KTH Electrical Engineering

Distributed Optimization with Nonconvexities and Limited Communication

SINDRI MAGNÚSSON

Licentiate Thesis
Stockholm, Sweden 2016

KTH Royal Institute of Technology
School of Electrical Engineering
SE-100 44 Stockholm
SWEDEN

TRITA-EE 2016:006
ISSN 1653-5146
ISBN 978-91-7595-854-5

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie licentiatesexamen i electro och systemteknik fredag den 19 februari 2016 klockan 10.00 i Osquarsbacke 14, sal E3, Huvudbyggnaden, KTH Campus.

Tryck: Universitetsservice US AB

Abstract

In economical and sustainable operation of cyber-physical systems, a number of entities need to often cooperate over a communication network to solve optimization problems. A challenging aspect in the design of robust distributed solution algorithms to these optimization problems is that as technology advances and the networks grow larger, the communication bandwidth used to coordinate the solution is limited. Moreover, even though most research has focused distributed convex optimization, in cyberphysical systems nonconvex problems are often encountered, e.g., localization in wireless sensor networks and optimal power flow in smart grids, the solution of which poses major technical difficulties. Motivated by these challenges this thesis investigates distributed optimization with emphasis on limited communication for both convex and nonconvex structured problems. In particular, the thesis consists of four articles as summarized below.

The first two papers investigate the convergence of distributed gradient solution methods for the resource allocation optimization problem, where gradient information is communicated at every iteration, using limited communication. In particular, the first paper investigates how distributed dual descent methods can perform demand-response in power networks by using one-way communication. To achieve the one-way communication, the power supplier first broadcasts a coordination signal to the users and then updates the coordination signal by using physical measurements related to the aggregated power usage. Since the users do not communicate back to the supplier, but instead they only take a measurable action, it is essential that the algorithm remains primal feasible at every iteration to avoid blackouts. The paper demonstrates how such blackouts can be avoided by appropriately choosing the algorithm parameters. Moreover, the convergence rate of the algorithm is investigated. The second paper builds on the work of the first paper and considers more general resource allocation problem with multiple resources. In particular, a general class of quantized gradient methods are studied where the gradient direction is approximated by a finite quantization set. Necessary and sufficient conditions on the quantization set are provided to guarantee the ability of these methods to solve a large class of dual problems. A lower bound on the cardinality of the quantization set is provided, along with specific examples of minimal quantizations. Furthermore, convergence rate results are established that connect the fineness of the quantization and number of iterations needed to reach a predefined solution accuracy. The results provide a bound on the number of bits needed to achieve the desired accuracy of the optimal solution.

The third paper investigates a particular nonconvex resource allocation problem, the Optimal Power Flow (OPF) problem, which is of central importance in the operation of power networks. An efficient novel method to address the general nonconvex OPF problem is investigated, which is based on the Alternating Direction Method of Multipliers (ADMM) combined with sequential convex approximations. The global OPF problem is decomposed into smaller problems associated to each

bus of the network, the solutions of which are coordinated via a light communication protocol. Therefore, the proposed method is highly scalable. The convergence properties of the proposed algorithm are mathematically and numerically substantiated. The fourth paper builds on the third paper and investigates the convergence of distributed algorithms as in the third paper but for more general nonconvex optimization problems. In particular, two distributed solution methods, including ADMM, that combine the fast convergence properties of augmented Lagrangian-based methods with the separability properties of alternating optimization are investigated. The convergence properties of these methods are investigated and sufficient conditions under which the algorithms asymptotically reache the first order necessary conditions for optimality are established. Finally, the results are numerically illustrated on a nonconvex localization problem in wireless sensor networks.

The results of this thesis advocate the promising convergence behaviour of some distributed optimization algorithms on nonconvex problems. Moreover, the results demonstrate the potential of solving convex distributed resource allocation problems using very limited communication bandwidth. Future work will consider how even more general convex and nonconvex problems can be solved using limited communication bandwidth and also study lower bounds on the bandwidth needed to solve general resource allocation optimization problems.

Keywords: Distributed Optimization, Resource Allocation, Power Networks, Limited Communication, Nonconvex Optimization, Wireless Sensor Networks, Cyberphysical Systems.

Acknowledgments

First of all, I would like to thank my supervisor Associate Professor Carlo Fischione for his unlimited support, insightful guidance, and inspiring spirit. I am grateful to Prof. Bo Wahlberg for being my co-advisor. I would also like to express my deep gratitude to Pradeep Chathuranga Weeraddana for patiently guiding me through my first steps of research. During my studies I also had the great pleasure of working with Associate Professor Michael Rabbat, Professor Vahid Tarokh, Assistant Professor Na Li, Chinwendu Enyioha, and Kathryn Heal, thanks for all the fun and productive times we had together, I learned so much from all of you! In particular, I would like express sincere appreciation to Prof. Vahid Tarokh for inviting me to work in his group in Harvard University and making me feel so welcome during my stay there. I would also like to thank the Engblom Foundation for supporting my research at Harvard University. I would also like to think all my collages, at the Department of Automatic Control at the Electrical Engineering school at KTH Royal Institute of Technology, for contributing in making work so much fun. Finally, I would like thank my girlfriend, friends, and family for their unconditional support during my studies.

Sindri Magnússon
Stockholm, January 2016

Contents

List of Figures

List of Acronyms

ADMM	Alternating Direction Method of Multipliers
CDF	Cumulative Distribution Function
OPF	Optimal Power Flow
QGM	Quantized Gradient Methods
RA	Resource Allocation
WSN	Wireless Sensor Network

Part I

Thesis Overview

Chapter 1

Introduction

Cyber physical systems, consisting of networked computing and sensing devices used to control physical systems, appear in numerous modern engineering applications. Examples range from large systems such as smart cities, power grids, and vehicular networks, to smaller systems such as robotic teams, smart homes, and inter-body wireless sensor networks. The different network entities, e.g., computing, sensing, and actuator devices, are usually scattered throughout the physical system that they operate and coordinate through a communication network. The traditional way to operate these systems is to use a centralized control/decision making approach, where all the entities send their data to a central node which makes an executive decision. However, the centralized approach is generally not applicable in practice due to the constraints of real-world engineering systems, e.g., their large scale, the need for real-time operation, or communication limits. Moreover, the centralized approach may violate the autonomy and privacy of the individual entities, which does not comply with most practical applications. Therefore, reliable control/decision making algorithms for cyber physical systems must run in a distributed fashion where the entities cooperate by a communication network.

In addition to the distributed nature, a challenging aspect in the design of robust distributed algorithms is that as technology advances and the networks become larger the communication bandwidth used to coordinate these networks has fundamental limits or can be poorly available due to high costs. In particular, the computing, sensing, and actuator devices are always becoming cheaper, smaller, and more convenient to employ in larger quantities. At the same time, both academia and industry are constantly coming up with new ways to revolutionize networked infrastructures and cyber physical systems using the latest technology. Yet, a bottleneck for future developments is the communication bandwidth, which is a scarce or expensive resource. Therefore, distributed algorithms for cyber physical systems that are agnostic about the communication infrastructure are not sustainable. As a result, it is important to investigate communication efficient algorithms for completing the main tasks of cyber physical systems, such as optimization, decision making, or estimation.

The task of finding good operation points in cyberphysical systems can often be formulated as an optimization problem with a global network-wide objective function and problem data that are scattered between the network entities. Distributed algorithms for solving these optimization problems have been much investigated recently, but almost extensively for well behaved convex problems. On the contrary, distributed algorithms for the more challenging nonconvex problems have received little attention despite the emerging need for scalable solutions for many large scale nonconvex applications, e.g., localization in wireless sensor networks and optimal power flow in smart grids. Fortunately, most large scale nonconvex problems share the structures that are usually exploited to achieve distributed/scalable algorithms for their convex counterparts; however, the convergence properties of the solution algorithms for nonconvex problems are left largely unexplored. In this thesis, we investigate some prominent nonconvex optimization problems that appear in cyberhysical systems, as we survey in the next section that provides some motivating examples.

1.1 Motivating Examples

We now provide two motivating examples for the theory developed in this thesis.

1.1.1 Future Power Distribution Networks

Power networks are arguably among the largest infrastructures made by humans and an essential foundation for most modern technology. Power networks consists of two layers, *transmission networks* and *distribution networks*, which are connected by *electric substations* [2]. The transmission networks deliver power from big power plants to large factories and electrical substations. The distribution networks take over at the electrical substations and deliver the power within the cities to the end users. In the transmission networks, all the network entities are owned by the the power companies and possibly few other big corporations. Therefore, the information and the control actions needed for efficiently operating these networks are easily attainable. In fact, transmission networks are now operated at almost optimal efficiency after being heavily investigated for many decades. On the other hand, the distribution networks consist of large number of users, such as households and small business, whose electricity consumption is unpredictable. Therefore, distribution networks have traditionally been operated using crude information using statistical inference of the aggregated/total power usage and without timely and efficiently affecting the users behaviour.

The rapid developments of modern electronics over the last decades have resulted in unprecedented growth of power demand in the distribution networks. At the same time, both legislators and consumers have largely rejected power generation that comes at the expense of nature. Therefore, the future power networks are confronted with the contradicting goals of increasing the generation and causing less environmental impacts. The distribution networks must be reformed to keep up

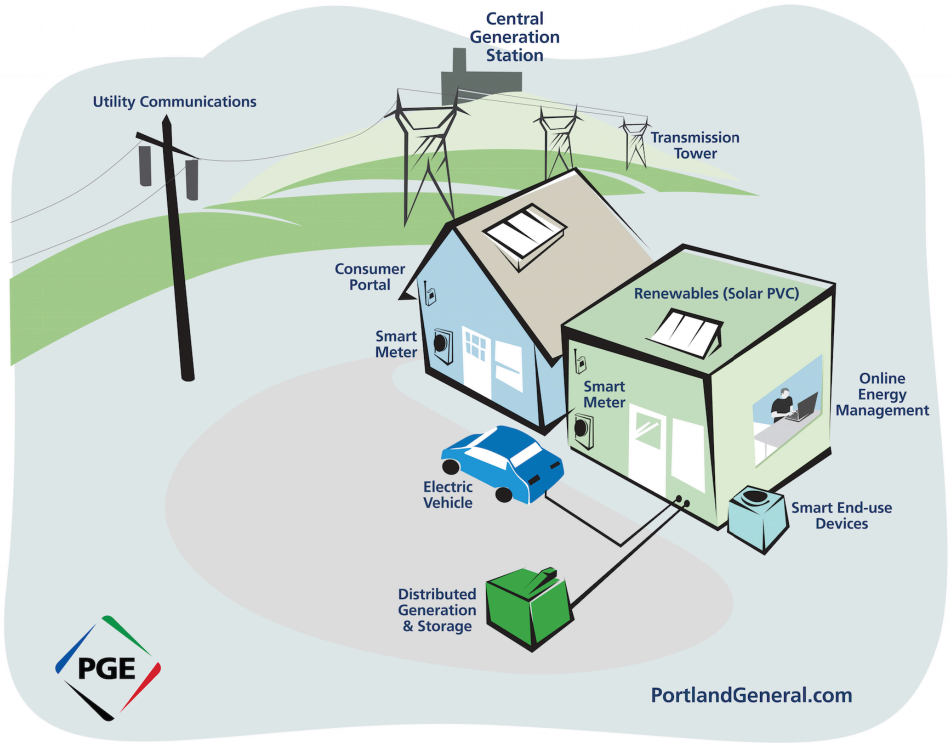


Figure 1.1: The future distribution network will consist of various devices that co-operate to achieve more economic and environmental friendly power distribution. A challenging aspect is how to achieve the best operation using limited communication bandwidth for coordination and communication among the devices. (source: https://www.flickr.com)

with the future developments. Fortunately, the many recent technical advancements such as i) cheaper, smaller, and more effective sensing-, computing-, and actuator devices, ii) the promising evaluation of fast and distributed signal processing and control algorithms, and iii) more reliable communication, have the potential to revolutionize today's distribution networks. In particular, in the future smart distribution networks [3] (smart grids) users will be equipped with smart meters which are small computational devices that can communicate with the power operator to help consumers use power in a more economical way. The smart meter will both i) affect/control the users behaviour positively by incentivising them to consider the state of the distribution network when using power, e.g., lowering the price when the total demand is low, and ii) increase the efficiency of the distribution network by gathering the needed information and coordinating with other network entities.

Moreover, the household appliances will be more autonomous and use power in more economic and environmental friendly manner, in sync with the user. The future distribution network will also generate power from many distributed renewable generators such as wind turbines. The consumers will also be able to inject power into the grid from various renewable sources such as solar panels, training bikes, or from the batteries of their electrical vehicles.

While the vision of the next generation distribution networks has the potential to revolutionize todays power networks, many technical problems must first be solved. With the smart grid now in its infancy, industry and academia will investigate new hardware that will be integrated to the grid and solve advanced tasks with the new technology. With the distribution networks growing, due to the integration of all the new technologies, it is essential to carefully consider from the start how the different tasks can be accomplished with minimal communication so the smart grid will evolve in a sustainable manner and will not crash under the future growing huge scale. Another reason for considering efficient communication is that no sophisticated communication infrastructure has yet been integrated into todays distribution networks - nevertheless power networks are equipped with natural communication infrastructure, i.e., power line communication [4, 5], which can already be used but has limited communication bandwidth. The power line communication has the potential to kickstart many of the early integrations of smart grid before a dedicated communication network is merged with the grid, which could take a long time since industry will carefully consider any steps towards integrating costly new infrastructure into the grid. Therefore, it is essential to address the communication challenges right away both for sake of early integration of smart grid and also so that the smart grid can keep up with future challenges.

Another central problem in power networks is the optimal power flow problem, which finds the best of feasible power flows through the network, where what is best depends on the engineering need. The problem is well studied in transmission networks, but due to lack of control and information in traditional distribution networks the problem has only recently become relevant. A challenge in solving the optimal power flow problem in distribution networks is that DC-optimal power flow, an approximation of the problem which works well in transmission networks, is not appropriate for the distribution network due to their low voltage. Therefore, when solving power flow problems in distribution networks the full AC optimal power flow problem needs to be considered, which turns out to be highly nonconvex. As a result, it is important to find distribution solution method to nonconvex power flow problems in the distribution networks to operate them more efficiently.

1.1.2 Wireless Sensor Networks

Wireless sensor networks (WSN) [6] are one of the fundamental parts of cyber physical systems, where the autonomous devices used to monitor/control the physical system cooperate over a wireless medium. In many cyber physical systems, wireless communication is the only feasible way to coordinate the network devices due to

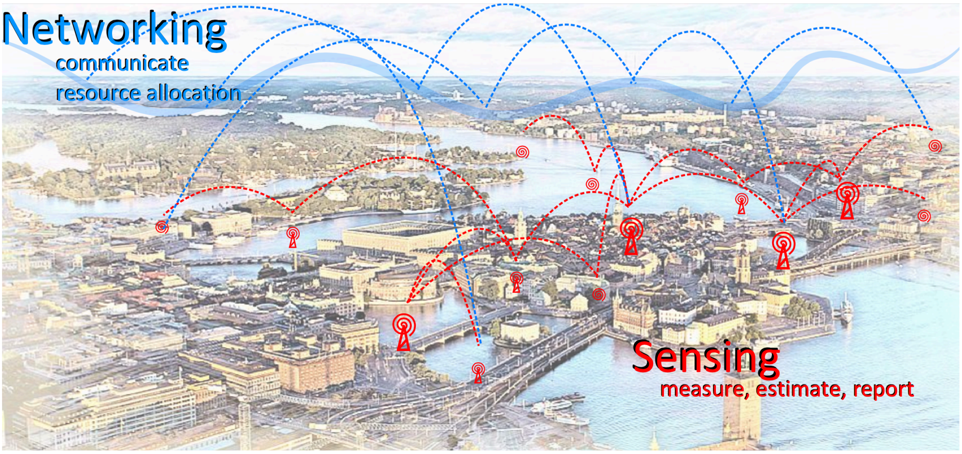


Figure 1.2: Wireless sensor network operating a city. Some of the devices coordinate the power flows through the city, others the vehicular traffic and water distribution. (source: Yuzhe Xu's Licentiate Thesis [1])

environmental constraints, such as when the networks are mobile (e.g., vehicle networks [7]), are underwater [8], or are attached or even inside the human body (e.g., body area networks [9]). Even when wired communications are feasible, wireless communication can still be a better option due to the many benefits of WSNs, e.g., they don't need an existing communication infrastructure and are usually easier to employ. However, to achieve all the benefits of wireless operation of cyber physical systems, lightweight communications are essential because communication over wireless channels is a scarce resource and has to be well coordinated due interferences and messages collisions, or may be very expensive. Moreover, the wireless devices are usually battery powered with no or limited power sources and therefore economical usage of communication is essential to prolong the lifetime of the networks.

Localization and tracking is a fundamental task in many WSN's, since the physical location of the network devices usually has large statistical impact on the sensed information and is also relevant in controlling the network. Moreover, in many WSN applications, such as indoor positioning systems, the localization/tracking is the sensing task of the network. Using the Global Positioning Systems (GPS) is often unfeasible due non-line of sight to a satellite, e.g., in indoor positioning, or unsatisfactory accuracy, e.g., in interbody wireless sensor networks, in addition to being generally unattractive due to the power and cost constraints of the individual sensors. A more attractive option is self-localization of the network [10], where the locations of the devices are estimated from the known locations of reference nodes and distance measurements between communication neighbors in the network. Localization using distance measurements is a nonconvex optimization problem. Therefore, it is essential to find distributed solution method that solve such challenging nonconvex optimization cooperatively among the network devices using efficient communication schemes.

1.2 Problem Formulation

In this thesis we consider general problems of the form

$$\begin{aligned} &\underset{x}{\text{minimize}} && \sum_{i=1}^{N} f_i(\mathbf{x}_i), \\ &\text{subject to} && (\mathbf{x}_1, \cdots, \mathbf{x}_N) \in \mathcal{X}, \qquad \mathbf{x}_k \in \mathcal{X}_k, \end{aligned} \tag{1.1}$$

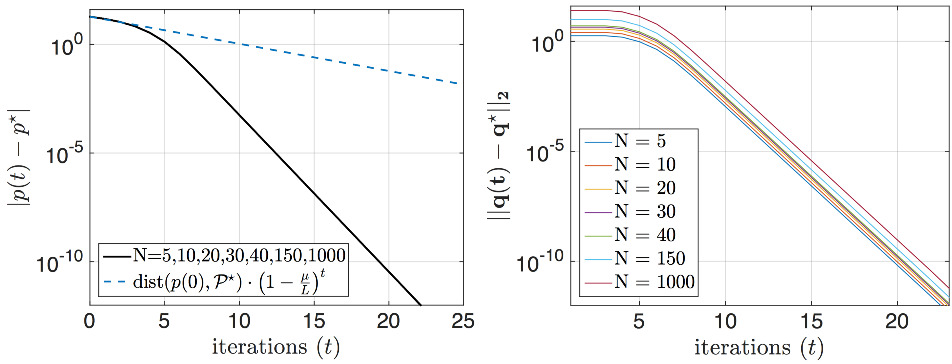
where $i = 1, \cdots N$, are users in a cyberphysical system, or nodes in a network. The objective function is separable between the usurers where the objective function part of user $i = 1, \cdots, N$, f_i, represents his/her preference. The constraints are coupled between the users, which can for example represent shared resources. We articulate in Chapter 2 the theoretical background on problems of the form of (1.1). In this thesis, we consider distributed solution methods for (1.1), where the users keep their private objective function part, f_k, and constraints, $\mathcal{X}_k$, and communicate to reach the solution, over possibly a bandwidth limited channel. Moreover, we consider some nonconvex instances of problem (1.1).

1.3 Outline and Contribution of the Thesis

This thesis is based on four published and submitted papers, we now list the contribution of each of them.

Distributed Resource Allocation Using One-Way Communication with Applications to Power Networks

The first paper investigates the solution to problem (1.1) based on distributed dual descent power allocation algorithm in power networks that only use one way communication. At every iteration of the algorithm the power supplier i) first broadcasts a scalar coordinating/price signal to the users and ii) then he/she measures the aggregate power usage given that pricing signal. Therefore, the algorithm operates by communicating only one scalar message at every iteration of the algorithm. Since the users actually update their power consumption while the algorithm is running, it is possible that the total power usage excites the supplier power capacity causing blackouts. To avoid such blackouts the resource allocation problem must remain feasible at every iteration of the algorithm, i.e., the power usage must always be below the suppliers capacity. We show how the algorithm parameters can be chosen to ensure primal feasibility while the algorithm runs and hence avert costly blackouts. To ensure the primal feasibility at every iteration, it is not possible to choose



(a) Dual Iterates

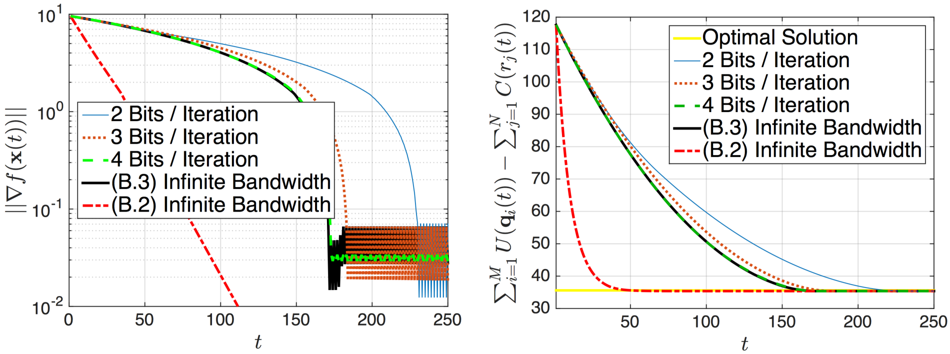
(b) Primal Iterates

Figure 1.3: The figure depicts the convergence of the one-way communication dual decomposition, reported in the first paper, for solving an instance of (1.1) where the number of users is $N = 5, 10, 20, 30, 40, 150, 1000$. Figure 1.3a and 1.3b depict the the distance from the dual and primal variables, at each iteration, to the optimal dual and primal variable, respectively. The blue dotted line in Figure 1.3a depicts a theoretical bound on the convergence provided in the paper. For illustration purposes the primal problems have been constructed so that the dual problem does not change when number of users increases and hence the dual convergence shown in Figure 1.3a is the same for all N. The paper discusses how the primal problem can be regulated so the dual problem does not change significantly when the number of users increase, indicating superior scalability properties.

the step sizes in the dual descent algorithm that give the optimal convergence rate $\mathcal{O}(1/t^2)$, where t is the iteration index, for the given the structure of the dual problem, which is convex with Lipschitz continuous gradients, and the algorithm generally has the convergence rate $\mathcal{O}(1/t)$. Nevertheless, we show that under mild structure on the resource allocation problem, a linear convergence rate $\mathcal{O}(c^t)$, with $c \in [0, 1[$, is achieved. Moreover, we provide additional problem structure where the linear convergence rate is independent of the number of users, hence demonstrating a superior scalability properties. Finally, we illustrate the results using numerical simulations.

This article has been accepted to appear in:

- S. Magnússon, K. Heal, C. Enyioha, N. Li, C. Fischione, and V. Tarokh, Distributed Resource Allocation Using One-Way Communication with Applications to Power Networks, Accepted to Appear in the 50th Annual Conference on Information Sciences and Systems (CISS) 2016, Princeton, NJ, USA.



(a) Gradient

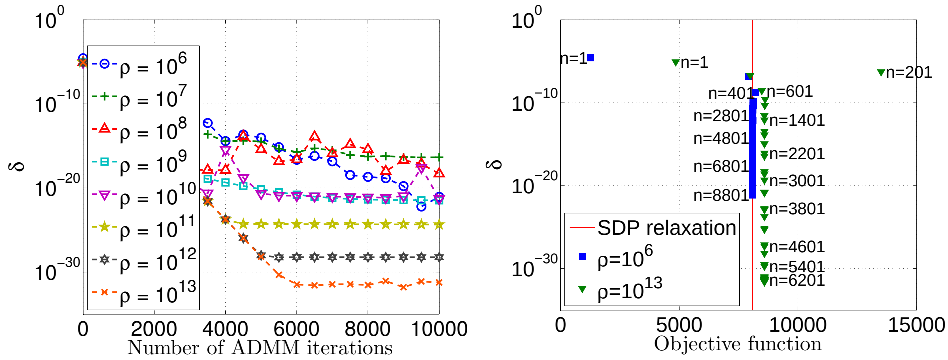
(b) Objective Function Value

Figure 1.4: The figure depicts iterations to obtain the solution to optimization problem (1.1) based on the quantized gradient methods from the second paper reported in the thesis. The vertical axis in Figures 1.4a and 1.4b depicts the norm of the gradient and the primal objective function, respectively, at every iteration. The green dotted line is an approximation of the gradient direction method (B.3) where only 4 bits are communicated per iteration, but yet achieves almost the same performance.

Convergence of Limited Communications Gradient Methods

The second paper investigates distributed gradient methods, where the gradient is communicated at every iteration of the algorithm, when bandwidth is limited. In particular, the paper considers quantized gradient methods (QGM) where the gradient descent direction is projected to a finite quantization set $\mathcal{D}$ before being communicated. The paper investigates necessary and sufficient conditions that ensure the quantization set $\mathcal{D}$ be *proper*, in the sense that the QGMs can minimize any convex function $f : \mathbb{R}^N \to \mathbb{R}$ with Lipschitz continuous gradients and non-empty, bounded set of minimizers. We use this characterization to provide examples of proper quantization sets $\mathcal{D}$. We also show that if $|\mathcal{D}| \leq N$ then $\mathcal{D}$ cannot be proper and there exists an optimization problem, from the aforementioned class, which QGMs can not solve. Moreover, we show that there exists proper quantization sets with $|\mathcal{D}| = N + 1$, hence the minimal cardinality of a proper quantization set is $N + 1$, which can be communicated using $\log_2(N + 1)$ bits. We provide a bound on the number of iterations needed to achieve any ϵ accuracy on the optimal solution that depends on the fineness of the quantization set $\mathcal{D}$. Specifically, the bound on number of iterations decreases when the quantization set becomes finer. We also show that, when the step-sizes are non summable but square summable, then the iterates of QGMs converge to the set of optimal values. Finally, we demonstrate how the theory can be applied to a resource allocation problem in power networks.

This paper has been submitted to:



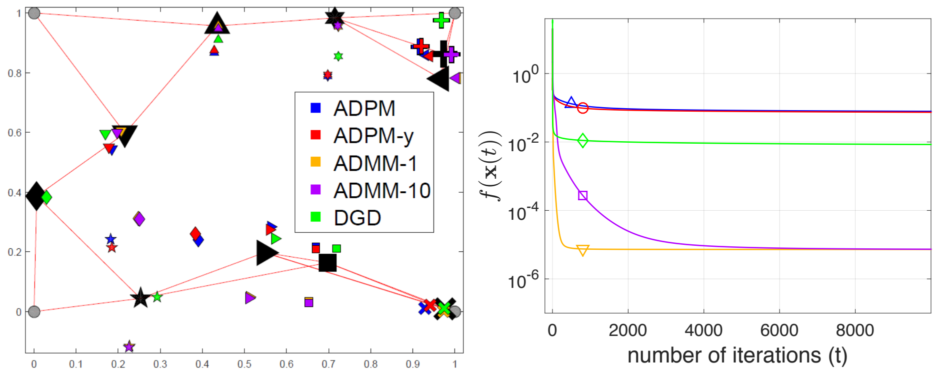
(a) Consensus (b) Objective Function

Figure 1.5: The figure depicts the convergence of ADMM for the nonconvex OPF problem, of the form (1.1), studied in the third paper. The algorithm is distributed between the nodes/buses of a power network where each node keeps a private estimate of its neighbours voltages and δ is a measure of the consensus/consistency between the voltage estimates. ρ is an algorithm parameter that penalizes violations of the consensus constraint. Figure 1.5a depicts δ over the course of the algorithm for different ρ's. The figure shows that the algorithm always reaches consensus among the nodes with high numerical accuracy and larger ρ enforce more consistency. Figure 1.5b depicts the objective function value compared with the consensus, δ, where n indicates the iteration number. The figure shows that when $\rho = 10^6$ then the algorithm converges almost to the red vertical line depicting a lower bound on the optimal value given by a relaxation.

- S. Magnússon, K. Heal, C. Enyioha, N. Li, C. Fischione, and V. Tarokh, Convergence of Limited Communications Gradient Methods, submitted to IEEE American Control Conference (ACC) 2016.

A Distributed Approach for the Optimal Power Flow Problem Based on ADMM and Sequential Convex Approximations

The third paper considers the optimal power flow (OPF) problem, which plays a central role in operating electrical networks. The problem consists in finding the optimal flow of power through a power networks. It is nonconvex and NP hard. Therefore, designing efficient algorithms of practical relevance is crucial, though their global optimality is not guaranteed. The paper develops an efficient novel method to address the general nonconvex OPF problem. The proposed method is based on the alternating direction method of multipliers combined with sequential convex approximations. The global OPF problem is decomposed into smaller problems associated to each bus of the network, the solutions of which are coordinated via a light communication protocol. Therefore, the proposed method is



(a) Location Estimates (b) Objective Function

Figure 1.6: The figure depicts the the convergence of the distributed algorithms studied in the forth reported paper when (1.1) represents a nonconvex estimation problem, i.e., localization based on noise distance measurements. Figure 1.6a depicts the problem setup, where the sensors, black markers, estimate their own location by measuring the distances to their neighbours in the network and communicating over the network. The grey circles are anchor nodes that know their own location and the coloured markers denote the estimations obtained from different algorithms. Figure 1.6b depicts the objective function, i.e., a penalty of violating the distance measurements, at every iteration. Since the problem is nonconvex the algorithms can converge to different local minima.

highly scalable. The convergence properties of the proposed algorithm are mathematically substantiated. Finally, the algorithm is evaluated on a number of test examples, where the convergence properties are investigated and the performance is compared with a global optimal method.

The paper has been published in:

- S. Magnússon, P. C. Weeraddana, C. Fischione, "A Distributed Approach for the Optimal Power-Flow Problem Based on ADMM and Sequential Convex Approximations,", Control of Network Systems, IEEE Transactions on, vol.2, no.3, pp.238-253, Sept. 2015

On the Convergence of Alternating Direction Lagrangian Methods for Nonconvex Structured Optimization Problems

The forth reported paper investigates the convergence of distributed methods for nonconvex structured optimization problems. In particular, the paper investigates two distributed solution methods that combine the fast convergence properties of augmented Lagrangian-based methods with the separability properties of alternating optimization. The first method is adapted from the classic quadratic penalty function method and is called the Alternating Direction Penalty Method (ADPM). Unlike the original quadratic penalty function method, in which single-step optimizations are adopted, ADPM uses an alternating optimization, which in turn makes it scalable. The second method is the well-known Alternating Direction Method of Multipliers (ADMM). It is shown that ADPM for nonconvex problems asymptotically converges to a primal feasible point under mild conditions. Additional conditions ensuring that ADPM asymptotically reaches the standard first order necessary conditions for local optimality are introduced. In the case of the ADMM, novel sufficient conditions under which the algorithm asymptotically reaches the standard first order necessary conditions are established. Based on this, complete convergence of ADMM for a class of low dimensional problems are characterized. Finally, the results are illustrated by applying ADPM and ADMM to a nonconvex localization problem in wireless sensor networks.

The chapter is based on the following papers:

- S. Magnússon, P. C. Chathuranga, M. Rabbat, C. Fischione, "On the Convergence of Alternating Direction Lagrangian Methods for Nonconvex Structured Optimization Problems," Accepter, to Appear, in Control of Network Systems, IEEE Transactions on, 2016

- S. Magnússon, P. C. Chathuranga, M. Rabbat, C. Fischione, "On the convergence of an alternating direction penalty method for nonconvex problems," in Signals, Systems and Computers, 2014 48th Asilomar Conference on, pp.793-797, 2-5 Nov. 2014

Chapter 2

Background

This chapter summarizes the background theory used in the contribution of the thesis. In particular, Section 2.2 introduces general resource allocation problems that are studied in this thesis. Section 2.3 introduces the optimal power flow problem, a specific resource allocation problem that plays an important role in the operation of power networks. In the following Sections 2.4 and 2.5 we discuss distributed solution methods for the studied problems. In particular, Section 2.4 discusses dual decomposition and 2.5 discusses the alternating direction method of multipliers (ADMM).

2.1 Notation

We use the following notation in this chapter. Vectors and matrices are represented by boldface lower and upper case letters, respectively. The set of real n vectors and $n \times m$ matrices are denoted by $\mathbb{R}^n$ and $\mathbb{R}^{n \times m}$, respectively, and $\mathbb{C}$ represents the set of complex numbers. Otherwise, we use calligraphy letters to represent sets. The superscript $(\cdot)^{\mathrm{T}}$ stands for transpose. j denotes the imaginary number $\sqrt{-1}$. $\mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ denotes the diagonal block matrix with $\mathbf{A}_1, \ldots, \mathbf{A}_n$ on the diagonal. $||\cdot||$ denotes the 2-norm. The $\mathtt{dom}f$ is the domain of the function $f : \mathbb{R}^n \to \mathbb{R}^m$. $\mathtt{int}\mathcal{X}$ denotes the interior of the set $\mathcal{X}$.

2.2 Resource Allocation

The question of how to distribute limited resources between number of entities is central to every society, even so that it is the theme of a whole field of social science, i.e., economics. The question also has deep roots in engineering since many engineering systems are designed to distribute some limited resources between users, e.g., power, data/bandwidth, water, etc. To design engineering systems that can distribute the limited resources in some sensible manner the question must be expressed formally as a mathematical problem, so it is clear which resource allocations are more desirable than others. A standard approach is to formulate the resource allocation problem as an optimization problem by associating a utility function f_i to each user $i \in \mathcal{N}$, where $\mathcal{N} = \{1 \cdots, N\}$ is a set of N users. Then the resource allocation problem can be formulated as follows.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \sum_{k=1}^{N} f_k(\mathbf{x}_k) \\ \text{subject to} \quad & \mathbf{g}(\mathbf{x}_1, \cdots, \mathbf{x}_N) \leq 0, \quad (\text{or} \quad \mathbf{g}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = 0) \\ & \mathbf{x}_k \in \mathcal{X}_k, \quad \text{for all } k = 1, \cdots, N, \end{aligned} \tag{2.1}$$

where the function $\mathbf{g}$ is associated to the resources that are shared among the users and $\mathcal{X}_k$ is a local constraint associated with user $k \in \mathcal{N}$ which expresses his/her preferences. We refer to the constraint $\mathbf{g}(\mathbf{x}_1, \cdots, \mathbf{x}_N) \leq 0$ (or $\mathbf{g}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = 0$) as the coupling constraints since they couple the allocations between the users. The coupling constraint can be an equality or an inequality constraint depending on weather the resources must be used up or not. The utility functions f_k can represent the preferences of user $k \in \mathcal{N}$ for the allocated resources or can be designed to achieve various types of equilibrium points or fairness among the users [11, 12].

An illustrative and practical instance of (2.1) is when one resource is shared among all the users as follows:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \sum_{k=1}^{N} f_k(x_k) \\ \text{subject to} \quad & \sum_{k=1}^{N} x_k \leq Q, \quad \left(\text{or} \quad \sum_{k=1}^{N} x_k = Q\right) \\ & x_k \in \mathcal{X}_k, \quad \text{for all } k = 1, \cdots, N, \end{aligned} \tag{2.2}$$

Resource allocation problem (2.2) can, for an example, appear in power networks where limited power is available and needs to be shared among the users. Here, Q denotes the power available in the network, x_k the power allocated to user $k \in \mathcal{N}$, and $\sum_{k=1}^{N} x_k$ the aggregated power usages. The set $\mathcal{X}_k$ and the function f_k represent the preferences of user $k \in \mathcal{N}$. More complicated variants of (2.1) can have more resources or even coupling between the different resources.

In power networks, it is usually not only important to find a good/fair allocation between the users of the available power, but it is also important that the network can support that allocation. This is because the power in power networks flows according to physical laws, i.e., Ohm's and Kirchoff's laws, represented by power flow equations. We next demonstrate a special case of (2.1) that includes these physical properties of power networks.

2.3 Optimal Power Flow Problem

Consider a power network $(\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ and $\mathcal{E}$ are the users/vertices and power lines/edges, respectively. We represent the power flow through each user $k \in \mathcal{N}$ by the power flow equation

$$g_k(\mathbf{z}_{\mathcal{N}_k}) = 0,$$

where $\mathbf{z}_{\mathcal{N}_k} = (\mathbf{z}_i)_{i \in \mathcal{N}_k}$, $\mathcal{N}_k = \{j | (k, j) \in \mathcal{E}\} \cup \{k\}$, and the vector $\mathbf{z}_i$ keeps all the needed physical quantities of user i, such as voltage and power injection. Now the optimal power flow problem consists in finding the optimal flows that satisfies the power flow equations. More formally, the problem can be written as

$$\begin{array}{ll} \underset{\mathbf{z}_1, \cdots, \mathbf{z}_N}{\text{minimize}} & \sum_{k=1}^{N} f_k(\mathbf{z}_k) \\ \text{subject to} & g_k\left(\mathbf{z}_{\mathcal{N}_k}\right) = 0, \quad \text{for all } k \in \mathcal{N}, \\ & \mathbf{z}_k \in \mathcal{X}_k, \quad k \in \mathcal{N}. \end{array} \tag{2.3}$$

Notice that (2.3) is a special case of (2.1) with

$$\mathbf{g}(\mathbf{z}_1, \cdots, \mathbf{z}_N) = (g_k(\mathbf{z}_{\mathcal{N}_k}))_{k \in \mathcal{N}}.$$

The utility functions f_k and the local constraints $\mathcal{X}_k$ can vary depending on the engineering need. For example, the utilities can be chosen so that (2.3) minimizes the power loss or generation costs to satisfy demand of the users given by $\mathcal{X}_k$ or the utilities can be constructed so (2.3) finds some fair allocation of power between the households.

We now formally express the power flow equations $\mathbf{g}(\mathbf{z}_1, \cdots, \mathbf{z}_N) = \mathbf{0}$, see any textbook on power system analysis for more details, e.g., [2, 13]. To express the power flows set $\mathbf{z}_k = (s_k, v_k)$ where $s_k, v_k \in \mathbb{C}$ are the power and voltage at user (or bus) $k \in \mathcal{N}$, respectively, then

$$g_k(\mathbf{z}_{\mathcal{N}_k}) = s_k - v_k \sum_{i \in \mathcal{N}_k} y_{ki} v_i, \tag{2.4}$$

where $y_{ki} = g_{ki} + j b_{ki} \in \mathbb{C}$, with $g_{ki}, b_{ki} \in \mathbb{R}$, is the admittance in the flow line $(k, i) \in \mathcal{E}$. The power flow equations are also commonly expressed with the voltage in polar coordinates. In that case, we set $\mathbf{z}_k = (p_k, q_k, |v_k|, \theta_k)$ where $s_k = p_k + j q_k$, with $p_k, q_k \in \mathbb{R}$, $v_k = |v_k| e^{j\theta_k}$, and the power flow equations (2.4) reduce to

$$g_k(\mathbf{z}_{\mathcal{N}_k}) = \left[\begin{array}{c} p_k - \sum_{i \in \mathcal{N}_k} |v_i||v_k| \left(g_{ki} \cos(\theta_k - \theta_i) + b_{ki} \sin(\theta_k - \theta_i)\right) \\ q_k - \sum_{i \in \mathcal{N}_k} |v_i||v_k| ((g_{ki} \sin(\theta_k - \theta_i) - b_{ki} \cos(\theta_k - \theta_i)) \end{array} \right]. \tag{2.5}$$

The power flow equations (2.4), (2.5), and (2.3) are nonlinear which renders the optimization problem (2.3) nonconvex.

As argued in the introduction, it is essential to solve problems such as (2.1), (2.2), and (2.3) in a distributed manner. We next review standard distributed methods for these problems.

2.4 Dual Decomposition

We start by recalling the demonstration on how problems on the form of (2.1) can be decomposed between the users in $\mathcal{N}$ by using duality theory, for more comprehensive overview see, for example, [14, Chapter 6]. The dual function D of the problem (2.1) with respect to the coupling constraint is obtain by maximizing the Lagrangian function, given by

$$L(\mathbf{x}, \mathbf{p}) = \sum_{k=1}^{N} f_k(\mathbf{x}_k) - \mathbf{p}^{\mathrm{T}} \mathbf{g}(\mathbf{x}_1, \cdots, \mathbf{x}_N), \tag{2.6}$$

which results in

$$D(\mathbf{p}) = \underset{\mathbf{x} \in \mathcal{X}}{\text{maximize}} \; L(\mathbf{x}, \mathbf{p}), \tag{2.7}$$

$$= L(\mathbf{x}(\mathbf{p}), \mathbf{p}), \tag{2.8}$$

where $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, $\mathcal{X} = \prod_{i=1}^{N} \mathcal{X}_i$, and

$$\mathbf{x}(\mathbf{p}) = \underset{\mathbf{x} \in \mathcal{X}}{\text{argmax}} \;\; L(\mathbf{x}, \mathbf{p}). \tag{2.9}$$

The dual problem of (2.1) with respect to the coupling constraint $\mathbf{g}$ is then given by

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & D(\mathbf{p}) \\ \text{subject to} & \mathbf{p} \geq \mathbf{0}. \end{array} \tag{2.10}$$

The dual function, D, is always convex and therefore the dual problem (2.10) is always convex, even when the primal problem (2.1) is nonconvex. Therefore, D has a gradient (or a subgradient) at every interior point of the domain of D [15, Theorem 1.7], i.e., at every $\mathbf{p} \in \texttt{int dom } D$, and the (sub)gradients are given by

$$\nabla D(\mathbf{p}) = \mathbf{g}(\mathbf{x}_1(\mathbf{p}), \cdots, \mathbf{x}_N(\mathbf{p})). \tag{2.11}$$

If optimization problem (2.9) has a unique solution, then (2.11) is a gradient, otherwise every solution $\mathbf{p}$ of (2.9) provides a subgradient (2.11). When D is everywhere differentiable, which is for example the case when f_i are strictly concave for all $i \in \mathcal{N}$ and $\mathbf{g}$ is convex, then (2.10) can be solved by using the dual descent method

$$\mathbf{p}(t{+}1) = \mathbf{p}(t) + \gamma(t) \nabla D(\mathbf{p}(t)) \tag{2.12}$$

with appropriate step-size choice $\gamma(t) \in \mathbb{R}_+$, see [14]. When D is not differentiable but subgradients exist everywhere, which is for example the case if $\mathcal{X}$ is compact and f_i and $\mathbf{g}$ are continuous, then (2.10) can be solved using subgradient method which follows the recursion (2.12) using subgradients.

The interesting aspect about the dual descent method (2.12) is that a decomposition structure in the coupling constraint $\mathbf{g}$ can be used to decouple the recursion (2.12) between the users. For example if $\mathbf{g}$ is separable between the users, i.e.,

$$\mathbf{g}(\mathbf{x}) = \sum_{k=1}^{N} \mathbf{g}_k(\mathbf{x}_k),$$

then optimization problem in (2.7) is fully separable between the users and the (sub)gradient $\nabla D(\mathbf{p})$ can be computed without coordination where each user $k \in \mathcal{N}$ solves the local problem

$$\mathbf{x}_k(\mathbf{p}) = \underset{\mathbf{x}_k \in \mathcal{X}_k}{\operatorname{argmax}} f_i(\mathbf{x}_k) - \mathbf{p}^{\mathrm{T}}\mathbf{g}_k(\mathbf{x}_k).$$

However, as we are interested in solving the primal problem (2.1) but not the dual problem (2.10), the dual descent method is only usable if the primal solution can be constructed from the optimal dual solution $\mathbf{p}^\star$, e.g., if the duality gap is zero and (2.7) has a unique solution for $\mathbf{p}^\star$.

The dual descent method can be unstable in some cases since the dual gradient might not exist everywhere, e.g., outside of the domain of D where $D(\mathbf{p}) = \infty$. Moreover, to ensure that the primal optimal solution can be constructed from the dual optimal solution, strong assumptions must be made on the primal problem, such as strongly convexity. In addition, the convergence of the dual ascent is heavily dependent on the step size choice $\gamma(t)$. These drawbacks of the dual decomposition have motivated the more robust variant of the dual descent method which is now introduced.

2.5 Alternating Direction Method of Multipliers

Consider now a different setup where two user wish to cooperatively solve an optimization problem, we see later that this setup naturally generelizes to a multi user case, e.g., to (2.1). The two users wish to solve a problem of the form

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{subject to} & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}. \end{array} \tag{2.13}$$

where the information keept by the two users is ($\mathbf{x}$, f, $\mathbf{A}$) and ($\mathbf{z}$, g, $\mathbf{B}$), respectively, and both users know $\mathbf{c}$. A method for addressing problems of the from (2.13) cooperatively between the two users is the Alternating Direction Method of Multipliers (ADMM), a variant of the dual descent method where the dual function is obtained from a regularized Lagrangian function given by

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{p}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{p}^{\mathrm{T}}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \rho||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2. \tag{2.14}$$

Unlike the standard Lagrangian, the regularized Lagrangian is not separable in $\mathbf{x}$ and $\mathbf{z}$, even though the primal problem (2.13) has the useful decomposition structure in $\mathbf{x}$ and $\mathbf{z}$. Therefore, the usual primal variable update (cf. (2.7)) cannot be performed in parallel by the two users without communication. Instead, ADMM approximately solves the primal problem in two steps, first with respect to the $\mathbf{x}$ variable and then with respect to the $\mathbf{z}$ variable before updating the dual variable, as follows

$$\mathbf{x}(t+1) = \underset{\mathbf{x}}{\operatorname{argmin}}\ L_\rho(\mathbf{x}, \mathbf{z}(t), \mathbf{p}(t)), \tag{2.15}$$

$$\mathbf{z}(t+1) = \underset{\mathbf{z}}{\operatorname{argmin}}\ L_\rho(\mathbf{x}(t+1), \mathbf{z}, \mathbf{p}(t)), \tag{2.16}$$

$$\mathbf{p}(t+1) = \mathbf{p}(t) + \rho\left(\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t+1) - \mathbf{c}\right). \tag{2.17}$$

Note that the users need to communicate $\mathbf{A}\mathbf{x}(t+1)$ and $\mathbf{B}\mathbf{z}(t+1)$ over the course of the algorithm and the dual variable $\mathbf{p}$ can be maintained by either or both users. Compared to the dual decomposition, the ADMM has very good convergence properties and is guaranteed to converge to the optimal value of (2.13) for any $\rho > 0$ if f and g are closed, proper, and convex and L_0 (the standard Lagrangian) has a saddle point [16].

Problem (2.13) can also cover multiuser scenarios. For instance, by letting the $\mathbf{x}$ variable be a private variable of the different users and $\mathbf{z}$ be a coupling variable that ensures consensus. Consider for example the optimal power flow problem given in (2.3). Let $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, where $\mathbf{x}_k$ is a local copy that user k keeps of his/her own variable $\mathbf{z}_k$ and it's neighbours $\mathbf{z}_i$ for $i \in \mathcal{N}_k$, where $\mathcal{N}_k = \{i \in \mathcal{N} | (k,i) \in \mathcal{E}\}$. For convenience we use the notation $\mathbf{x}_k^k$ to denote the component of $\mathbf{x}_k$ associated with $\mathbf{z}_k$. Now the problem (2.3) can be formulated equivalently on the form of (2.13), as follows

$$\begin{array}{ll}
\underset{\mathbf{x}_1,\cdots,\mathbf{x}_N}{\text{minimize}} & \displaystyle\sum_{k=1}^{N} f_k(\mathbf{x}_k^k) \\
\text{subject to} & \mathbf{x} - \mathbf{E}\mathbf{z} = \mathbf{0} \\
& g_k(\mathbf{x}_k) = 0, \quad \text{for all } k \in \mathcal{N}, \\
& \mathbf{x}_k \in \mathcal{X}_k, \quad k \in \mathcal{N},
\end{array} \tag{2.18}$$

where $\mathbf{E} = (\mathbf{E}_1, \cdots, \mathbf{E}_N)$ is a binary matrix where component $\mathbf{E}_k$ captures the coupling constraints $\mathbf{x}_k = \mathbf{z}_{\mathcal{N}_k} = \mathbf{E}_k\mathbf{z}$. Notice that (2.18) is on the form of (2.13) where the coupling constraint is $\mathbf{x} - \mathbf{E}\mathbf{z} = \mathbf{0}$. Moreover, the $\mathbf{x}$-update of the ADMM (cf. (2.15)) can be performed in a fully distributed manner between the users $\mathcal{N}$ where each user $k \in \mathcal{N}$ solves the local subproblem

$$\begin{array}{ll}
\underset{\mathbf{x}_k}{\text{minimize}} & f_k(\mathbf{x}_k^k) + \mathbf{p}^{\mathrm{T}}(t)\left(\mathbf{x}_k - \mathbf{E}\mathbf{z}(t)\right) + \dfrac{\rho}{2}||\mathbf{x}_k - \mathbf{E}\mathbf{z}(t)|| \\
\text{subject to} & g_k(\mathbf{x}_k) = 0, \\
& \mathbf{x}_k \in \mathcal{X}_k.
\end{array} \tag{2.19}$$

The $\mathbf{z}$-update of the ADMM (cf. (2.16)) reduces to a simple averaging between neighbors, i.e.,

$$\mathbf{z}_k = \frac{1}{|\mathcal{N}_k|} \sum_{i \in \mathcal{N}_k} \mathbf{x}_i,$$

and can be achieved by a simple coordination between the neighbors in the network. The dual variable update (cf. (2.17)) can then be achieved locally by each user.