ \lambda^\top x - \|x\|_1 - \frac{c}{2}\|x\|^2 \\ &= -\inf_x \ \|x\|_1 + \frac{c}{2}\|x\|^2 - \lambda^\top x \\ &= -\sum_{i=1}^{n} \inf_{x_i} \ |x_i| + \frac{c}{2}x_i^2 - \lambda_i x_i \,. \end{aligned} \tag{B.1}$$

Applying the optimality condition for convex problems to the problem in the ith component,

$$0 \in \partial|x_i| + c\,x_i - \lambda_i \,. \tag{B.2}$$

When $x_i > 0$, $\partial|x_i| = \{1\}$, and (B.2) becomes $x_i = (\lambda_i - 1)/c$. This happens when $\lambda_i > 1$, otherwise the expression would give a negative x_i. Similarly, when $x_i < 0$, $\partial|x_i| = \{-1\}$, and (B.2) becomes $x_i = (\lambda_i + 1)/c$. This expression is negative when $\lambda_i < -1$. Finally, when $x_i = 0$, $\partial|x_i| = [-1, 1]$, and (B.2) becomes the condition under which $x_i = 0 \colon |\lambda_i| \leq 1$. This explains expression (3.20).

ℓ_2-norm plus quadratic regularization. Here, we derive expression (3.45), which is a closed-form expression for the conjugate of the function $h(x) = \|x\| + (c/2)\|x\|^2$, where $\|\cdot\|$ is the ℓ_2-norm.

The convex conjugate of h is

$$h^\star(\lambda) = \sup_x \ \lambda^\top x - \|x\| - \frac{c}{2}\|x\|^2 \tag{B.3}$$

$$= -\inf_x \ \|x\| + \frac{c}{2}\|x\|^2 - \lambda^\top x\,. \tag{B.4}$$

The subgradient of the norm function is

$$\partial\|x\| = \begin{cases} B(0,1) & , x = 0 \\ \frac{x}{\|x\|} & , x \neq 0\,, \end{cases}$$

where $B(0,1) = \{x \,:\, \|x\| \leq 1\}$ is the ball with radius 1, centered at the origin. The optimality conditions for (B.4) then tell us that $x = 0$ if $\|\lambda\| \leq 1$ and that, for $x \neq 0$,

$$0 = \frac{x}{\|x\|} + c\,x - \lambda\,. \tag{B.5}$$

From (B.5), we first find the norm of x and then compute an expression for x. To find the norm of x, first rewrite (B.5) as $\lambda = (1/\|x\| + c)x$, and compute the squared norm of both sides of the equation. This yields

$$\left(\frac{1}{\|x\|} + c\right)^2 \|x\|^2 = \|\lambda\|^2 \qquad \Longleftrightarrow \qquad 1 + 2c\|x\| + c^2\|x\|^2 = \|\lambda\|^2\,,$$

which is a quadratic expression on $\|x\|^2$. Solving the quadratic equation, gives us $\|x\| = (\|\lambda\| - 1)/c$, which is positive because $\|\lambda\| > 1$. Replacing in (B.5) gives

$$x = \frac{1}{c}\Big(1 - \frac{1}{\|\lambda\|}\Big)\lambda\,.$$

To compute the value (B.4), just take the inner product of (B.5) with x and subtract $(c/2)\|x\|^2$ to both sides of the equation. This gives

$$-\frac{c}{2}\|x\|^2 = \|x\| + \frac{c}{2}\|x\|^2 - \lambda^\top x\,.$$

Using the expression for the norm of x, we get

$$h^\star(\lambda) = \frac{1}{2c}\Big(\|\lambda\|^2 - 2\|\lambda\| + 1\Big)\,,$$

for $\|\lambda\| > 1$. This explains (3.45).

Appendix C

ADMM-based Algorithm For The Connected Class: Derivation

In this appendix, we derive Algorithm 4, an ADMM-based algorithm presented in Chapter 4 that was proposed in [47] to solve (P) with a star-shaped variable, but that can be easily generalized for a generic connected variable. The purpose of this appendix is to derive that algorithm. It applies the 2-block ADMM to the reformulation (4.20), reproduced here for convenience:

$$\begin{array}{ll} \underset{\{\bar{x}_l\}_{l=1}^n, \{\bar{z}_l\}_{l=1}^n}{\text{minimize}} & f_1(x_{S_1}^{(1)}) + f_1(x_{S_2}^{(2)}) + \cdots + f_1(x_{S_P}^{(P)}) \\ \text{subject to} & x_l^{(p)} = z_l^{\{p,j\}}\,, \quad l \in S_p \cap S_j\,, \quad j \in \mathcal{N}_p\,, \quad p = 1, \ldots, P\,, \end{array} \tag{C.1}$$

whose variable consists of $\bar{x} := \{\bar{x}_l\}_{l=1}^n$, and $\bar{z} := \{\bar{z}_l\}_{l=1}^n$. The augmented Lagrangian of (C.1) is

$$L_\rho(\bar{x}, \bar{z}; \bar{\lambda}) = \sum_{p=1}^P f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{p=1}^P \sum_{j \in \mathcal{N}_p} \sum_{l \in S_p \cap S_j} \left[\lambda_l^{pj\top}(x_l^{(p)} - z_l^{\{p,j\}}) + \frac{\rho}{2}\Big\|x_l^{(p)} - z_l^{\{p,j\}}\Big\|^2 \right], \tag{C.2}$$

Note that λ_l^{pj} and λ_l^{jp} are associated to different constraints. The 2-block ADMM (2.18)-(2.20) applied to this problem translates into

$$\bar{x}^{k+1} = \arg\min_{\bar{x}} \; L_\rho(\bar{x}, \bar{z}^k; \bar{\lambda}^k) \tag{C.3}$$

$$\bar{z}^{k+1} = \arg\min_{\bar{z}} \; L_\rho(\bar{x}^{k+1}, \bar{z}; \bar{\lambda}^k) \tag{C.4}$$

$$\lambda_l^{pj,k+1} = \lambda_l^{pj,k} + \rho\Big(x_l^{(p),k+1} - z_l^{\{p,j\},k+1}\Big) \tag{C.5}$$

$$\lambda_l^{jp,k+1} = \lambda_l^{jp,k} + \rho\Big(x_l^{(j),k+1} - z_l^{\{p,j\},k+1}\Big)\,. \tag{C.6}$$

As in the derivation of Algorithm 2 in Appendix A, we first analyze the minimization with respect to $\bar{z}$, (C.4); then, we analyze the minimization with respect to $\bar{x}$, (C.3); and, finally, we will see how to simplify the updates of the dual variables (C.5) and (C.6).

Minimization in $\bar{z}$. No function f_p depends on any component of $\bar{z}$, which means that (C.4) is an unconstrained quadratic program and, thus, has a closed-form solution. Furthermore, it decomposes across each component. In particular the minimization with respect to $z_l^{\{p,j\}} = z_l^{\{j,p\}}$ is

$$\begin{aligned}
z_l^{\{p,j\},k+1} &= \underset{z_l^{\{p,j\}}}{\arg\min}\ \lambda_l^{pj,k\top}(x_l^{(p),k+1} - z_l^{\{p,j\}}) + \frac{\rho}{2}\left\|x_l^{(p),k+1} - z_l^{\{p,j\}}\right\|^2 \\
&\qquad\qquad + \lambda_l^{jp,k\top}(x_l^{(j),k+1} - z_l^{\{p,j\}}) + \frac{\rho}{2}\left\|x_l^{(j),k+1} - z_l^{\{p,j\}}\right\|^2 \\
&= \underset{z_l^{\{p,j\}}}{\arg\min}\ -\left(\lambda_l^{pj,k} + \lambda_l^{jp,k}\right)^\top z_l^{\{p,j\}} + \rho\left\|z_l^{\{p,j\}}\right\|^2 - \rho\, x_l^{(p),k+1\top} z_l^{\{p,j\}} - \rho\, x_l^{(j),k+1\top} z_l^{\{p,j\}} \\
&= \underset{z_l^{\{p,j\}}}{\arg\min}\ -\left(\lambda_l^{pj,k} + \lambda_l^{jp,k} + \rho(x_l^{(p),k+1} + x_l^{(j),k+1})\right)^\top z_l^{\{p,j\}} + \rho\left\|z_l^{\{p,j\}}\right\|^2 \\
&= \frac{\lambda_l^{pj,k} + \lambda_l^{jp,k}}{2\rho} + \frac{x_l^{(p),k+1} + x_l^{(j),k+1}}{2}\,. \qquad \text{(C.7)}
\end{aligned}$$

Replacing (C.7) in (C.5) and (C.6), we get, respectively,

$$\begin{aligned}
\lambda_l^{pj,k+1} &= \lambda_l^{pj,k} + \rho\left(x_l^{(p),k+1} - \frac{\lambda_l^{pj,k} + \lambda_l^{jp,k}}{2\rho} - \frac{x_l^{(p),k+1} + x_l^{(j),k+1}}{2}\right) \\
&= \frac{\lambda_l^{pj,k} - \lambda_l^{jp,k}}{2} + \frac{\rho}{2}(x_l^{(p),k+1} - x_l^{(j),k+1}) \qquad \text{(C.8)} \\
\lambda_l^{jp,k+1} &= \lambda_l^{jp,k} + \rho\left(x_l^{(j),k+1} - \frac{\lambda_l^{pj,k} + \lambda_l^{jp,k}}{2\rho} - \frac{x_l^{(p),k+1} + x_l^{(j),k+1}}{2}\right) \\
&= \frac{\lambda_l^{jp,k} - \lambda_l^{pj,k}}{2} + \frac{\rho}{2}(x_l^{(j),k+1} - x_l^{(p),k+1}) \qquad \text{(C.9)}
\end{aligned}$$

Summing (C.8) with (C.9), we get

$$\lambda_l^{pj,k+1} + \lambda_l^{jp,k+1} = 0\,,$$

which holds for all $k \geq 1$. Let assume that it also holds for $k = 0$, i.e., λ_l^{pj} and λ_l^{jp} are initialized with symmetric values. Then, the first term in (C.7) is zero, and updating $z_l^{\{p,j\}}$ simplifies to

$$z_l^{\{p,j\},k+1} = \frac{x_l^{(p),k+1} + x_l^{(j),k+1}}{2}\,. \qquad \text{(C.10)}$$

Minimization in $\bar{x}$. We now turn to the minimization in $\bar{x}$ (C.3). If we fix each $z_l^{\{p,j\}}$ at $z_l^{\{p,j\},k}$ and each λ_l^{pj} at $\lambda_l^{pj,k}$, the augmented Lagrangian (C.2) is the sum of p terms, where the pth term depends only on x_{S_p}. Thus, problem (C.3) decomposes into P optimization problems that can be solved in parallel. The problem associated with node p is

$$
\begin{aligned}
x_{S_p}^{(p),k+1} &= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{j\in\mathcal{N}_p}\sum_{l\in S_p\cap S_j}\left[\lambda_l^{pj,k^\top}(x_l^{(p)} - z_l^{\{p,j\},k}) + \frac{\rho}{2}\Big\|x_l^{(p)} - z_l^{\{p,j\},k}\Big\|^2\right] \\
&= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{j\in\mathcal{N}_p}\sum_{l\in S_p\cap S_j}\left[\lambda_l^{pj,k^\top}x_l^{(p)} + \frac{\rho}{2}\Big\|x_l^{(p)}\Big\|^2 - \rho z_l^{\{p,j\},k^\top}x_l^{(p)}\right] \qquad \text{(C.11)} \\
&= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{l\in S_p}\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\left[\Big(\lambda_l^{pj,k} - \rho z_l^{\{p,j\},k}\Big)^\top x_l^{(p)} + \frac{\rho}{2}\Big\|x_l^{(p)}\Big\|^2\right] \qquad \text{(C.12)} \\
&= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{l\in S_p}\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\Big(\lambda_l^{pj,k} - \rho z_l^{\{p,j\},k}\Big)^\top x_l^{(p)} + \frac{\rho}{2}\sum_{l\in S_p}\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\Big\|x_l^{(p)}\Big\|^2 \\
&= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{l\in S_p}\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\Big(\lambda_l^{pj,k} - \rho z_l^{\{p,j\},k}\Big)^\top x_l^{(p)} + \frac{\rho}{2}\sum_{l\in S_p} D_{p,l}\Big\|x_l^{(p)}\Big\|^2 \qquad \text{(C.13)} \\
&= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{l\in S_p}\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\Big(\lambda_l^{pj,k} - \frac{\rho}{2}(x_l^{(p),k} + x_l^{(j),k})\Big)^\top x_l^{(p)} + \frac{\rho}{2}\sum_{l\in S_p} D_{p,l}\Big\|x_l^{(p)}\Big\|^2 \qquad \text{(C.14)} \\
&= \underset{x_{S_p}^{(p)}}{\arg\min}\ f_p\Big(x_{S_p}^{(p)}\Big) + \sum_{l\in S_p}\Big(\gamma_l^{(p),k} - \frac{\rho}{2}(D_{p,l}x_l^{(p),k} + \sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l} x_l^{(j),k})\Big)^\top x_l^{(p)} + \frac{\rho}{2}\sum_{l\in S_p} D_{p,l}\Big\|x_l^{(p)}\Big\|^2, \qquad \text{(C.15)}
\end{aligned}
$$

where $D_{p,l}$ is the degree of node p in the subgraph induced by x_l, $\mathcal{G}_l$. From (C.11) to (C.12), we used the fact that, for a fixed node p, $\sum_{j\in\mathcal{N}_p}\sum_{l\in S_p\cap S_j}(\cdot) = \sum_{l\in S_p}\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}(\cdot)$. From (C.13) to (C.14), we used (C.10). Finally, from (C.14) to (C.15), we defined $\gamma_l^{(p),k} = \sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\lambda_l^{pj,k}$.

Update of the dual variables. Note from (C.15) that node p depends only on $\gamma_l^{(p),k} = \sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\lambda_l^{pj,k}$ and not on the individual λ_l^{pj}s. Using (C.5), the update of $\gamma_l^{(p),k}$ comes as

$$
\begin{aligned}
\gamma_l^{(p),k+1} &= \sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\lambda_l^{pj,k+1} \\
&= \underbrace{\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\lambda_l^{pj,k}}_{=\gamma_l^{(p),k}} + \rho\Big(x_l^{(p),k+1} - z_l^{\{p,j\},k+1}\Big) \\
&= \gamma_l^{(p),k} + \rho\sum_{j\in\mathcal{N}_p\cap\mathcal{V}_l}\Big(x_l^{(p),k+1} - z_l^{\{p,j\},k+1}\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \gamma_l^{(p),k} + \rho \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} \Big(x_l^{(p),k+1} - \frac{x_l^{(p),k+1} + x_l^{(j),k+1}}{2} \Big) \\
&= \gamma_l^{(p),k} + \frac{\rho}{2} \sum_{j \in \mathcal{N}_p \cap \mathcal{V}_l} \Big(x_l^{(p),k+1} - x_l^{(j),k+1} \Big), \qquad \text{(C.16)}
\end{aligned}
$$

where we have used (C.10). We thus see that (C.16) corresponds to step 6 of Algorithm 4, while (C.15) corresponds to step 4.

Bibliography

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[2] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2nd ed., 1999.

[3] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*. MPS-SIAM Series on Optimization, 2001.

[4] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Intern. Conf. Information Proc. in Sensor Networks (IPSN)*, pp. 20–27, 2004.

[5] D. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," tech. rep., LIDS-2848, 2010.

[6] I. Akyildiz, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, pp. 393–422, 2002.

[7] M. DeGroot, "Reaching a consensus," *J. American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[8] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53, pp. 65–78, 2004.

[9] T. Erseghe, D. Zennaro, E. Dall'Anese, and L. Vangelista, "Fast consensus by the alternating direction multipliers method," *IEEE Trans. Signal Processing*, vol. 59, no. 11, pp. 5523–5537, 2011.

[10] A. Olshevsky and J. Tsitsiklis, "Convergence speed in distributed consensus and averaging," *SIAM Review*, vol. 53, no. 4, pp. 747–772, 2011.

[11] B. Oreshkin, M. Coates, and M. Rabbat, "Optimization and analysis of distributed averaging with short node memory," *IEEE Trans. Signal Processing*, vol. 58, no. 5, pp. 2850–2865, 2010.

[12] S. Kar and J. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Processing*, vol. 57, no. 1, pp. 355–369, 2009.

[13] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[14] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[15] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statistical. Soc., Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[17] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[18] D. Šlijak, *Large-Scale Dynamic Systems*. Dover Publications, 2007.

[19] E. Camponogara, D. Jia, B. Krogh, and S. Talukdar, "Distributed model predictive control," *IEEE Control Syst. Mag.*, vol. 22, no. 1, pp. 44–52, 2002.

[20] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[21] B. Krishnamachari, *Networking Wireless Sensors*. Cambridge University Press, 2005.

[22] D. Han and X. Yuan, "A note on the alternating direction method of multipliers," *J. Optimization Theory and Appl.*, vol. 155, no. 1, pp. 227–238, 2012.

[23] R. Glowinski and A. Marrocco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de dirichelet non linéaires," *Revue Française d'Automatique, Informatique, et Recherche Opérationelle*, vol. 9, no. 2, pp. 41–76, 1975.

[24] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," *Computers and Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[25] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc WSNs with noisy links - Part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[26] H. Zhu, G. Giannakis, and A. Cano, "Distributed in-network channel decoding," *IEEE Trans. Signal Processing*, vol. 57, no. 10, pp. 3970–3983, 2009.

[27] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

[28] M. Snir, S. Otto, S. Hess-Lederman, D. Walker, and J. Dongarra, *MPI: The Complete Reference*. MIT Press, 1996.

[29] P. Fischione, P. Park, and K. Johansson, *Wireless Network Based Control*, ch. Design Principles of Wireless Sensor Network Protocols for Control Applications. Springer, 2011.

[30] J. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer, 2004.

[31] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[32] B. Bollobás, *Modern Graph Theory*. Springer, 2008.

[33] C. Tan, D. Palomar, and M. Chiang, "Distributed optimization of coupled systems with applications to network utility maximization," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, pp. 981–984, 2006.

[34] F. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Info. Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[36] G. Dantzig and P. Wolfe, "Decomposition principle for linear programs," *Operations Research*, vol. 8, no. 1, pp. 101–111, 1960.

[37] J. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numerische Mathematik*, vol. 4, pp. 238–252, 1962.

[38] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, no. 3, pp. 399–417, 1963.

[39] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. AC-31, no. 9, pp. 803–812, 1986.

[40] C. Soares, J. Xavier, and J. Gomes, "DCOOL-NET: Distributed cooperative localization for sensor networks." preprint: `http://arxiv.org/abs/1211.7277`, 2012.

[41] P. Forero, A. Cano, and G. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 707–724, 2011.

[42] J. Bazerque and G. Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity," *IEEE Trans. Signal Processing*, vol. 58, no. 3, pp. 1847–1862, 2010.

[43] P. Forero, A. Cano, and G. Giannakis, "Consensus-based distributed support vector machines," *J. Machine Learning Research*, vol. 11, pp. 1663–1707, 2010.

[44] A. Navia-Vázquez, D. Gutiérrez-González, E. Parrado-Hernández, and J. Navarro-Abellán, "Distributed support vector machines," *IEEE Trans. Neural Networks*, vol. 17, no. 4, pp. 1091–1097, 2006.

[45] K. Flouri, B. Beferull-Lozano, and P. Tsakalides, "Distributed consensus algorithms for SVM training in wireless sensor networks," in *European Signal Proc. Conf. (Eusipco)*, 2008.

[46] C. Conte, T. Summers, M. Zeilinger, M. Morari, and C. Jones, "Computational aspects of distributed optimization in model predictive control," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 6819–6824, 2012.

[47] V. Kekatos and G. Giannakis, "Distributed robust power system state estimation," *IEEE Trans. Power Systems*, vol. 28, no. 2, pp. 1617–1626, 2012.

[48] M. Kraning, E. Chu, J. Lavaei, and S. Boyd, "Dynamic network energy management via proximal message passing," *Found. Trends in Optimization*, vol. 1, no. 2, pp. 70–122, 2013.

[49] T. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method." preprint: `http://arxiv.org/abs/1304.5590`, 2013.

[50] E. Dall'Anese, H. Zhu, and G. Giannakis, "Distributed optimal power flow for smart microgrids," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1464–1475, 2013.

[51] I. Necoara, V. Nedelcu, and I. Dumitrache, "Parallel and distributed optimization methods for estimation and control in networks," *Journal of Process Control*, vol. 21, pp. 756–766, 2011.

[52] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[53] L. Vandenberghe, "Dual decomposition," Spring 2011-12. Lecture Notes, Optimization Methods for Large-Scale Systems (EE-236C), UCLA.

[54] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[55] L. Vandenberghe, "Subgradient method," Spring 2011-12. Lecture Notes, Optimization Methods for Large-Scale Systems (EE-236C), UCLA.

[56] S. Boyd and A. Mutapic, "Subgradient methods," Winter 2007. Lecture Notes, Convex Optimization II (EE364b), Stanford University.

[57] A. Beck, *Convergence Rate Analysis of Gradient Based Algorithms*. PhD thesis, Tel-Aviv University, 2002.

[58] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Im. Sc.*, vol. 2, no. 1, pp. 183–202, 2009.

[59] A. Beck and M. Teboulle, *Convex Optimization in Signal Processing and Communications*, ch. Gradient-based algorithms with applications to signal-recovery problems. Cambridge University Press, 2010.

[60] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization." Submitted to SIAM J. Optim., 2008.

[61] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program., Ser. A*, pp. 1–39, 2013.

[62] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[63] A. d'Aspremont, "Smooth optimization with approximate gradient," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1171–1183, 2008.

[64] L. Vandenberghe, "Fast proximal gradient methods," Spring 2011-12. Lecture Notes, Optimization Methods for Large-Scale Systems (EE-236C), UCLA.

[65] A. Zakarian, *Nonlinear Jacobi and ϵ-relaxation methods for parallel network optimization*. PhD thesis, University of Wisconsin, Madison, 1995.

[66] J. Mota, "Distributed algorithms for sparse approximation," Master's thesis, Instituto Superior Técnico, Technical University of Lisbon, Portugal, 2008. `http://users.isr.ist.utl.pt/~jmota/`.

[67] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optimization Theory and Appl.*, vol. 109, no. 3, pp. 475–494, 2001.

[68] M. Hestenes, "Multiplier and gradient methods," *J. Optimization Theory and Appl.*, vol. 4, no. 5, pp. 303–320, 1969.

[69] M. Powell, *Optimization*, ch. A method for nonlinear constraints in minimization problems. Academic Press, 1969.

[70] R. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research*, vol. 1, no. 2, pp. 97–116, 1976.

[71] J. Eckstein, *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*. PhD thesis, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1989.

[72] D. Bertsekas, "Multiplier methods: A survey," *Automatica*, vol. 12, pp. 133–145, 1976.

[73] D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.

[74] R. Rockafellar, "The multiplier method of Hestenes and Powell applied to convex programming," *J. Optimization Theory and Appl.*, vol. 12, no. 6, pp. 555–562, 1973.

[75] J. Eckstein, "Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," tech. rep., Rutcor Research Report, 32-2012, 2012.

[76] B. Martinet, "Regularisation d'inequations variationelles par approximations successives," *Revue Française d'Informatique et de Recherche Operationelle*, vol. 4, no. R-3, pp. 154–158, 1970.

[77] B. Martinet, "Determination approchée d'un point fixe d'une application pseudo-contractante. cas de l'application prox," *Comptes Rendus de l'Academie des Sciences (Paris)*, vol. 274, no. A, pp. 163–165, 1972.

[78] R. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control and Optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[79] J. Douglas and H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Trans. American Math. Society*, vol. 82, pp. 421–439, 1956.

[80] P. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Num. Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[81] D. Gabay, *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*, ch. Applications of the method of multipliers to variational inequalities. North-Holland: Amsterdam, 1983.

[82] J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, pp. 293–318, 1992.

[83] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions." preprint: `http://arxiv.org/abs/1112.2295`, 2011.

[84] J. Eckstein and D. Bertsekas, "An alternating direction method for linear programming," tech. rep., LIDS-P-1967, 1990.

[85] B. He and X. Yuan, "On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method," *SIAM J. Numer. Anal.*, vol. 50, no. 2, pp. 700–709, 2012.

[86] B. He and X. Yuan, "On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers." preprint: `http://www.optimization-online.org/DB_HTML/2012/01/3318.html`, 2012.

[87] T. Goldstein, B. O'Donoghue, and S. Setzer, "Fast alternating direction optimization methods," tech. rep., CAM report 12-35, UCLA, 2012.

[88] D. Boley, "Linear convergence of ADMM on a model problem," tech. rep., TR 12-009, Dept. Computer Science, University of Minnesota, 2012.

[89] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems." preprint: `http://arxiv.org/abs/1306.2454`, 2013.

[90] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," tech. rep., Rice University, Dept. Computational and Applied Mathematics, 2012.

[91] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control and Optimization*, vol. 29, no. 1, pp. 119–138, 1991.

[92] D. Goldfarb and S. Ma, "Fast multiple splitting algorithms for convex optimization," tech. rep., Department of IEOR, Columbia Univ., 2009.

[93] D. Goldfarb, S. Ma, and K. Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex functions," tech. rep., Department of IEOR, Columbia Univ., 2010.

[94] E. Wei and A. Ozdaglar, "On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers." preprint: `http://arxiv.org/abs/1307.8254`, 2013.

[95] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Im. Proc.*, vol. 20, no. 3, pp. 681–695, 2011.

[96] A. Martins, M. Figueiredo, P. Aguiar, N. Smith, and E. Xing, "An augmented Lagrangian approach to constrained MAP inference," in *Proc. 28th Intern. Conf. Machine Learning, Bellevue, WA, USA*, 2011.

[97] B. He, M. Tao, and X. Yuan, "Alternating direction method with Gaussian back substitution for separable convex programming," *SIAM J. Optim.*, vol. 22, no. 2, pp. 313–340, 2012.

[98] M. Hong and Z. Luo, "On the linear convergence of the alternating direction method of multipliers." preprint: `http://arxiv.org/abs/1208.3922`, 2013.

[99] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "Distributed basis pursuit." preprint: `http://arxiv.org/abs/1009.1128v2`, version 2, July, 2011.

[100] A. Nedić and A. Ozdaglar, "On the rate of convergence of distributed subgradient methods for multi-agent optimization," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 4711–4716, 2007.

[101] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, 2009.

[102] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Math. Program., Ser. B*, vol. 129, pp. 255–284, 2011.

[103] A. Nedić and A. Ozdaglar, *Convex Optimization in Signal Processing and Communications*, ch. Cooperative distributed multi-agent optimization. Cambridge University Press, 2010.

[104] S. Ram, A. Nedić, and V. Veeravalli, "Asynchronous gossip algorithms for stochastic optimization," in *Joint 48th IEEE Conf. Decision and Control and 28th Chinese Control Conf., Shanghai, P.R. China*, pp. 3581–3586, 2009.

[105] J. Tsitsiklis, *Problems in decentralized decision making and computation.* PhD thesis, Massachussets Institute of Technology, 1984.

[106] M. Rabbat and R. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE J. Selected Areas in Communications*, vol. 23, no. 4, pp. 798–808, 2005.

[107] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM J. Optim.*, vol. 20, no. 3, pp. 1157–1170, 2009.

[108] M. Rabbat, R. Nowak, and J. Bucklew, "Generalized consensus algorithms in networked systems with erasure links," in *IEEE Workshop Signal Proc. Advances in Wireless Comunications*, pp. 1088–1092, 2005.

[109] B. Johansson, T. Keviczky, M. Johansson, and K. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 4185–4190, 2008.

[110] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Trans. Autom. Control*, vol. 57, no. 1, pp. 151–164, 2012.

[111] M. Zhu and S. Martínez, "On distributed optimization under inequality and equality constraints via penalty primal-dual methods," in *American Control Conf.*, pp. 2434–2439, 2010.

[112] J. Chen and A. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012.

[113] D. Jakovetić, J. Xavier, and J. Moura, "Fast distributed gradient methods." preprint: `http://arxiv.org/abs/1112.2972`, 2011.

[114] J. Duchi, A. Argawal, and M. Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, 2012.

[115] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Math. Program., Ser. B*, vol. 120, pp. 221–259, 2009.

[116] E. Ghadimi, I. Shames, and M. Johansson, "Accelerated gradient methods for networked optimization." preprint: `http://arxiv.org/abs/1211.2132`, 2012.

[117] B. Polyak, *Introduction to Optimization*. Optimization Software, 1987.

[118] A. Ruszczynski, "Augmented Lagrangian decomposition for sparse convex optimization," *Inter. Inst. Applied Systems Analysis*, 1992.

[119] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "Basis pursuit in sensor networks," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, pp. 2916–2919, 2011.

[120] D. Jakovetić, J. Xavier, and J. Moura, "Cooperative convex optimization in networked systems: Augmented Lagrangian algorithms with directed gossip communication," *IEEE Trans. Signal Processing*, vol. 59, no. 8, pp. 3889–3902, 2011.

[121] I. Necoara, "Random coordinate descent algorithms for multi-agent convex optimization over networks," *IEEE Trans. Autom. Control*, vol. 58, no. 8, pp. 2001–2012, 2013.

[122] J. Moreau, "Fonctions convexes duales et points proximaux dans un espace Hilbertien," *Comptes Rendus de l'Academie des Sciences (Paris), Série A*, vol. 255, pp. 2897–2899, 1962.

[123] P. Combettes and J. Pesquet, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ch. Proximal Splitting Methods in Signal Processing, pp. 185–212. Springer, New York, 2011.

[124] G. Mateos, J. Bazerque, and G. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.

[125] W. Shi, Q. Ling, G. Wu, and W. Yin, "Linearly convergent decentralized consensus optimization with the alternating direction method of multipliers," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, pp. 4613–4617, 2013.

[126] W. Shi, Q. Ling, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization." preprint: `http://arxiv.org/abs/1307.5561`, 2013.

[127] Q. Ling, M. Tao, W. Yin, and X. Yuan, "A multi-block alternating direction method with parallel splitting for decentralized consensus optimization," *EURASIP J. Wireless Comm. and Networking*, vol. 338, 2012.

[128] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers." preprint: `http://arxiv.org/abs/1303.2837`, 2013.

[129] F. Kelly, "Charging and rate control for elastic traffic," *European Trans. Telecommunications*, vol. 8, pp. 33–37, 1997.

[130] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[131] S. Low and D. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," *IEEE/ACM Trans. Networking*, vol. 7, no. 6, pp. 861–874, 1999.

[132] S. Low, L. Peterson, and L. Wang, "Understanding Vegas: a duality model," *Journal of the ACM*, vol. 49, no. 2, pp. 207–235, 2002.

[133] S. Shakkottai and R. Srikant, "Network optimization and control," *Found. Trends Networking*, vol. 2, no. 3, pp. 271–379, 2007.

[134] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: a mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[135] S. Athuraliya and S. Low, "Optimization flow control with Newton-like algorithm," *J. Telecomm. Syst.*, vol. 15, pp. 345–358, 2000.

[136] D. Bertsekas, "Centralized and distributed Newton methods for network optimization and extensions," tech. rep., LIDS-2866, 2011.

[137] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed Newton method for network utility maximization," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 1816–1821, 2010.

[138] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed Newton method for network utility maximization, I: Algorithm," tech. rep., LIDS-2832, 2011.

[139] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle, "Optimal distributed gradient methods for network resource allocation problems." preprint: `http://web.mit.edu/asuman/www/documents/NUM-FGM.pdf`, 2013.

[140] L. Zadeh and B. Whalen, "On optimal control and linear programming," *IRE Trans. Autom. Control*, vol. 7, no. 4, pp. 45–46, 1962.

[141] A. Propoi, "Use of LP methods for synthesizing sampled-data automatic systems," *Automn Remote Control*, vol. 24, 1963.

[142] C. García, D. Prett, and M. Morari, "Model predictive control: Theory and practice – a survey," *Automatica*, vol. 25, no. 3, pp. 335–348, 1989.

[143] L. Acar, "Some examples for the decentralized receding horizon control," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 1356–1359, 1992.

[144] H. Fawal, D. Georges, and G. Bornard, "Optimal control of complex irrigation systems via decomposition-coordination and the use of augmented Lagrangian," in *IEEE Conf. on Systems, Man, and Cybernetics*, pp. 3874–3879, 1998.

[145] D. Jia and B. Krogh, "Distributed model predictive control," in *American Control Conf.*, pp. 2767–2772, 2001.

[146] T. Keviczky, F. Borrelli, and G. Balas, "Decentralized receding horizon control for large scale dynamically decoupled systems," *Automatica*, vol. 42, pp. 2105–2115, 2006.

[147] A. Venkat, J. Rawlings, and S. Wright, "Stability and optimality of distributed model predictive control," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 6680–6685, 2005.

[148] X. Zhang, M. Burger, X. Bresson, and S. Osher, "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM J. Imaging Sciences*, vol. 3, no. 3, pp. 253–276, 2010.

[149] X. Zhang, M. Burger, and S. Osher, "A unified primal-dual algorithm framework based on Bregman iteration," *J. Sci. Comput.*, vol. 46, pp. 20–46, 2011.

[150] Y. Wakasa, M. Arakawa, K. Tanaka, and T. Akashi, "Distributed model predictive control via dual decomposition," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 381–386, 2008.

[151] E. Camponogara and H. Scherer, "Distributed optimization for model predictive control of linear dynamic networks with control-input and output constraints," *IEEE Trans. Aut. Sc. Engin.*, vol. 8, no. 1, 2011.

[152] T. Summers and J. Lygeros, "Distributed model predictive consensus via the alternating direction method of multipliers," in *Allerton Conf. Communications, Control, and Computing*, pp. 79–84, 2012.

[153] G. Dantzig, *Linear Programming and Extensions.* Princeton University Press, 1963.

[154] R. Ahuja, T. Magnanti, and J. Orlin, "Some recent advances in network flows," *SIAM Review*, vol. 33, no. 2, pp. 175–219, 1991.

[155] A. Goldberg, E. Tardos, and R. Tarjan, "Network flow algorithms," tech. rep., CS-TR-216-89, Dept. Computer Science, Stanford University, CA, 1989.

[156] D. Bertsekas, *Network Optimization: Continuous and Discrete Models.* Athena Scientific, 1998.

[157] D. Bertsekas and D. El Baz, "Distributed asynchronous relaxation methods for convex network flow problems," *SIAM J. Control and Optimization*, vol. 25, no. 1, pp. 74–85, 1987.

[158] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Communications*, vol. 52, no. 7, pp. 1136–1144, 2004.

[159] J. Trdlička and Z. Hanzálek, "Distributed multi-commodity network flow algorithm for energy optimal routing in wireless sensor networks," *Radioengineering*, vol. 19, no. 4, pp. 579–588, 2010.

[160] A. Jadbabaie, A. Ozdaglar, and M. Zargham, "A distributed Newton method for network optimization," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 2736–2741, 2009.

[161] M. Zargham, A. Ribeiro, A. Jadbabaie, and A. Ozdaglar, "Accelerated dual descent for network optimization," in *American Control Conf.*, pp. 2663–2668, 2011.

[162] M. Zargham, A. Ribeiro, A. Ozdaglar, and A. Jadbabaie, "Accelerated dual descent for network flow optimization." preprint: `http://www.seas.upenn.edu/~zargham/ADDextended_PR1.pdf`, 2012.

[163] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Processing*, vol. 61, no. 10, pp. 2718–2723, 2013.

[164] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "D-ADMM: A distributed algorithm for compressed sensing and other separable optimization problems," in *IEEE Intern. Conf. Acoustics, Speech, and Sig. Processing (ICASSP)*, pp. 2869–2872, 2012.

[165] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "Consensus on colored networks," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 5116–5121, 2012.

[166] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "Distributed basis pursuit," *IEEE Trans. Signal Processing*, vol. 60, no. 4, pp. 1942–1956, 2012.

[167] M. Garey and D. Johnson, *Computers and Intractability.* W. H. Freeman and Co., 1979.

[168] F. Kuhn and R. Wattenhofer, "On the complexity of distributed graph coloring," in *in proceed. of Principles of distributed computing*, pp. 7–15, 2006.

[169] D. Leith and P. Clifford, "Convergence of distributed learning algorithms for optimal wireless channel allocation," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 2980–2985, 2006.

[170] K. Duffy, N. Connell, and A. Sapozhnikov, "Complexity analysis of a decentralised graph colouring algorithm," *Information Processing Letters*, vol. 107, pp. 60–63, 2008.

[171] N. Linial, "Locality in distributed graph algorithms," *SIAM J. Comput.*, vol. 21, no. 1, pp. 193–201, 1992.

[172] J. Kurose and K. Ross, *Computer networking: A top-down approach featuring the internet.* Addison Wesley, 3rd ed., 2005.

[173] J. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Conf. Mobile Computing Netw. (Mobicom)*, pp. 1–12, 2010.

[174] A. Jadbabaie, J. Lin, and S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[175] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Info. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[176] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[177] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[178] J. Predd, S. Kulkarni, and V. Poor, "A collaborative training algorithm for distributed learning," *IEEE Trans. Info. Theory*, vol. 55, no. 4, pp. 1856–1871, 2009.

[179] E. Candès, "Compressive sampling," pp. 1–20, European Math. Society, Proc. Inter. Congress of Mathematicians, Madrid, Spain, 2006.

[180] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[181] R. Baraniuk, "Compressive sensing," *IEEE Sig. Proc. Mag.*, vol. 24, no. 4, pp. 118–121, 2007.

[182] K. Bryan and T. Leise, "Making do with less: An introduction to compressed sensing," *SIAM Review*, vol. 55, no. 3, pp. 547–566, 2013.

[183] J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Info. Theory*, vol. 51, no. 10, pp. 3601–3608, 2005.

[184] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Info. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[185] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Info. Theory*, vol. 51, no. 3, pp. 1030–1051, 2006.

[186] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[187] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[188] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathem.*, vol. 61, no. 8, pp. 1025–1045, 2008.

[189] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus de l'Academie des Sciences (Paris), Série I*, no. 346, pp. 589–592, 2008.

[190] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[191] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Info. Theory*, vol. 56, no. 1, pp. 520–544, 2010.

[192] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Intern. Conf. Information Proc. in Sensor Networks (IPSN)*, pp. 134–142, 2006.

[193] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 92–101, 2008.

[194] V. Cevher, M. Duarte, and R. Baraniuk, "Distributed target localization via spatial sparsity," in *European Signal Proc. Conf. (Eusipco)*, 2008.

[195] A. Schmidt, *Scalable Sensor Network Field Reconstruction with Robust Basis Pursuit.* PhD thesis, Carnegie Mellon University, 2013.

[196] J. Romberg, R. Neelamani, C. Krohn, J. Krebs, M. Deffenbaugh, and J. Anderson, "Efficient seismic forward modeling using simultaneous random sources and sparsity," in *Soc. Expl. Geophysicists Annual Meeting*, pp. 2107–2110, 2008.

[197] M. Friedlander and P. Tseng, "Exact regularization of convex programs," *SIAM J. Optim.*, vol. 18, no. 4, pp. 1326–1350, 2007.

[198] O. Mangasarian and R. Meyer, "Nonlinear perturbation of linear programs," *SIAM J. Contr. Optim.*, vol. 17, no. 6, pp. 745–752, 1979.

[199] M. Friedlander, "Exact regularization of linear programs," tech. rep., TR-2005-31, Dept. Computer Science, Univ. of British Columbia, 2006.

[200] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[201] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 409–10, 1998.

[202] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[203] M. Penrose, *Random Geometric Graphs*. Oxford University Press, 2004.

[204] G. van Rossum *et al.*, "Python programming languange." `http://www.python.org/`.

[205] A. Hagberg, D. Schult, and P. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Python Science Conference (SciPy)*, pp. 11–15, 2008.

[206] W. Stein *et al.*, *Sage Mathematics Software (Version 5.8)*. The Sage Development Team, 2013. http://www.sagemath.org.

[207] A. Nedić, A. Ozdaglar, and P. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control*, vol. 55, no. 4, pp. 922–938, 2010.

[208] E. Berg and M. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Sci. Comput.*, vol. 31, no. 2, pp. 890–912, 2008.

[209] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[210] E. Berg, M. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and O. Yilmaz, "Sparco: a testing framework for sparse reconstruction," tech. rep., Dept. Computer Science, University of British Columbia, Vancouver, 2007.

[211] M. Raydan, "The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem," *SIAM J. Optim.*, vol. 7, no. 1, pp. 26–33, 1997.

[212] The Mathworks Inc., "Optimization toolbox." `http://www.mathworks.com/products/optimization/`, 2012.

[213] A. Frank and A. Asuncion, *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, 2010.

[214] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "Distributed optimization with local domains: Applications in MPC and network flows." submitted to IEEE Trans. Autom. Control, preprint: `http://arxiv.org/abs/1305.1885`, 2013.

[215] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "Distributed ADMM for model predictive control and congestion control," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 5110–5115, 2012.

[216] J. Mota, J. Xavier, P. Aguiar, and M. Püschel, "A unified algorithmic approach to distributed optimization." accepted at IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2013.

[217] R. Raffard, C. Tomlin, and S. Boyd, "Distributed optimization for cooperative agents: Application to formation flight," in *IEEE Intern. Conf. Decision and Control (CDC)*, pp. 2453–2459, 2004.

[218] P. Moroşan, R. Bourdais, D. Dumur, and J. Buisson, "Building temperature regulation using a distributed model predictive control," *Energy and Buildings*, vol. 42, pp. 1445–1452, 2010.

[219] A. Abur and A. Expósito, *Power System State Estimation*. Marcel Dekker, 2004.

[220] G. Giannakis, V. Kekatos, N. Gatsis, S.-J. Kim, H. Zhu, and B. Wollenberg, "Monitoring and optimization for power grids: a signal processing perspective," *IEEE Sig. Proc. Mag.*, vol. 30, no. 5, pp. 107–128, 2013.

[221] K. Baker, G. Hug, and X. Li, "Optimal integration of intermittent energy sources using distributed multi-step optimization," in *Power and Energy Soc. Gen. Meeting*, pp. 1–8, 2012.

[222] G. Hug, *Coodinated Power Flow Control to Enhance Steady-State Security in Power Systems*. PhD thesis, Swiss Federal Institue of Technology Zurich, 2008.

[223] M. Garey, R. Graham, and D. Johnson, "The complexity of computing Steiner minimal trees," *SIAM J. Appl. Math.*, vol. 32, no. 4, pp. 835–859, 1997.

[224] D. Williamson, "The primal-dual method for approximating algorithms," *Math. Program.*, vol. 91, no. B, pp. 447–478, 2002.

[225] M. Goemans and D. Williamson, *Approximation algorithms for NP-hard problems*, ch. The primal-dual method for approximation algorithms and its application to network design problems. PWS Publishing Company, 1997.

[226] G. Robins and A. Zelikovsky, "Improved Steiner tree approximation in graphs," in *ACM-SIAM Symposium Discrete Algs.*, pp. 770–779, 2000.

[227] L. Drummond, M. Santos, and E. Uchoa, "A distributed dual ascent algorithm for Steiner problems in multicast routing," *Networks, Wiley Periodicals*, vol. 53, no. 2, pp. 170–183, 2009.

[228] A. Sadeh, "Distributed primal-dual approximation algorithms for network design problems," Master's thesis, Dept. Mathematics and Computer Science, The Open University of Israel, 2008.

[229] L. Vandenberghe, "The proximal mapping," Spring 2011-12. Lecture Notes, Optimization Methods for Large-Scale Systems (EE-236C), UCLA.

[230] E. Birgin, J. Martinez, and M. Raydan, "Nonmonotone spectral projected gradient methods on convex sets," *SIAM J. Optim.*, vol. 10, no. 4, pp. 1196–1211, 2000.

[231] M. Andersen, J. Dahl, and L. Vandenberghe, "CVXOPT." `http://abel.ee.ucla.edu/cvxopt/index.html`, 2012.