Large-Scale Complex Systems: From Antenna Circuits to Power Grids

Thesis by

Javad Lavaei

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2011

(Defended May 11, 2011)

To my wife, Somayeh Sojoudi

Acknowledgements

When I was a master's student in Canada, I always wished to closely collaborate with John Doyle and Richard Murray, who were two of the world-famous professors in the area of control. My dream came true when I met John Doyle in the 2006 IEEE Conference on Decision and Control. Indeed, he invited me to give a visit to Caltech in order to discuss my research interests with him and Richard Murray, which led to my admission to the interdisciplinary department of Control & Dynamical Systems. It has been a great honor for me to have John Doyle as my primary PhD advisor and Richard Murray as my PhD co-advisor.

I owe my deepest gratitude to John Doyle and Richard Murray for their guidance, support and encouragement. They taught me how to conduct high-impact interdisciplinary research and inspired me to work on a broad range of projects. I was always amazed by their invaluable insights, ideas and suggestions. I am forever indebted to my advisors for shaping my academic life. I also feel fortunate to have had the opportunity to closely collaborate with Steven Low during my PhD studies. I would like to thank him for motivating me to work on the two important areas of energy systems and communication networks. He was a constant source of inspirational ideas and discussions. Half of this dissertation has been developed under the great supervision of Steven Low.

I would like to thank my old friend, Aydin Babakhani, and his former advisor, Ali Hajimiri, for introducing me to the interesting world of electrical circuits and antenna devices. Part II of this dissertation is the result of my collaboration with Aydin and Ali. I am also grateful to Sean Meyn, Anders Rantzer, Jerrold Marsden, Daniel Kirschen, Chris Dent, Ross Baldick, Mani Chandy and Adam Wierman for fruitful discussions on different parts of this dissertation. Last, but not least, I want to thank my wonderful wife and colleague, Somayeh Sojoudi, for her unconditional love, support and understanding. Our collaboration on several joint projects since 6 years ago has made my academic life special!

Abstract

This dissertation is motivated by the lack of scalable methods for the analysis and synthesis of different large-scale complex systems appearing in electrical and computer engineering. The systems of interest in this work are power networks, analog circuits, antenna systems, communication networks and distributed control systems. By combining theories from control and optimization, the high-level objective is to develop new design tools and algorithms that explicitly exploit the physical properties of these practical systems (e.g., passivity of electrical elements or sparsity of network topology). To this end, the aforementioned systems are categorized intro three classes of systems, and then studied in Parts I, II, and III of this dissertation, as explained below:

Power networks: In Part I of this work, the operation planning of power networks using efficient algorithms is studied. The primary focus is on the optimal power flow (OPF) problem, which has been studied by the operations research and power communities in the past 50 years with little success. In this part, it is shown that there exists an efficient method to solve a practical OPF problem along with many other energy-related optimization problems such as dynamic OPF or security-constrained OPF. The main reason for the successful convexification of these optimization problems is also identified to be the physical properties of a power circuit, especially the passivity of transmission lines.

Circuits and Systems: Motivated by different applications in power networks, electromagnetics and optics, Part II of this work studies the fundamental limits associated with the synthesis of a particular type of linear circuit. It is shown that the optimal design of the parameters of this type of circuit can be performed in polynomial time if the circuit is passive and there are sufficient number of controllable (unknown) parameters. This result introduces a trade-off between the design simplicity and the implementation complexity for an important class of linear circuits. As an application of this methodology, the design of smart antennas is also studied; the goal is to devise an intelligent wireless communication device in order to avoid co-channel interference, power consumption in undesired directions and security issues. Since the existing smart antennas are either hard to program or hard to implement, a new type of smart antenna is synthesized by utilizing tools from algebraic geometry, control, communications, and circuits, which is both easy to program and easy to implement.

Distributed Computation: The first problem tackled in Part III of this work is a very simple type of distributed computation, referred to as *quantized consensus*, which aims to compute the average of a set of numbers using a distributed algorithm subject to a quantization error. It is shown that quantized consensus is reached by means of a recently proposed gossip algorithm, and the convergence time of the algorithm is also derived. The second problem studied in Part III is a more advanced type of distributed computation, which is the distributed resource allocation problem for the Internet. The existing distributed resource allocation algorithms aim to maximize the utility of the network only at the equilibrium point and ignore the transient behavior of the network. To address this issue, it is shown that optimal control theory provides powerful tools for designing distributed resource allocation algorithms with a guaranteed real-time performance.

The results of this work can all be integrated to address real-world interdisciplinary problems, such as the design of the next generation of the electrical power grid, named the Smart Grid.

Contents

Chapter 1

Introduction

Large-scale complex systems naturally appear in many areas of electrical and computer engineering, such as circuits, power networks, communication networks, electromagnetics and distributed computation. Although these systems are different in nature, they benefit from some common features, namely sparsity in the topology or physical properties imposed by the laws of physics. The main purpose of this dissertation is to investigate whether any universal property of these practical large-scale complex systems makes it possible to study these systems using scalable methods in polynomial time. In other words, the high-level objective is to investigate how the physics of a real-world system can be used to simplify its analysis or synthesis. Another objective is to develop new tools and algorithms for the study of such systems based on combining theories from control and optimization. Motivated by the current challenges of this century, the systems of interest in this work are broadly categorized into three classes: power networks, circuits and systems, and distributed computation.

This dissertation is composed of three parts, where each part studies one aforementioned class of large-scale complex systems. In the following Sections 1.1, 1.2, and 1.3, we will first introduce the systems studied in each part of this work and then outline our contribution. After introducing these important categories of systems, we will discuss in Section 1.4 how the results of different parts of this work can be integrated to address an important interdisciplinary problem.

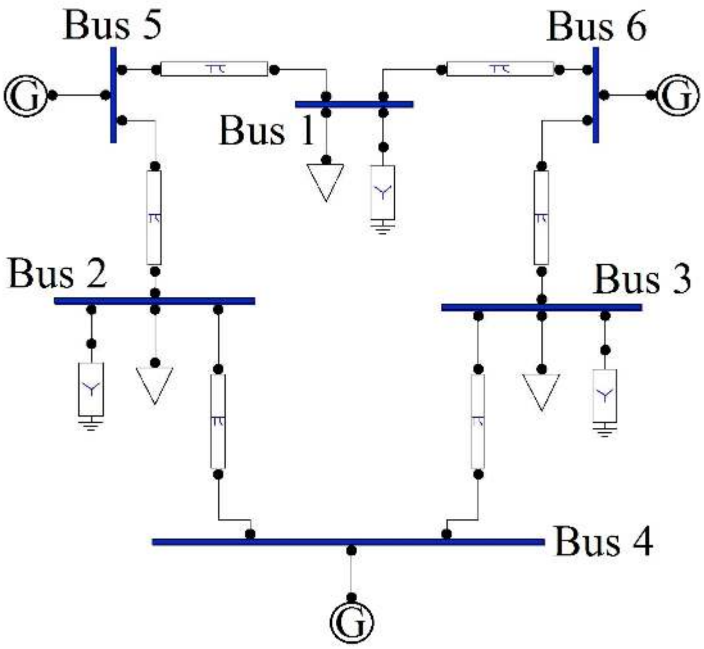


Figure 1.1: An example of a power network

1.1 Part I: Power Networks

A power grid is an interconnected network for delivering electricity from suppliers to consumers. There are a number of optimization problems related to power grids, e.g., network flow, unit commitment, and economic dispatch, which are solved periodically in practice on different time scales (from a few minutes to several days) to set the decision parameters for operation planning and pricing. These resource allocation and planning problems seem to be very hard to solve for power systems due to the nonlinearity of the physical laws imposed by electrical devices.

Part I of this dissertation aims to solve important optimization problems associated with power networks using scalable global optimization techniques. This is explained in more details in the sequel.

1.1.1 Optimal Power Flow Problem

Consider a power network consisting of a set of buses that are connected to each other via transmission lines, transformers or other power electronic devices. Assume that some buses, referred to as *load buses*, are connected to specific loads with given values, whereas the remaining buses, known as *generator buses*, are connected to generators. Figure 1.1 illustrates a 6-bus power network with the load buses 1–3 and the generator buses 4–6. It is often the case that there exist an infinite number of possibilities to generate power by the generators in order to supply the given loads. Hence, a question arises as to how much power should be generated by each generator. In a broader context, it is needed to find an optimal operating point of a power network. This problem is referred to as the optimal power flow (OPF) problem.

The OPF problem is a fundamental optimization problem in the power area, which aims to optimize the decision variables for a power network (e.g., voltage magnitudes at generator buses, values of capacitor banks, and transformer tap ratios) to satisfy the demand and physical constraints [72]. This optimization problem is often solved every 5–15 minutes in practice to decide how to operate the network and charge the consumers accordingly. Started by the work [21] in 1962, the OPF problem has been extensively studied in the literature and numerous algorithms have been proposed for solving this highly nonconvex problem [40, 106, 112], including linear programming, Newton Raphson, quadratic programming, nonlinear programming, Lagrange relaxation, interior point methods, artificial intelligence, artificial neural network, fuzzy logic, genetic algorithm, evolutionary programming and particle swarm optimization [72, 73, 74, 83]. A good number of these methods are based on the Karush-Kuhn-Tucker (KKT) necessary conditions, which can guarantee only the local optimality of the solution, in light of the nonconvexity of the OPF problem [113]. The existing algorithms are not robust, lack performance guarantees, and may not find a global optimum.

During the past few years, the idea of upgrading today's transmission grid into a Smart Grid has been seriously considered by the electric power industry, state and federal regulators, government agencies, and academics [25]. At the high level, the goal is to modernize the electricity transmission and distribution systems to maintain a reliable and secure electricity infrastructure that can meet future demand growth. The aforementioned issues for the OPF problem become more critical in a smart grid because of two reasons: (i) a tremendous increase in the size of the corresponding OPF problem, and (ii) the necessity to solve the OPF problem on a shorter time scale to respond to the intermittency of renewable resources.

Chapter 2 of this dissertation deals with the classical OPF problem. Although an arbitrary OPF problem might not be solvable in polynomial time unless P=NP, the objective is to show that a physically structured OPF problem (corresponding to a practical power network) can be solvable in polynomial time using convex optimization. To this end, a semidefinite programming (SDP) optimization is proposed, which is the dual of an equivalent form of the OPF problem. A global optimum solution to the OPF problem can be retrieved from a solution of this convex dual problem whenever the duality gap is zero. A necessary and sufficient condition is derived to guarantee the existence of no duality gap for the OPF problem. This condition is satisfied by the standard IEEE benchmark systems with 14, 30, 57, 118, and 300 buses as well as several randomly generated systems. Since this condition is hard to study, a sufficient zero-duality-gap condition is also proposed that holds widely in practice, leading to finding a global solution to the OPF problem in polynomial time. The successful convexification of the OPF problem can be traced back to the physical properties of transformers and transmission lines (e.g., passivity).

1.1.2 Energy-Related Optimization Problems

There are several important optimization problems arising in power systems, which are obtained from a single or a set of classical OPF problems by adding more constraints and/or variables. As an example, the security-constrained optimal power flow (SCOPF) problem is one of such optimizations. This problem aims to optimize the performance of a power system under the normal condition such that the load and physical constraints are still satisfied after every pre-specified contingency [19]. Although energy-related optimization problems based on OPF are NP-hard in general, a question arises as whether the physical properties of a power network induce any useful structure on the corresponding optimization problems so that they can be solved (globally) in polynomial time.

Chapter 3 of this dissertation addresses the above question and studies several OPF-based optimization problems appearing in power networks, which are all nonconvex due to the nonlinearity of certain physical quantities (e.g., active power, reactive power, and magnitude of voltage). It is shown that zero duality gap for the classical OPF problem implies zero duality gap for several other optimization problems in power systems such as OPF with variable shunt elements and transformer ratios, dynamic OPF, SCOPF, and scheduling of renewable resources [117]. These findings have the potential to change the way the fundamental optimization problems are solved for power grids in real time.

1.2 Part II: Circuits and Systems

There are many problems in circuits, electromagnetics, optics, and power networks that can be reduced to the analysis and synthesis of some linear systems in the frequency domain. These systems, in the circuit theory, consist of passive elements including resistors, inductors, capacitors, ideal transformers, and ideal gyrators [76]. Since the seminal work [15], there has been remarkable progress in characterizing such passive (dissipative) systems using the concept of positive real functions. As witnessed in Part I of this dissertation, passivity is also the key reason behind the tractability of a practical OPF problem.

The emerging optimization tools developed by control theorists, such as linear matrix inequalities (LMIs) [13] and sum-of-squares (SOS) [84], have been successfully applied to a number of fundamental problems in circuits. For instance, the work [110] is one of the earliest papers connecting the convex optimization theory to circuit design, whose objective is to optimize the dominant time constant of a linear resistor-capacitor circuit using semidefinite programming. The recent paper [36] proposes an LMI optimization to check whether a given multi-port network can be realized using a pre-specified set of linear time-invariant components (namely an inductor and small-signal model of a transistor).

Part II of this dissertation studies how a circuit can be synthesized efficiently using advanced optimization techniques, e.g., LMI, when the circuit is known to be passive. To this end, a general circuit synthesis problem is first tackled and the results obtained are then applied to an antenna design problem. In what follows, we will introduce these problems and subsequently summarize our contributions.

1.2.1 Analog Circuits

Consider the simple filter drawn in Figure 1.2. Assume that the goal is to find the numerical values of the impedances Z_1 to Z_5 in such a way that the input-output gain of the filter is maximized at a pre-specified frequency ω_0. To this end, one can re-organize the elements of this filter to obtain the equivalent model given in Figure 1.3, where the known elements are clustered in the block "linear passive network" and the unknown components are grouped in the block "control unit". Under this setting, the objective reduces to designing the control unit in Figure 1.3 so that the magnitude of the observed output of the circuit is maximized.

Motivated by the above example, consider the circuit given in Figure 1.4 consisting of

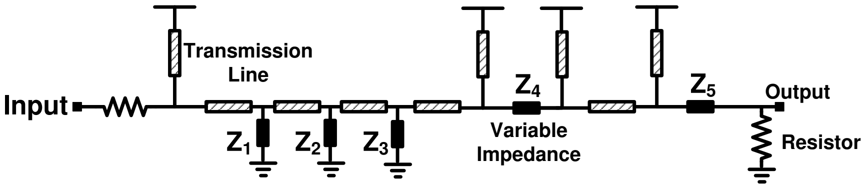


Figure 1.2: A distributed circuit with variable impedances

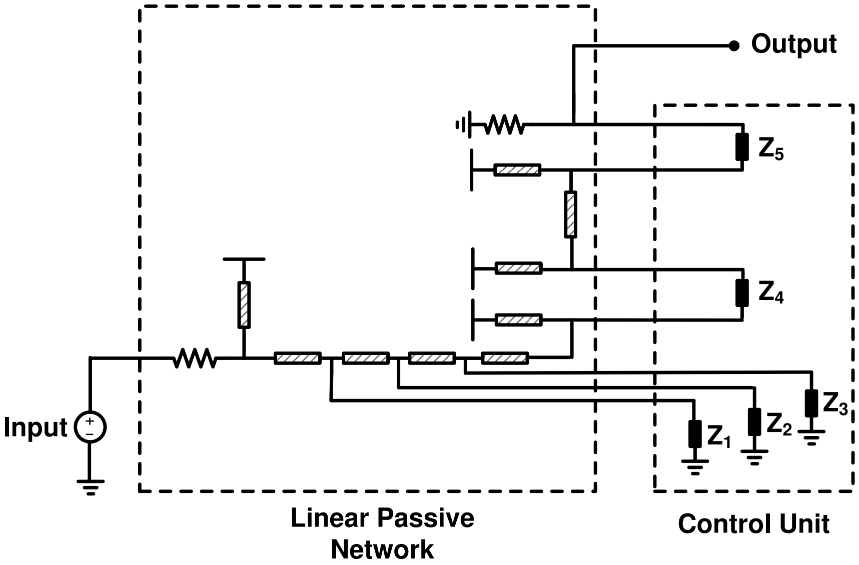


Figure 1.3: A re-arrangement of the circuit depicted in Figure 1.2

a known linear multi-port passive network and an unknown control unit. The objective is to design the control unit in such a way that certain linear and convex constraints on the input and outputs of the circuit are satisfied. Note that several problems in analog circuits, electromagnetics, and optics can be reformulated as this circuit design problem. Moreover, the OPF problem with variable shunt elements studied in Chapter 3 can also be cast as a very similar circuit problem.

In Chapter 4 of this dissertation, the circuit given in Figure 1.4 is studied with the aim of understanding what type of control unit simplifies the design process. To this end, it is shown that finding a control unit in the form of a switching network, which is desirable for antenna applications, to meet the design specifications is an NP-complete problem. Instead, the design of a control unit in the from of an unstructured passive network can be cast as a semidefinite optimization. Since this passive network may require many components

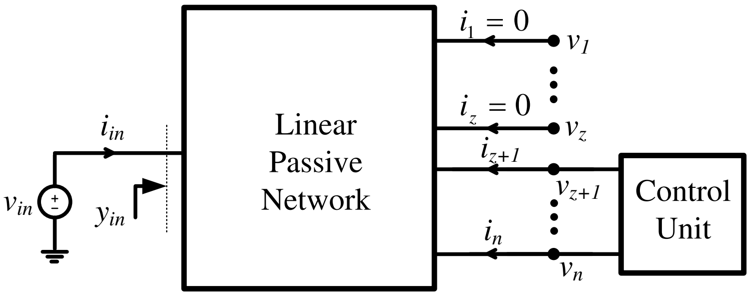


Figure 1.4: A generic linear circuit with an unknown control unit

(electrical elements) for its implementation, the design of a sparse passive controller is also studied. To this end, a rank-minimization problem is obtained that can be handled using the convex-based heuristic method proposed in [27]. It is verified that this heuristic method is able to solve the rank minimization problem exactly in certain practical examples.

1.2.2 Antenna Systems

Conventional antennas for wireless transmission, e.g., omni-directional antennas, radiate in almost all directions. In order to save power, improve security and avoid co-channel interference, much attention has been paid to designing smart transmitting/receiving antenna systems that can steer the beam towards a desired direction and make nulls in undesired directions [66]. There are two main types of smart antennas as follows:

- *Array (active) antenna system:* This type of smart antenna comprises multiple active radiating elements for varying the relative phases and amplitudes of the respective signals in order to generate a desired radiation pattern. Although this type of antenna is easy to program, its implementation is costly. The reason is that a satisfactory radiation pattern can be attained only if several active elements are used and in addition the size of the antenna system is several multiples of the signal wavelength [30, 120].

- *Passive antenna system:* This type of smart antenna employs only one active element surrounded by a number of tunable passive parasitic elements [79]. Although a passive antenna can be implemented fairly easily, its programming is very hard. Indeed, the radiation pattern of this type of antenna is a nonlinear function of the passive elements to be optimized. As a consequence of this nonlinearity property, it has been reported in [97] that an optimal programming of such an antenna (based on heuristic methods) can take as long as 4 weeks or more.

Chapter 5 of this dissertation aims to build on the results developed in [2] and Chapter 4 to propose a new type of smart antenna system, referred to as *passively controllable smart (PCS) antenna*, which is both easy to program and easy to implement. A PCS antenna system is composed of a main dipole (transmitting) antenna, a number of reflectors (or a patch array), and a variable (tunable) passive controller. Since changing the parameters of the passive controller modifies the radiation pattern generated at the far field, this act is regarded as *programming* of the PCS antenna. To study the programming capabilities of a PCS antenna, a number of receiving nodes are placed around the PCS antenna, which are all equipped with short dipole antennas for signal reception. It is shown that a pre-determined set of voltages can be sent to the receiving nodes if and only if the vector of voltages satisfies an LMI problem. Using this result, it is proved that a pre-specified radiation pattern can be generated for the receiving antennas if and only if the associated vector of voltages belongs to an ellipsoidal region. This region characterizes both the individual signals that can be sent to different directions as well as the correlation among them. Based on the obtained properties, it is shown how the PCS antenna can be programmed to transmit data to an intended node in such a way that many of the unintended nodes receive a zero signal (no signal) simultaneously. Finally, an on-chip wavelength-size passive antenna is designed in a few seconds (rather than 4 weeks) that can steer the beam to different directions and make nulls in at least 8 directions concurrently.

1.3 Part III: Distributed Computation

During the past few decades, there has been a particular interest in the area of distributed computation, which aims to compute some quantity over a network of processors in a decentralized fashion [107, 108]. The distributed averaging problem, as a particular case, is concerned with computing the average of numbers owned by the agents of a group [82]. Motivated by a variety of applications, this problem has been investigated through the notion of *consensus* in several papers [57, 103]. A more advanced type of distributed computation appears in the network resource allocation problem, where the goal is to find optimal transmission rates for a group of users sharing the links of a communication network.

Part III of this dissertation studies both the consensus and the distributed resource allocation problems. We will explain the details of these problems together with our contributions in the sequel.

1.3.1 Consensus

Several applications of the consensus problem have been reported in the literature; for instance, the synchronization of coupled oscillators, arising in biophysics, neurobiology, and systems biology, has been studied in [57] and [103] to explore how to reach a consensus on the frequencies of some agents. Moreover, the problem of aligning the heading angles of a group of mobile agents (e.g., a flock of birds) has been treated in [47]. Given a sensor network comprising a set of sensors measuring the same quantity in a noisy environment, the problem of consensus on state estimates has been discussed in [100].

Consider the distributed averaging problem in which the values associated with a set of agents are to be averaged in a distributed way. Since all agents cannot update their numbers synchronously in many practical situations, gossip algorithms have been widely exploited by researchers to handle the averaging problem asynchronously [107]. This type of algorithm selects a pair of agents at each time instant and updates their values based on some averaging policy. The consensus problem in the context of gossip algorithms has been thoroughly investigated in the literature [12]. In light of communication constraints, the data being exchanged between a pair of agents is normally quantized. This has given rise to the emergence of quantized gossip algorithms. The notion of *quantized consensus* has been introduced in [52] for the case when quantized values (integers) are to be averaged over a connected network with digital communication channels. This result has been extended in [28] to the case when the quantization is uniform, and the initial values of the agents are real (as opposed to being integer).

The problem of quantized consensus via gossip algorithm is studied in Chapter 6 of this dissertation. In this chapter, a weighted connected graph is considered together with a set of scalars sitting on its vertices. The weight of each edge represents the probability of establishing a communication between its corresponding vertices through the updating procedure. First, it is shown that a quantized consensus is reached under the stochastic gossip algorithm proposed in [28], for a wide range of updating parameters and an arbitrary quantizer. The convergence time of the gossip algorithm is then studied. More precisely, consider the expected value of the time at which a quantized consensus is reached, and take its maximum over all possible initial states belonging to a given hypercube. Lower and upper bounds on this quantity are provided for a uniform quantizer, which turn out to be related to the Laplacian of the weighted graph. The upper bound is then minimized in order to obtain the best weights resulting in a small convergence time. To do so, a convex optimization problem is proposed, which can be solved by a semidefinite program.

1.3.2 Distributed Resource Allocation

As a special type of distributed computation, the network resource allocation problem has been extensively studied in the context of congestion control ever since the first congestion collapse occurred in the Internet [46, 98, 101, 23]. The main idea behind the existing congestion control (resource allocation) algorithms is more or less the same: each user measures some feedback signal, such as packet loss or queueing delay, and accordingly adapts its transmission rate. From a mathematical standpoint, the available resource allocation algorithms, namely the primal, dual, and primal/dual algorithms, are particularly designed to solve the underlying problem in a distributed way asymptotically and ignore the transient behavior of the network. More precisely, the existing congestion control algorithms are obtained through the following steps [98]:

i) A static utility maximization problem is introduced and then the corresponding KKT conditions are derived.

ii) Since the solution of the KKT conditions is to be found in a distributed way, a dynamical updating algorithm is derived to solve the KKT conditions asymptotically.

This technique ignores the real-time behavior of the network. As a result, the link capacity constraints can, for instance, be violated in this period. Furthermore, the current algorithms have not been derived in such a way that they can be generalized systematically to include real-time constraints such as a link capacity requirement.

In Chapter 7 of this dissertation, the congestion control problem is revisited from the viewpoint of the optimal control theory. Indeed, it is proved that the resource allocation controllers proposed by the primal, dual, and primal/dual algorithms all maximize some meaningful dynamical behaviors. More precisely, there exist natural utility functionals whose maximization leads to these celebrated controllers. This result opens the possibility of tackling network problems directly as optimal control problems, which not only take the dynamics into account, but which also allow including physical constraints. Two other applications of dealing with utility functionals directly are: (i) in deducing the stability of the control system for free, and (ii) in gaining insight into how to perform joint routing and congestion control. The ideas developed in this chapter open up the possibility of designing optimal transmission protocols guaranteeing the real-time performance of the network, something which can never be realized by the existing congestion controllers.

1.4 Design of Smart Grid as an Interdisciplinary Problem

In order to have a sustainable future, the area of energy systems has received much attention in the past few years. In particular, a huge effort has been made to upgrade the existing electricity grid into a smart grid, which not only has optimized performance, efficiency and reliability, but is also friendly to the environment. A smart grid can be envisaged as a modernized power network whose control and operation rely on information technology. To design a smart grid, many considerations should be taken into account, including the following [25]:

- The grid should be able to accommodate distributed generation so that many small distributed generation units (rather than only a few hundred bulk generators) can exist over each area.

- A part of the energy in the grid must be penetrated based on renewable resources such as solar and wind.

- The grid should be associated with an efficient demand-response strategy so that consumers can actively participate in energy savings.

- Large-scale charging of electric vehicles should be possible in the gird.

- Guaranteed reliability and efficiency are needed for the operation of the grid, especially when a fault happens in the network.

A typical smart grid with the aforementioned properties includes all types of the large-scale complex systems studied in this dissertation. Therefore, the results of this work are all useful for the design of a smart grid, as described below:

- There are several nonlinear and nonconvex optimization problems associated with a smart grid, from operation planning to pricing and optimal charging of electric vehicles. Part I of this dissertation proposes an efficient method to handle those optimization problems.

- Part II of this work has applications in an important part of a smart grid, referred to as *advanced metering infrastructure* (AMI), which is in charge of measuring, collecting and analyzing energy usage of consumers [16]. Indeed, as opposed to an electromechanical meter, every house should be equipped with a smart meter with communication capabilities so that it will periodically transmit the load profile to a utility company in a wireless way. However, since reading this wireless signal reveals important information about the activities inside the house, this has created serious security and privacy concerns [70, 6]. The ideas developed in Part II can be used for designing efficient communication devices (e.g., on-chip directional antenna) to guarantee the security at the physical layer.

- Two independent results are derived in Part III of this dissertation, which are both useful for the control and operation of a smart grid. To be more precise, the state of a smart grid should be estimated in a distributed way and this problem is related to the first result of Part III on quantized consensus [26, 114]. In addition, a smart grid is composed of a great number of wireless devices which want to transmit delay-sensitive data over shared communication channels. Hence, the transmission rates of these devices must be controlled by some appropriate protocols [16]. The second result of Part III is on designing network protocols with a guaranteed real-time performance, which can be used for the transmission control of delay-sensitive signals in a smart grid.

Part I

Power Networks

Chapter 2

Zero Duality Gap in Optimal Power Flow Problem

The optimal power flow (OPF) problem is nonconvex and generally hard to solve. In this chapter, we propose a semidefinite programming (SDP) optimization, which is the dual of an equivalent form of the OPF problem. A global optimum solution to the OPF problem can be retrieved from a solution of this convex dual problem whenever the duality gap is zero. A necessary and sufficient condition is provided here to guarantee the existence of no duality gap for the OPF problem. This condition is satisfied by the standard IEEE benchmark systems with 14, 30, 57, 118, and 300 buses as well as several randomly generated systems. Since this condition is hard to study, a sufficient zero-duality-gap condition is also derived. This sufficient condition holds for IEEE systems after small resistance (10^{-5} per unit) is added to every transformer that originally assumes zero resistance. We investigate this sufficient condition and justify that it holds widely in practice. The main underlying reason for the successful convexification of the OPF problem can be traced back to the modeling of transformers and transmission lines as well as the non-negativity of physical quantities such as resistance and inductance.

2.1 Introduction

The optimal power flow (OPF) problem deals with finding an optimal operating point of a power system that minimizes an appropriate cost function such as generation cost or transmission loss subject to certain constraints on power and voltage variables [72]. Started by the work [21] in 1962, the OPF problem has been extensively studied in the literature and numerous algorithms have been proposed for solving this highly nonconvex problem [40, 106, 112], including linear programming, Newton Raphson, quadratic programming, nonlinear programming, Lagrange relaxation, interior point methods, artificial intelligence, artificial neural network, fuzzy logic, genetic algorithm, evolutionary programming and particle swarm optimization [72, 73, 74, 83]. A good number of these methods are based on the Karush-Kuhn-Tucker (KKT) necessary conditions, which can only guarantee a locally optimal solution, in light of the nonconvexity of the OPF problem [113]. This nonconvexity is partially due to the nonlinearity of physical quantities such as active power, reactive power and voltage magnitude. In the past decade, much attention has been paid to devising efficient algorithms with guaranteed performance for the OPF problem. For instance, the recent papers [67] and [48] propose nonlinear interior-point algorithms for an equivalent current injection model of the problem. An improved implementation of the automatic differentiation technique for the OPF problem is studied in the recent work [49]. In an effort to convexify the OPF problem, it is shown in [43] that the load flow problem of a radial distribution system can be modeled as a convex optimization problem in the form of a conic program. Nonetheless, the results fail to hold for a meshed network, due to the presence of arctangent equality constraints [44]. Nonconvexity appears in more sophisticated power problems such as the stability constrained OPF problem where the stability at the operating point is an extra constraint [29, 17] or the dynamic OPF problem where the dynamics of the generators are also taken into account [117, 116]. The recent paper [4] also proposes a convex relaxation to solve the OPF problem efficiently and tests its results on IEEE systems. Some of the results derived in the present work are related to this well-known convex relaxation. However, [4] drops the rank constraint of the original OPF without any justification in order to obtain the SDP formulation. We have derived the conditions under which the SDP relaxation is exact.

As will be shown in this chapter, the OPF problem is NP-hard in the worst case. Our recent work also proves that a closely related problem of finding an optimal operating point of a radiating antenna circuit is an NP-complete problem, by reducing the number partitioning problem to the antenna problem [61]. The goal of the present work is to exploit the physical properties of power systems and obtain a polynomial-time algorithm to find a global optimum of the OPF problem.

In this chapter, we suggest solving the dual of an equivalent form of the OPF problem (referred to as the Dual OPF problem), rather than the OPF problem itself. This dual problem is a convex semidefinite program and therefore can be solved efficiently (in polynomial time). However, the optimal objective value of the dual problem is only a lower bound on the optimal value of the original OPF problem and the lower bound may not be tight (in presence of a nonzero duality gap) [13]. A globally optimal solution to the OPF problem can be recovered from a solution to the Dual OPF problem if the duality gap is zero, meaning that strongly duality holds between these two optimizations. In this chapter, we derive a necessary and sufficient condition to guarantee the existence of no duality gap. Interestingly, this condition is satisfied for all the five IEEE benchmark systems archived at [109] with 14, 30, 57, 118, and 300 buses, in addition to several randomly generated systems. In other words, these practical systems can all be convexified naturally via the new formulation proposed here. In order to study why the duality gap is zero for the IEEE systems, we derive a sufficient zero-duality-gap condition, which reveals many useful properties of power systems. This sufficient condition holds for IEEE systems after a small perturbation in a few entries of the admittance matrix, in order to make the graph corresponding to the resistive part of the power network strongly connected.

To study the sufficient zero-duality-gap condition provided, we first consider a resistive network with only resistive loads. The OPF problem in this special case is also NP-hard. We exploit some physical properties of power circuits and prove that the duality gap is zero for a modified version of the OPF problem. Later on, we show that this modified OPF problem is expected to have the same solution as the OPF problem. The results are then extended to general networks with no constraints on reactive loads. It is shown that by fixing the real part of the admittance matrix Y, there is an unbounded region so that if the imaginary part of Y belongs to that region, the duality gap is zero. In other words, it is shown that there is an unbounded set of network topologies for which the duality gap is zero for all possible values of loads and physical limits. The results are then extended to a general OPF problem. It is worth mentioning that we have proved in [60] that zero duality gap for the classical OPF problem studied here implies zero duality gap for a general OPF-based problem in which there could be more variables (such as transformer ratios and variable shunt elements) and more constraints (such as dynamic or contingency constraints). Hence, the results of this work make it possible to convexify several fundamental power problems that have been studied for about half a century.

Notations: The following notations are used throughout this chapter:

- i : The imaginary unit
- **R**: The set of real numbers
- Re{·} and Im{·}: The operators returning the real and imaginary parts of a complex matrix
- $*$: The conjugate transpose operator
- T : The transpose operator
- $\succeq$ and $\preceq$: The matrix inequality signs in the positive semidefinite sense (i.e., given two symmetric matrices A and B, $A \succeq B$ implies $A - B$ is a positive semidefinite matrix, meaning that its eigenvalues are all nonnegative)
- Tr: The matrix trace operator
- $|\cdot|$: The absolute value operator

2.2 OPF Problem: Formulation and Computational Complexity

2.2.1 Problem Formulation

Consider a power network with the set of buses $\mathcal{N} := \{1, 2, ..., n\}$, the set of generator buses $\mathcal{G} \subseteq \mathcal{N}$ and the set of flow lines $\mathcal{L} \subseteq \mathcal{N} \times \mathcal{N}$. Define the parameters of the system as follows:

- $P_{D_k} + Q_{D_k}\mathrm{i}$: The given apparent power of the load connected to bus $k \in \mathcal{N}$ (this number is zero whenever bus k is not connected to any load)
- $P_{G_k} + Q_{G_k}\mathrm{i}$: The apparent power of the generator connected to bus $k \in \mathcal{G}$
- V_k: Complex voltage at bus $k \in \mathcal{N}$
- P_{lm}: Active power transferred from bus $l \in \mathcal{N}$ to the rest of the network through line $(l, m) \in \mathcal{L}$

- S_{lm}: Apparent power transferred from bus $l \in \mathcal{N}$ to the rest of the network through line $(l,m) \in \mathcal{L}$

- $f_k(P_{G_k}) = c_{k2}P_{G_k}^2 + c_{k1}P_{G_k} + c_{k0}$: Quadratic cost function with the given nonnegative coefficients accounting for the cost of active power generation at bus $k \in \mathcal{G}$

Let $\mathbf{V}$, $\mathbf{P}_g$, and $\mathbf{Q}_g$ denote the unknown sets $\{V_k\}_{k\in\mathcal{N}}$, $\{P_{G_k}\}_{k\in\mathcal{G}}$, and $\{Q_{G_k}\}_{k\in\mathcal{G}}$, respectively. The classical OPF problem aims to minimize $\sum_{k\in\mathcal{G}} f_k(P_{G_k})$ over the unknown parameters $\mathbf{V}$, $\mathbf{P}_g$, and $\mathbf{Q}_g$ subject to the power balance equations at all buses and the physical constraints

$$P_k^{\min} \leq P_{G_k} \leq P_k^{\max}, \qquad \forall k \in \mathcal{G} \tag{2.1a}$$

$$Q_k^{\min} \leq Q_{G_k} \leq Q_k^{\max}, \qquad \forall k \in \mathcal{G} \tag{2.1b}$$

$$V_k^{\min} \leq |V_k| \leq V_k^{\max}, \qquad \forall k \in \mathcal{N} \tag{2.1c}$$

$$|S_{lm}| \leq S_{lm}^{\max}, \qquad \forall (l,m) \in \mathcal{L} \tag{2.1d}$$

$$|P_{lm}| \leq P_{lm}^{\max}, \qquad \forall (l,m) \in \mathcal{L} \tag{2.1e}$$

$$|V_l - V_m| \leq \Delta V_{lm}^{\max}, \qquad \forall (l,m) \in \mathcal{L} \tag{2.1f}$$

where $P_k^{\min}, P_k^{\max}, Q_k^{\min}, Q_k^{\max}, V_k^{\min}, V_k^{\max}, S_{lm}^{\max}, P_{lm}^{\max}, \Delta V_{lm}^{\max}$ are some given real numbers such that $S_{lm}^{\max} = S_{ml}^{\max}$ and $P_{lm}^{\max} = P_{ml}^{\max}$. Note that some of the constraints stated in (2.1) may not be needed for a practical OPF problem, in which case the undesired constraints can be removed by setting the corresponding lower/upper bounds as infinity. For instance, the line flow constraints (2.1d) and (2.1e) might not be necessary simultaneously or the constraint (2.1f) could be redundant, depending on the situation. Although not stated explicitly, we assume throughout this work that the OPF problem is feasible and that $\mathbf{V} = 0$ does not satisfy its constraints.

Derive the circuit model of the power network by replacing every transmission line and transformer with their equivalent Π models [72]. In this circuit model, let y_{kl} denote the mutual admittance between buses k and l, and y_{kk} denote the admittance-to-ground at bus k, for every $k, l \in \mathcal{N}$ (note that $y_{kl} = 0$ if $(k,l) \notin \mathcal{L}$). Let Y represent the admittance matrix of this equivalent circuit model, which is an $n \times n$ complex-valued matrix whose (k,l) entry is equal to $-y_{kl}$ if $k \neq l$ and $y_{kk} + \sum_{m\in\mathcal{N}(k)} y_{km}$ otherwise, where $\mathcal{N}(k)$ denotes the set of all buses that are directly connected to bus k. Define the current vector $\mathbf{I} :=$

$\begin{bmatrix} I_1 & I_2 & \cdots & I_n \end{bmatrix}^T$ as $Y\mathbf{V}$. Note that I_k represents the net current injected to bus $k \in \mathcal{N}$.

It is shown in Appendix 2.7.2 that the OPF problem is NP-hard, which implies that an arbitrary (general) OPF problem may not be solvable in polynomial time. However, the goal is to show that an OPF problem corresponding to a practical power network is structured in such a way that it might be solved efficiently in polynomial time even if it could have multiple local minima with a nonconvex (disconnected) feasibility region.

2.3 New Approach to Solving OPF

By denoting the standard basis vectors in $\mathbf{R}^n$ as $e_1, e_2, ..., e_n$, let a number of matrices be defined now for every $k \in \mathcal{N}$ and $(l, m) \in \mathcal{L}$:

$$\begin{aligned}
Y_k &:= e_k e_k^T Y, \qquad Y_{lm} := (\bar{y}_{lm} + y_{lm}) e_l e_l^T - (y_{lm}) e_l e_m^T \\
\mathbf{Y}_k &:= \frac{1}{2} \begin{bmatrix} \mathrm{Re}\{Y_k + Y_k^T\} & \mathrm{Im}\{Y_k^T - Y_k\} \\ \mathrm{Im}\{Y_k - Y_k^T\} & \mathrm{Re}\{Y_k + Y_k^T\} \end{bmatrix} \\
\mathbf{Y}_{lm} &:= \frac{1}{2} \begin{bmatrix} \mathrm{Re}\{Y_{lm} + Y_{lm}^T\} & \mathrm{Im}\{Y_{lm}^T - Y_{lm}\} \\ \mathrm{Im}\{Y_{lm} - Y_{lm}^T\} & \mathrm{Re}\{Y_{lm} + Y_{lm}^T\} \end{bmatrix} \\
\bar{\mathbf{Y}}_k &:= \frac{-1}{2} \begin{bmatrix} \mathrm{Im}\{Y_k + Y_k^T\} & \mathrm{Re}\{Y_k - Y_k^T\} \\ \mathrm{Re}\{Y_k^T - Y_k\} & \mathrm{Im}\{Y_k + Y_k^T\} \end{bmatrix} \\
\bar{\mathbf{Y}}_{lm} &:= \frac{-1}{2} \begin{bmatrix} \mathrm{Im}\{Y_{lm} + Y_{lm}^T\} & \mathrm{Re}\{Y_{lm} - Y_{lm}^T\} \\ \mathrm{Re}\{Y_{lm}^T - Y_{lm}\} & \mathrm{Im}\{Y_{lm} + Y_{lm}^T\} \end{bmatrix} \\
M_k &:= \begin{bmatrix} e_k e_k^T & 0 \\ 0 & e_k e_k^T \end{bmatrix} \\
M_{lm} &:= \begin{bmatrix} (e_l - e_m)(e_l - e_m)^T & 0 \\ 0 & (e_l - e_m)(e_l - e_m)^T \end{bmatrix} \\
\mathbf{X} &:= \begin{bmatrix} \mathrm{Re}\{\mathbf{V}\}^T & \mathrm{Im}\{\mathbf{V}\}^T \end{bmatrix}^T
\end{aligned}$$

where $\bar{y}_{lm}$ denotes the value of the shunt element at bus l associated with the Π model of the line (l, m). For every $k \in \mathcal{N}$, define $P_{k,\mathrm{inj}}$ and $Q_{k,\mathrm{inj}}$ as the net active and reactive powers injected to bus k, i.e.,

$$\begin{aligned} P_{k,\text{inj}} &:= P_{G_k} - P_{D_k}, && \forall k \in \mathcal{G} \\ Q_{k,\text{inj}} &:= Q_{G_k} - Q_{D_k}, && \forall k \in \mathcal{G} \\ P_{k,\text{inj}} &:= -P_{D_k}, && \forall k \in \mathcal{N}\backslash\mathcal{G} \\ Q_{k,\text{inj}} &:= -Q_{D_k}, && \forall k \in \mathcal{N}\backslash\mathcal{G}. \end{aligned}$$

Lemma 1 *The following relations hold for every $k \in \mathcal{N}$ and $(l,m) \in \mathcal{L}$:*

$$P_{k,inj} = Tr\left\{\mathbf{Y}_k\mathbf{X}\mathbf{X}^T\right\} \tag{2.2a}$$

$$Q_{k,inj} = Tr\left\{\bar{\mathbf{Y}}_k\mathbf{X}\mathbf{X}^T\right\} \tag{2.2b}$$

$$P_{lm} = Tr\left\{\mathbf{Y}_{lm}\mathbf{X}\mathbf{X}^T\right\} \tag{2.2c}$$

$$|S_{lm}|^2 = \left(Tr\left\{\mathbf{Y}_{lm}\mathbf{X}\mathbf{X}^T\right\}\right)^2 + \left(Tr\left\{\bar{\mathbf{Y}}_{lm}\mathbf{X}\mathbf{X}^T\right\}\right)^2 \tag{2.2d}$$

$$|V_k|^2 = Tr\left\{M_k\mathbf{X}\mathbf{X}^T\right\} \tag{2.2e}$$

$$|V_l - V_m|^2 = Tr\left\{M_{lm}\mathbf{X}\mathbf{X}^T\right\}. \tag{2.2f}$$

Proof: See Appendix 2.7.3.

Extend the definition of $P_k^{\min}, P_k^{\max}, Q_k^{\min}, Q_k^{\max}$ from $k \in \mathcal{G}$ to every $k \in \mathcal{N}$, with $P_k^{\min} = P_k^{\max} = Q_k^{\min} = Q_k^{\max} = 0$ if $k \in \mathcal{N}\backslash\mathcal{G}$. Using Lemma 1, one can formulate the OPF problem in terms of $\mathbf{X}$ as follows.

OPF problem formulated in X: Minimize

$$\sum_{k\in\mathcal{G}}\left\{c_{k2}\left(\text{Tr}\left\{\mathbf{Y}_kW\right\} + P_{D_k}\right)^2 + c_{k1}\left(\text{Tr}\left\{\mathbf{Y}_kW\right\} + P_{D_k}\right) + c_{k0}\right\} \tag{2.3}$$

over the variables $\mathbf{X} \in \mathbf{R}^{2n}$ and $W \in \mathbf{R}^{2n\times 2n}$ subject to the following constraints for every $k \in \mathcal{N}$ and $(l,m) \in \mathcal{L}$

$$P_k^{\min} - P_{D_k} \leq \operatorname{Tr}\{\mathbf{Y}_k W\} \leq P_k^{\max} - P_{D_k} \tag{2.4a}$$

$$Q_k^{\min} - Q_{D_k} \leq \operatorname{Tr}\{\bar{\mathbf{Y}}_k W\} \leq Q_k^{\max} - Q_{D_k} \tag{2.4b}$$

$$(V_{k,\min})^2 \leq \operatorname{Tr}\{M_k W\} \leq (V_k^{\max})^2 \tag{2.4c}$$

$$\operatorname{Tr}\{\mathbf{Y}_{lm} W\}^2 + \operatorname{Tr}\{\bar{\mathbf{Y}}_{lm} W\}^2 \leq (S_{lm}^{\max})^2 \tag{2.4d}$$

$$\operatorname{Tr}\{\mathbf{Y}_{lm} W\} \leq P_{lm}^{\max} \tag{2.4e}$$

$$\operatorname{Tr}\{M_{lm} W\} \leq (\Delta V_{lm}^{\max})^2 \tag{2.4f}$$

$$W = \mathbf{X}\mathbf{X}^T. \tag{2.4g}$$

Note that the constraint $|P_{lm}| \leq P_{lm}^{\max}$ in the original OPF problem is changed to $P_{lm} \leq P_{lm}^{\max}$ in order to derive (2.4e). This modification can be done in light of the relations $P_{lm} + P_{ml} \geq 0$ and $P_{lm}^{\max} = P_{ml}^{\max}$. The above OPF formulation is not quadratic in $\mathbf{X}$, due to the objective function being of degree 4 with respect to the entries of $\mathbf{X}$ as well as the constraint (2.4d). However, one can define some auxiliary variables to reformulate the OPF problem in a quadratic way with respect to X. To this end, Schur's complement formula yields that the constraint (2.4d) can be replaced by

$$\begin{bmatrix} -(S_{lm,\max})^2 & \operatorname{Tr}\{\mathbf{Y}_{lm} W\} & \operatorname{Tr}\{\bar{\mathbf{Y}}_{lm} W\} \\ \operatorname{Tr}\{\mathbf{Y}_{lm} W\} & -1 & 0 \\ \operatorname{Tr}\{\bar{\mathbf{Y}}_{lm} W\} & 0 & -1 \end{bmatrix} \preceq 0. \tag{2.5}$$

On the other hand, given a scalar α_k for some $k \in \mathcal{G}$, the constraint $f_k(P_{G_k}) < \alpha_k$ is equivalent to (by Schur's complement formula)

$$\begin{bmatrix} c_{k1}\operatorname{Tr}\{\mathbf{Y}_k W\} - \alpha_k + a_k & \sqrt{c_{k2}}\ \operatorname{Tr}\{\mathbf{Y}_k W\} + b_k \\ \sqrt{c_{l2}}\ \operatorname{Tr}\{\mathbf{Y}_k W\} + b_k & -1 \end{bmatrix} \preceq 0 \tag{2.6}$$

where $a_k := c_{k0} + c_{k1} P_{D_k}$ and $b_k := \sqrt{c_{k2}} P_{D_k}$.

Using (2.5) and (2.6), one can reformulate the OPF problem formalized in (2.3) and (2.4) in a quadratic way. This leads to Optimization 1 given below, which is equivalent to the OPF problem.

Optimization 1: Minimize $\sum_{k\in\mathcal{G}}\alpha_k$ over the scalar variables α_k's and the matrix variables $\mathbf{X}$ and W subject to the constraints (2.4a), (2.4b), (2.4c), (2.4e), (2.4f), (2.4g), (2.5), and (2.6).

The variable $\mathbf{X}$ can be eliminated from Optimization 1 by using the fact that a given matrix W can be written as $\mathbf{X}\mathbf{X}^T$ for some (nonzero) vector $\mathbf{X}$ if and only if W is both positive semidefinite and rank 1. Hence, Optimization 2 proposed below is an equivalent form of Optimization 1 whose variables are only W and α_k's for $k \in \mathcal{G}$.

Optimization 2: This optimization is obtained from Optimization 1 by replacing the constraint (2.4g), i.e., $W = \mathbf{X}\mathbf{X}^T$, with the new constraints $W \succeq 0$ and $\text{rank}\{W\} = 1$.

Notice that since Optimization 2 has a rank constraint, it is nonconvex. However, removing the constraint $\text{rank}\{W\} = 1$ from this optimization makes it a semidefinite program (SDP), which is a convex problem (see Appendix 2.7.1 for a brief overview of SDP). This gives rise to Optimization 3 presented below.

Optimization 3: This optimization is obtained from Optimization 2 by removing the rank constraint $\text{rank}\{W\} = 1$.

Optimization 3 is indeed an SDP relaxation of the OPF problem. Assume that this convex optimization problem has a rank-one optimal solution W^{opt}. Then, there exists a vector $\mathbf{X}^{\text{opt}}$ such that $W^{\text{opt}} = \mathbf{X}^{\text{opt}}(\mathbf{X}^{\text{opt}})^T$. In that case, $\mathbf{X}^{\text{opt}}$ is a global optimum of the OPF problem. However, since the OPF problem is NP-hard in general, Optimization 3 does not always have a rank-one solution. We numerically solved this optimization problem for IEEE test systems with 14, 30, 57, 118, and 300 buses using SEDUMI and noticed that each solution W^{opt} obtained always has rank two. The next lemma explains the reason why this occurs for IEEE systems.

Lemma 2 *If Optimization 3 has a rank-one solution, then it must have an infinite number of rank-two solutions.*

Proof: See Appendix 2.7.3.

Lemma 2 states that Optimization 3 might have a rank-one solution that cannot be easily identified by solving it numerically. However, using the method proposed later in this work, one can verify that Optimization 3 always has a rank-one solution for all aforementioned IEEE test systems. This implies that these power systems can be convexified by a convex relaxation technique. However, the focus of this chapter will not be on Optimization 3 due to the following reasons:

- The number of scalar variables of Optimization 3 is quadratic with respect to n (in light of the non-sparse structure of the matrix variable W_c). Hence, solving this optimization problem might be expensive and time-consuming for large values of n.

- Since Optimization 3 may have an infinite number of solutions (see Lemma 2), it is not clear how to numerically verify the existence of a rank-one solution.

- Optimization 3 has a general structure with several constraints from which it is hard to analytically study when this optimization has a rank-one solution.

In this chapter, we consider the dual of Optimization 3. To this end, define the following dual variables for every $k \in \mathcal{N}$ and $(l,m) \in \mathcal{L}$:

i) $\underline{\lambda}_k, \underline{\gamma}_k, \underline{\mu}_k$: Lagrange multipliers associated with the lower inequalities in (2.4a), (2.4b), and (2.4c), respectively.

ii) $\bar{\lambda}_k, \bar{\gamma}_k, \bar{\mu}_k$: Lagrange multipliers associated with the upper inequalities in (2.4a), (2.4b), and (2.4c), respectively.

iii) λ_{lm}, μ_{lm}: Lagrange multipliers associated with the inequalities (2.4e) and (2.4f), respectively.

iv) $r^1_{lm}, r^2_{lm}, ..., r^6_{lm}$: The matrix

$$\begin{bmatrix} r^1_{lm} & r^2_{lm} & r^3_{lm} \\ r^2_{lm} & r^4_{lm} & r^5_{lm} \\ r^3_{lm} & r^5_{lm} & r^6_{lm} \end{bmatrix}$$

is the Lagrange multiplier associated with the matrix inequality (2.5).

v) r^1_k, r^2_k: If $k \in \mathcal{G}$, the matrix

$$\begin{bmatrix} 1 & r^1_k \\ r^1_k & r^2_k \end{bmatrix} \tag{2.7}$$

is the Lagrange multiplier associated with the matrix inequality (2.6).

Let x and r denote the sets of all multipliers introduced in (i–iii) and (iv–v), respectively. Define some aggregate multipliers for every $k \in \mathcal{N}$ as follows

$$\begin{aligned}
\lambda_k &:= \begin{cases} -\underline{\lambda}_k + \bar{\lambda}_k + c_{k1} + 2\sqrt{c_{k2}}\, r_k^1 & \text{if } k \in \mathcal{G} \\ -\underline{\lambda}_k + \bar{\lambda}_k & \text{otherwise} \end{cases} \\
\gamma_k &:= -\underline{\lambda}_k + \bar{\lambda}_k \\
\mu_k &:= -\underline{\mu}_k + \bar{\mu}_k.
\end{aligned}$$

Furthermore, define the functions

$$\begin{aligned}
h(x,r) := \sum_{k \in \mathcal{N}} \Big\{ & \underline{\lambda}_k P_k^{\min} - \bar{\lambda}_k P_k^{\max} + \lambda_k P_{D_k} + \underline{\gamma}_k Q_k^{\min} - \bar{\gamma}_k Q_k^{\max} + \gamma_k Q_{D_k} \\
& + \underline{\mu}_k \left(V_k^{\min}\right)^2 - \bar{\mu}_k \left(V_k^{\max}\right)^2 \Big\} + \sum_{k \in \mathcal{G}} \left(c_{k0} - r_k^2\right) \\
& - \sum_{(l,m) \in \mathcal{L}} \Big\{ \lambda_{lm} P_{lm}^{\max} + \mu_{lm} \left(\Delta V_{lm}^{\max}\right)^2 + \left(S_{lm}^{\max}\right)^2 r_{lm}^1 + r_{lm}^4 + r_{lm}^6 \Big\}
\end{aligned}$$

and

$$\begin{aligned}
A(x,r) := & \sum_{k \in \mathcal{N}} \Big\{ \lambda_k \mathbf{Y}_k + \gamma_k \bar{\mathbf{Y}}_k + \mu_k M_k \Big\} \\
& + \sum_{(l,m) \in \mathcal{L}} \Big\{ \left(2 r_{lm}^2 + \lambda_{lm}\right) \mathbf{Y}_{lm} + 2 r_{lm}^3 \bar{\mathbf{Y}}_{lm} + \mu_{lm} M_{lm} \Big\}.
\end{aligned}$$

We propose an optimization problem in the sequel, which plays a central role in solving the OPF problem.

Optimization 4 (Dual OPF): Maximize the linear function $h(x,r)$ over the vectors $x \geq 0$ and r subject to the linear matrix inequalities

$$A(x,r) \succeq 0 \tag{2.8a}$$

$$\begin{bmatrix} r_{lm}^1 & r_{lm}^2 & r_{lm}^3 \\ r_{lm}^2 & r_{lm}^4 & r_{lm}^5 \\ r_{lm}^3 & r_{lm}^5 & r_{lm}^6 \end{bmatrix} \succeq 0, \quad \forall (l,m) \in \mathcal{L} \tag{2.8b}$$

$$\begin{bmatrix} 1 & r_k^1 \\ r_k^1 & r_k^2 \end{bmatrix} \succeq 0, \qquad \forall k \in \mathcal{G}. \tag{2.8c}$$

The next theorem presents some important properties of Optimization 4.

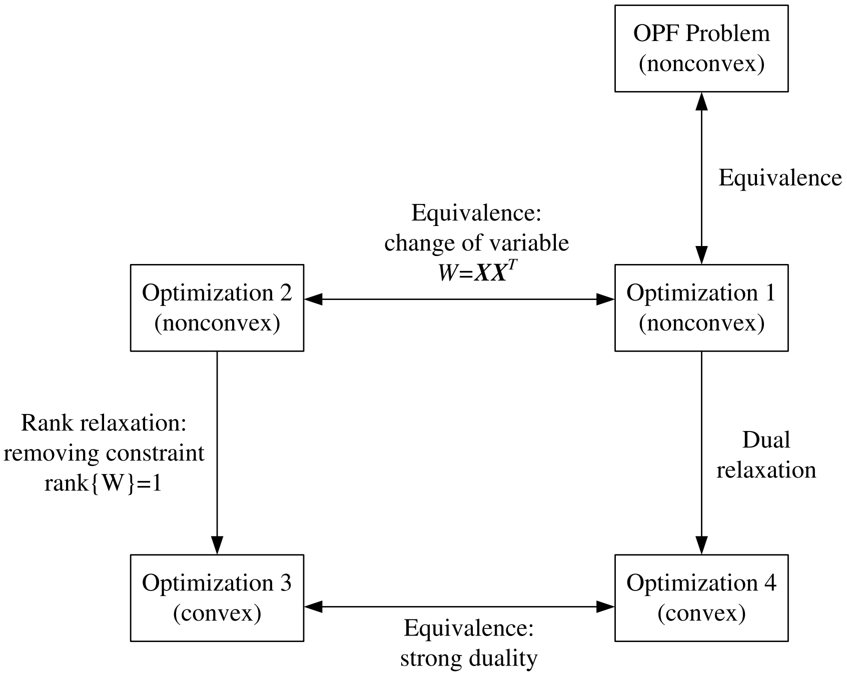


Figure 2.1: The relationship among OPF and Optimizations 1–4.

Theorem 1 *The following statements hold:*

i) Optimization 4 is the dual of the nonconvex problem of Optimization 1.

ii) Optimization 4 is the dual of Optimization 3 and strong duality holds between these optimizations. Moreover, the matrix variable W in Optimization 3 corresponds to a Lagrange multiplier for the inequality constraint $A(x, r) \succeq 0$ in Optimization 4.

Proof: See Appendix 2.7.3.

The relationship among the OPF problem and Optimizations 1–4 are illustrated in Figure 2.1. This chapter suggests solving Optimization 4, which is the dual of a reformulated OPF problem (i.e., Optimization 1) as well as the dual of a convex relaxation of the OPF problem (i.e., Optimization 3). Since Optimization 4 is an SDP, a globally optimization solution to this problem can be found in polynomial time. However, this solution can be used to retrieve a solution to the OPF problem only if the duality gap is zero for Optimization 1, meaning that the optimal objective values of Optimizations 1 and 4 are identical. The next theorem investigates this issue in more detail.

Theorem 2 *The following statements hold:*

i) The duality gap is zero for Optimization 1 if and only if the SDP Optimization 3 has a rank-one solution W^{opt}.

ii) The duality gap is zero for Optimization 1 if its dual (i.e., the SDP Optimization 4) has a solution (x^{opt}, r^{opt}) such that positive semidefinite matrix $A(x^{opt}, r^{opt})$ has a zero eigenvalue of multiplicity 2.

Proof: See Appendix 2.7.3.

Due to the reasons outlined right after Lemma 2, this chapter mainly focuses on Condition (ii) (as opposed to Condition (i)), whose usefulness will become clear later this work. The next corollary explains how to recover a solution to the OPF problem whenever this zero-duality-gap condition is satisfied.

Corollary 1 *If the zero-duality-gap condition (ii) given in Theorem 2 is satisfied, then the following properties hold:*

- *Given any nonzero vector $\begin{bmatrix} X_1^T & X_2^T \end{bmatrix}^T$ in the null space of $A(x^{opt}, r^{opt})$, there exist two real-valued scalars ζ_1 and ζ_2 such that $\mathbf{V}^{opt} = (\zeta_1 + \zeta_2 i)(X_1 + X_2 i)$ is a global optimum of the OPF problem.*

- *Given any arbitrary solution W^{opt} of Optimization 3, the rank of W^{opt} is at most 2. Moreover, if the matrix W^{opt} has rank 2, then the matrix $(\rho_1 + \rho_2)EE^T$ is a rank-one solution of Optimization 3, where ρ_1 and ρ_2 are the nonzero eigenvalues of W^{opt} and E is the unit eigenvector associated with ρ_1.*

Proof: See Appendix 2.7.3.

This work suggests the following strategy for finding a global optimum of the OPF problem.

Algorithm for Solving OPF:

1. Compute a solution $(x^{\text{opt}}, r^{\text{opt}})$ of Optimization 4, which is the dual of an equivalent form of the OPF problem.

2. If the optimal value $h(x^{\text{opt}}, r^{\text{opt}})$ is $+\infty$, then the OPF problem is infeasible.

3. Find the multiplicity of the zero eigenvalue of the matrix $A(x^{\text{opt}}, r^{\text{opt}})$ and denote it as ψ.

4. If ψ is greater than 2, it might not be possible to solve the OPF problem in polynomial time.

5. If ψ is less than or equal to 2, then use the method explained in Part (i) of Corollary 1 to find a globally optimal solution $\mathbf{V}^{\mathrm{opt}}$.

The main complexity of the above algorithm can be traced back to its Step 1, which requires solving the dual OPF problem. It is noteworthy that this optimization is an SDP problem and therefore can be solved in polynomial-time. We tested our algorithm on several randomly generated power systems with all types of constraints given in (2.1) and observed that this algorithm found a global optimum of the OPF problem for all trials. Then, we considered the IEEE test systems with 14, 30, 57, 118, and 300 buses, whose physical constraints are in the form of (2.1a)–(2.1d), and made the following observations:

- Optimization 3 always leads to a rank-two solution, from which a rank-one solution can be found using the technique delineated in Part (ii) of Corollary 1. Hence, Part (i) of Theorem 2 yields that the duality gap is zero for all these IEEE systems.

- Our algorithm based on the dual OPF works after a small perturbation of the matrix Y. More precisely, if a small resistance (10^{-5}) is added to each transformer that originally has zero resistance, the graph induced by the matrix $\mathrm{Re}\{Y\}$ will become connected for each aforementioned IEEE system. This perturbation makes ψ equal to 2.

Before studying why the OPF problem associated with a real power system is expected to be solvable using the algorithm proposed earlier, we make several important remarks below.

Remark 1 *The last step of the algorithm relies on Part (i) of Corollary 1, which states that there exist two real-valued scalars ζ_1 and ζ_2 such that $\mathbf{V}^{opt} = (\zeta_1 + \zeta_2 i)(X_1 + X_2 i)$. In order to find ζ_1 and ζ_2, two (linear) equations are required. The voltage angle at the swing bus being zero introduces one such equation. The second one can be formed by identifying the active voltage constraints. Indeed, if $\underline{\mu}_k^{opt}$ (respectively, $\bar{\mu}_k^{opt}$) turns out to be nonzero for some $k \in \mathcal{N}$, then the relation $|V_k^{opt}| = V_k^{\min}$ (respectively, $|V_k^{opt}| = V_k^{\max}$) must hold.*

Remark 2 *Optimization 4 has two interesting properties for a practical power system. Fist, since most of the constraints specified in (2.1) are likely to be inactive, the vectors x^{opt} and r^{opt} are sparse. Moreover, the number of variables of Optimization 4 is $O(|\mathcal{L}|) + O(|\mathcal{N}|)$, which is expected to be equal to $O(|\mathcal{N}|)$ due to the very sparse topology of real power systems. Note that solving Optimization 3 for very large-scale power networks might be too costly, in which case it is recommended to use some sub-gradient techniques [4, 91].*

Remark 3 *Optimization 4 has the interesting property that the given loads together with the physical limits on voltage and power parameters only appear in the objective function, whereas the network topology (the matrix Y) shows up in its linear matrix constraints. Therefore, there is a natural decomposition between the load profile and the network topology in Optimization 4. This useful property, besides the linearity of Optimization 4, makes it possible to solve many more sophisticated problems efficiently, such as the OPF problem with stochastic and time-varying loads, optimal network reconfiguration for minimizing power loss, etc.*

Remark 4 *Most of the algorithms proposed in the past decade to solve the OPF problem are built on the KKT conditions written for the original or a reformulated OPF problem. We highlight the differences between the Dual OPF and the KKT conditions in the following:*

- *The duality gap could be zero for an OPF problem whose feasibility region has several disjoint components (see Section 2.7.2). Hence, the OPF problem may have many local solutions, all of which satisfy the KKT conditions. In contrast, a global optimum of the OPF problem can be recovered by solving the Dual OPF in presence of no duality gap.*

- *The KKT conditions are based on both primal and dual variables (say $\mathbf{X}, x, r$), whereas the dual OPF depends only on the dual variables (say x, r).*

- *There is a constraint $A(x, r) \succeq 0$ in the dual OPF, and besides an optimal solution to the OPF problem satisfies the relation $A(x^{opt}, r^{opt})\mathbf{X}^{opt}$=0 . The constraint $A(x^{opt}, r^{opt})\mathbf{X}^{opt} = 0$ is part of the KKT conditions, implying that the matrix $A(x, r)$ should lose rank at optimality. However, the stronger constraint $A(x, r) \succeq 0$ is missing in the KKT conditions.*

Indeed, it can be shown that if the constraint $A(x,r) \succeq 0$ is incorporated into the KKT conditions, then the resulting conditions are able to find a global optimum of the OPF problem in the absence of a nonzero duality gap.

2.4 Zero Duality Gap for Power Systems

In this section, we study the zero-duality-gap condition (ii) given in Theorem 2 in more details to justify why this condition is expected to hold widely in practice. To this end, we first study the OPF problem for DC networks, which is indeed an NP-hard problem. This helps find the useful properties of the Dual OPF problem, which will later be used to explore the solvability of the OPF problem for AC networks.

2.4.1 Resistive Networks with Resistive Loads

As can be seen in Case (ii) of Section 2.7.2, the OPF problem is NP-hard even if the network is resistive and there are no reactive loads. This situation, which corresponds to DC power distribution, is itself important because (i) the active power loss in a power system is due to the resistive part of the network, and (ii) the study of this case reveals important facts about the general OPF problem. In this section, we prove the existence of no duality gap for DC networks under a mild assumption, which is expected to hold in reality.

Throughout this part, assume that the power system is a resistive network (i.e., $\text{Im}\{Y\} = 0$) and that all loads are resistive as well. In the formulation of the OPF problem, it was assumed that the (active) power to be delivered to the load of bus $k \in \mathcal{N}$ must be exactly equal to P_{D_k}. Let the OPF problem be changed to allow delivering any power more than P_{D_k} to the load of bus k. To this end, define P_{L_k} as the power delivered to the load of bus k and P_{D_k} as the desired power requested by the load of bus k. In the OPF problem, we have the constraints

$$P_{L_k} = P_{D_k}, \quad \forall k \in \mathcal{N}. \tag{2.9}$$

Modify the OPF problem by replacing the above constraints with the following

$$P_{L_k} \geq P_{D_k}, \quad \forall k \in \mathcal{N} \tag{2.10}$$

and name the resulting problem as *modified OPF problem*. Note that this variant of the OPF problem allows for the over-satisfaction of the loads. This idea has already been considered by some other papers too (see [45] and the references given therein). In what follows, we first study the modified OPF problem, and then explain why the OPF and modified OPF problems are expected to have the same solution.

Theorem 3 *The duality gap is zero for the modified OPF problem.*

Proof: It can be shown that the Dual (modified) OPF problem associated with the modified OPF problem is the same as Optimization 4 with the exception of having the extra constraints

$$\lambda_k \geq 0, \quad \forall k \in \mathcal{N}. \tag{2.11}$$

Let $(x^{\text{opt}}, r^{\text{opt}})$ denote a solution to the Dual modified OPF problem. The goal is to show that the multiplicity of the zero eigenvalue of $A(x^{\text{opt}}, r^{\text{opt}})$ is at most two. To this end, notice that the constraints (2.1b) and (2.1d) can be ignored due to the network and loads both being resistive (note that $S_{lk} = P_{lk}$ for DC networks). As a result,

$$\begin{aligned} &\gamma_k = 0, && \forall k \in \mathcal{N} \\ &r^1_{lm} = \cdots = r^6_{lm} = 0, && \forall (l,m) \in \mathcal{L}. \end{aligned}$$

Hence, the matrix $A(x^{\text{opt}}, r^{\text{opt}})$ can be expressed as

$$A(x^{\text{opt}}, r^{\text{opt}}) = \begin{bmatrix} T(x^{\text{opt}}, r^{\text{opt}}) & 0 \\ 0 & T(x^{\text{opt}}, r^{\text{opt}}) \end{bmatrix} \tag{2.12}$$

for some matrix $T(x^{\text{opt}}, r^{\text{opt}}) \in \mathcal{R}^{n \times n}$, where the (l,m) off-diagonal entry of $T(x^{\text{opt}}, r^{\text{opt}})$ is equal to

$$T_{lm}(x^{\text{opt}}, r^{\text{opt}}) = -\frac{y_{lm}}{2}\left(\lambda^{\text{opt}}_{lm} + \lambda^{\text{opt}}_{ml} + \lambda^{\text{opt}}_{l} + \lambda^{\text{opt}}_{m}\right) - \mu^{\text{opt}}_{lm} - \mu^{\text{opt}}_{ml}$$

if $(l,m) \in \mathcal{L}$ and is zero otherwise. On the other hand, since resistance is a nonnegative physical quantity, it can be shown that y_{lm} coming from the Π model of a transmission line or a transformer is always nonnegative. It follows from this fact together with the inequalities (2.11) and $x^{\text{opt}} \geq 0$ that all off-diagonal entries of the matrix $T(x^{\text{opt}}, r^{\text{opt}})$ are non-positive.

Assume for now that the graph of the power system is strongly connected, meaning that there exists a path between every two buses of the network [31]. Assume also that the nonnegative vector $(\lambda_1^{\text{opt}}, ..., \lambda_n^{\text{opt}})$ is strictly positive. These assumptions imply that the matrix $T(x^{\text{opt}}, r^{\text{opt}})$ is irreducible and its off-diagonal entries are non-positive. Hence, the Perron-Frobenius theorem yields that the smallest eigenvalue of $T(x^{\text{opt}}, r^{\text{opt}})$ is simple, and as a result of (2.12), the smallest eigenvalue of $A(x^{\text{opt}}, r^{\text{opt}})$ is repeated twice [31]. Since this matrix is positive semidefinite, this simply implies that the multiplicity of the zero eigenvalue of $A(x^{\text{opt}}, r^{\text{opt}})$ is at most 2. Thus, the duality gap is zero for the modified OPF problem, by virtue of Part (ii) of Theorem 2.

Now, suppose that the power network is strongly connected, but the nonnegative vector $(\lambda_1^{\text{opt}}, ..., \lambda_n^{\text{opt}})$ is not strictly positive. Perturb the constraint (2.11) as

$$\lambda_k \geq \varepsilon, \quad \forall k \in \mathcal{N}$$

for a small strictly positive number ε. Based on the above discussion, the duality gap is zero for the perturbed modified OPF problem and hence Optimization 3 has a rank-one solution, denoted by $W_\varepsilon^{\text{opt}}$ (see Part (i) of Theorem 2). Since $W_\varepsilon^{\text{opt}}$ has a bounded norm (due to the voltage constraints in the OPF problem), this matrix converges to a rank-one solution if ε tends to zero. Hence, Optimization 3 has a rank-one solution for $\varepsilon = 0$ and therefore it follows from Condition (i) of Theorem 2 that the duality gap is zero for the modified OPF problem. So far, it was assumed that the graph of the power system is connected. If not, it means that the OPF problem can be broken down into a number of decoupled OPF problems, each associated with a connected power sub-network. The proof is completed by repeating the aforementioned argument for each small-sized OPF problem. ■

Theorem 3 states that the duality gap becomes zero for the OPF problem if the load constraints are changed from equality to inequality, meaning that the over-satisfaction of the loads is permitted. It is important to study under what conditions the OPF and modified OPF problems have the same solution. This is addressed in the sequel in terms of the duals of these problems.

Lemma 3 *The duals of the OPF problem and the modified OPF problem have the same solution if the vector $(\lambda_1^{opt}, ..., \lambda_n^{opt})$ associated with the original (rather than the modified) OPF problem is nonnegative.*

Proof: As stated in the proof of Theorem 3, the dual of the modified OPF problem is the same as the dual of the OPF problem but with the additional constraints $\lambda_1, ..., \lambda_n \geq 0$. Therefore, if the optimal solution of the dual of the OPF problem satisfies these constraints, its means that the duals of the OPF and modified OPF problems have an identical solution. This completes the proof. ■

The following result can be easily derived from Lemma 3 and the proof of Theorem 3.

Corollary 2 *The duality gap is zero for the OPF problem if* $(\lambda_1^{opt}, ..., \lambda_n^{opt})$ *is nonnegative. Moreover, the sufficient zero-duality-gap condition given in Part (ii) of Theorem 2 holds for the OPF problem if the vector* $(\lambda_1^{opt}, ..., \lambda_n^{opt})$ *is strictly positive and the graph of the power network is strongly connected.*

Assume that the OPF and modified OPF problems have the same solution. Then, the duality gap is zero for the OPF problem, implying that Optimization 3 can solve the OPF problem exactly. However, in order for the Algorithm proposed here (based on Optimization 4) to solve the OPF problem, two conditions must hold. The first one is the connectivity of the power network that holds in reality. The second one requires that every nonnegative aggregate multiplier λ_k^{opt}, $k \in \mathcal{N}$, be strictly positive. This condition holds for a generic OPF problem because λ_k^{opt} being zero implies that the load constraint $P_{L_k} = P_{D_k}$ can be removed from the OPF problem without changing the solution, which signifies that the given value P_{D_k} is not important at all.

A practical power system is often maintained at a normal condition, where if a load bus requests to receive a certain amount of active power or more, the optimal strategy is to deliver exactly the *minimum* amount of power requested. This normal operation results from the fact that generated power is not supposed to be sold at a negative price (note that λ_k^{opt} in practice plays the role of nodal price for the load of bus $k \in \mathcal{N}$). However, an abnormal operation may occur if the physical limits in the OPF problems are very tight and unreasonable so that the OPF problem is over-constrained and each line has a huge amount of power loss on purpose. Under this circumstance, it is possible that the OPF and modified OPF problems achieve different solutions. The next theorem shows that this cannot occur if some of the constraints are removed from the OPF problem to avoid making it over-constrained by choosing inappropriate physical limits.

Theorem 4 *Consider a non-generator bus $k \in \mathcal{N} \backslash \mathcal{G}$. If the voltage constraints (2.1c) and (2.1f) associated with bus k and the flow constraint (2.1e) associated with every line connected to this bus are removed from the OPF problem, then λ_k^{opt} corresponding to this simplified OPF problem is nonnegative.*

Proof: The (k,k) entry of $A(x^{\rm opt}, r^{\rm opt})$, under the assumptions made in the theorem, can be written as

$$\lambda_k^{\rm opt} \left(y_{kk} + \sum_{l \in \mathcal{N}(k)} y_{kl} \right). \tag{2.13}$$

The proof follows from the following facts:

- The expression given in (2.13) must be nonnegative due to the positive semi-definiteness of $A(x^{\rm opt}, r^{\rm opt})$.

- Although y_{kk} might be negative, the overall term $y_{kk} + \sum_{l \in \mathcal{N}(k)} y_{kl}$ is always nonnegative (note that this term corresponds to the (k,k) entry of Y, which is the admittance of a passive network). ■

Consider a non-generator bus k. Since the load is known at this bus, extra constraints related to this bus can make the OPF problem infeasible or over-constrained if the limits are not defined properly. Note that the result of Theorem 4 can be easily generalized to generator buses as well. Hence, the multiplier $\lambda_k^{\rm opt}$ is expected to be nonnegative, something which is needed in Corollary 2 to guarantee the existence of no duality gap for the OPF problem.

In summary, in order to be able to solve the OPF problem in polynomial time, it suffices to have either of the following properties:

- The over-satisfaction of a load is allowed and therefore the modified OPF problem can be solved instead.

- The physical limits of the OPF problem are not chosen in such a way that the power system operates in an abnormal condition, where the active power is offered to a load at a negative price.

2.4.2 General Networks with No Reactive-Load Constraints

As before, consider the modified OPF problem obtained by: (i) replacing the equality constraint (2.9) with the inequality constraint (2.10), and (ii) ignoring the apparent line flow limits and taking only the active line flow limits into account. Assume that the matrix Y is complex, but any arbitrary (positive/negative) amount of reactive power can be injected to each bus $k \in \mathcal{N}$. In this case, the constraints (2.1b) can be ignored. On the other hand, one can write the matrix $A(x^{\text{opt}}, r^{\text{opt}})$ as

$$A(x^{\text{opt}}, r^{\text{opt}}) = \begin{bmatrix} T(x^{\text{opt}}, r^{\text{opt}}) & \bar{T}(x^{\text{opt}}, r^{\text{opt}}) \\ -\bar{T}(x^{\text{opt}}, r^{\text{opt}}) & T(x^{\text{opt}}, r^{\text{opt}}) \end{bmatrix} \tag{2.14}$$

for some real matrices $T(x^{\text{opt}}, r^{\text{opt}}), \bar{T}(x^{\text{opt}}, r^{\text{opt}}) \in \mathbf{R}^{n\times n}$. It can be concluded from the above relation and (2.12) that the matrix $\bar{T}(x^{\text{opt}}, r^{\text{opt}})$ becomes nonzero in the transition from resistive to general networks. Unlike the symmetric matrix $T(x^{\text{opt}}, r^{\text{opt}})$, the matrix $\bar{T}(x^{\text{opt}}, r^{\text{opt}})$ is skew-symmetric and therefore it cannot have only positive entries. This is an impediment to exploiting the Perron-Frobenius theorem. In what follows, we build on Theorem 3 to bypass this issue.

Given a small number $\varepsilon > 0$, consider the Dual OPF problem (Optimization 4) subject to the extra constraints

$$\|x\| \leq \frac{1}{\varepsilon}, \quad \|r\| \leq \frac{1}{\varepsilon}, \quad \varepsilon \leq \lambda_k \leq \frac{1}{\varepsilon}, \quad \forall k \in \mathcal{N} \tag{2.15}$$

where $\|\cdot\|$ is a vector norm. This optimization corresponds to the dual of a perturbed version of the modified OPF problem, which is referred to as *ε-modified OPF problem* here. Note that when ε goes to zero, the solution of this problem approaches that of the original modified OPF problem. To derive the next theorem, with no loss of generality, assume that the resistive part of the power network is strongly connected.

Theorem 5 *Given $\varepsilon > 0$, consider an arbitrary matrix $G \in \mathbf{R}^{n\times n}$, which satisfies all necessary properties for being the real part of the admittance matrix of a power network. There exists an unbound open set $\mathcal{T}_G$ in $\mathbf{R}^{n\times n}$ such that for every $\bar{G} \in \mathcal{T}_G$, the duality gap is zero for the ε-modified OPF problem with $Y = G + \bar{G}i$, regardless of the specific values of the loads and limits in the constraints (2.1).*

Proof: Write Y as $G + \bar{G}\mathrm{i}$, where G is a known matrix and $\bar{G}$ is a matrix variable. Now, the matrix $A(x, r)$ depends on the variable $\bar{G}$, in addition to x and r. To account for this dependence explicitly, we use the notation $A(x, r, \bar{G})$ instead of $A(x, r)$. Let $\mathcal{C}$ denote the set of all triple $(x, r, \bar{G})$ such that

- $A(x, r, \bar{G})$ as well as the matrices given in (2.8b) and (2.8c) are all positive semidefinite.
- The dimension of the null space of $A(x, r, \bar{G})$ is at least 3.
- The relations $x \geq 0$ and (2.15) are satisfied.

The way $\mathcal{C}$ is defined makes it a closed semi-algebraic set (note that the set $\mathcal{C}$ can be described by a number of polynomial inequalities). Recall that $\mathcal{C}$ belongs to the space associated with the variable $(x, r, \bar{G})$. Project this set on the subspace corresponding to its variable $\bar{G}$ and denote the resulting subset as $\mathcal{C}_G$. Define $\mathcal{T}_G$ as the complement of $\mathcal{C}_G$. Notice that the sufficient zero-duality-gap condition given in Theorem 2 is satisfied for the ε-modified OPF problem with $Y = G + \bar{G}\mathrm{i}$ as long as $\bar{G} \in \mathcal{T}_G$. The proof of this theorem follows from the facts given below:

- Since $\mathcal{C}$ is closed and bounded (due to the relations given in (2.15)), the projection set $\mathcal{C}_G$ is closed as well. Therefore, the complement of $\mathcal{C}_G$, i.e., $\mathcal{T}_G$, is an open set.
- Consider a diagonal matrix $\bar{G}$. It can be verified that the matrix $\bar{T}(x, r, \bar{G})$ is zero in this case, Thus, the matrix $A(x, r, \bar{G})$ has the block-diagonal structure (2.12), meaning that the non-resistive part of the network has disappeared. Hence, it can be inferred from the proof of Theorem 3 that the duality gap is zero in this case. As a result, $\bar{G}$ must belong to $\mathcal{T}_G$.
- The set of diagonal matrices is unbounded. ■

As done in the preceding subsection, the OPF and modified OPF problems are expected to have the same solution; otherwise the power system may not work in a normal condition. Note that the condition provided in Theorem 4 to guarantee the same solution for the OPF and modified OPF problems still holds for a general network with no constraints on reactive loads. In this subsection, we perturbed the modified OPF problem and defined an ε-modified OPF problem. Theorem 5 states that for every $\mathrm{Re}\{Y\}$ (that could be arbitrarily large or small), there exists an open, unbounded region for $\mathrm{Im}\{Y\}$ such that the algorithm proposed in this chapter can find a global optimum of the ε-modified OPF problem with $Y = \text{Re}\{Y\} + \text{Im}\{Y\}\text{i}$ in polynomial time. The importance of this result is as follows: **when the duality gap is zero for a topology Y, then the OPF problem corresponding to every possible load profiles and physical limits can be convexified.**

2.4.3 General Networks

In this part, we combine the ideas presented in the last two subsections to study the OPF problem associated with a general network. For simplicity in the presentation, remove the constraints $|S_{lm}| \leq S_{lm}^{\max}$ (where $(l,m) \in \mathcal{L}$), because of its similarity to the constraint $|P_{lm}| \leq P_{lm}^{\max}$. Consider the matrix $A(x^{\text{opt}}, r^{\text{opt}})$, which can be expressed as

$$A(x^{\text{opt}}, r^{\text{opt}}) = \begin{bmatrix} T(x^{\text{opt}}, r^{\text{opt}}) & \bar{T}(x^{\text{opt}}, r^{\text{opt}}) \\ -\bar{T}(x^{\text{opt}}, r^{\text{opt}}) & T(x^{\text{opt}}, r^{\text{opt}}) \end{bmatrix}$$

where $T(x^{\text{opt}}, r^{\text{opt}})$ is symmetric and $\bar{T}(x^{\text{opt}}, r^{\text{opt}})$ is skew-symmetric. As observed in both the resistive case and the general case with no reactive-load constraints, the duality gap can be pushed towards zero if the off-diagonal entries of $T(x^{\text{opt}}, r^{\text{opt}})$ are all non-positive. In what follows, we first study the sign structure of $T(x^{\text{opt}}, r^{\text{opt}})$.

As carried out in Section 2.4.1, define $P_{D_k} + Q_{D_k}\text{i}$ as the apparent power requested by load $k \in \mathcal{N}$ and $P_{L_k} + Q_{L_k}\text{i}$ as the apparent power delivered to load $k \in \mathcal{N}$. In the original OPF problem, the equalities

$$P_{L_k} = P_{D_k}, \quad Q_{L_k} = Q_{D_k}, \quad \forall k \in \mathcal{N} \tag{2.16}$$

must hold. If these equalities are replaced by the inequalities

$$P_{L_k} \geq P_{D_k}, \quad Q_{L_k} \geq Q_{D_k}, \quad \forall k \in \mathcal{N} \tag{2.17}$$

then the optimal solutions λ_k^{opt} and γ_k^{opt} corresponding to the dual of the modified OPF problem will both become nonnegative. On the other hand, the $(k,l) \in \mathcal{L}$ entry of $T(x^{\text{opt}}, r^{\text{opt}})$

can be obtained as

$$\begin{aligned} T_{kl}(x^{\text{opt}}, r^{\text{opt}}) = &- \frac{\text{Re}\{y_{kl}\}}{2} \left(\lambda_{kl}^{\text{opt}} + \lambda_{kl}^{\text{opt}} + \lambda_{k}^{\text{opt}} + \lambda_{l}^{\text{opt}} \right) \\ &+ \frac{\text{Im}\{y_{kl}\}}{2} \left(\gamma_{k}^{\text{opt}} + \gamma_{l}^{\text{opt}} \right) - \mu_{kl}^{\text{opt}} - \mu_{kl}^{\text{opt}}. \end{aligned} \tag{2.18}$$

With no loss of generality, assume that there exists no phase shifting transformer in the power system (for the analysis presented next, one may need to replace every phase shifting transformer with the model proposed in [60]). Due to the particular models of transmission lines and transformers as well as the non-negativity of resistance and capacitance, the matrix Y has the following two properties:

P1) The off-diagonal entries of the real part of Y are non-positive.

P2) The off-diagonal entries of the imaginary part of Y are nonnegative.

It follows from these properties and the relation (2.18) that the off-diagonal entries of $T(x^{\text{opt}}, r^{\text{opt}})$ are non-positive if $\lambda_k^{\text{opt}}, \gamma_k^{\text{opt}} \geq 0$, $\forall k \in \mathcal{N}$, or equivalently if the equality load constraints (2.16) are replaced by the inequality load constraints (2.17). Unlike $\lambda_1^{\text{opt}}, ..., , \lambda_n^{\text{opt}}$ that are expected to be all nonnegative, a few of $\gamma_1^{\text{opt}}, ..., , \gamma_n^{\text{opt}}$ might become negative. Indeed, it is known that the injection of a negative reactive power to a bus might reduce the optimal generation cost, especially when there exists a large capacitor bank at the same bus.

Hence, the sufficient condition $\lambda_k^{\text{opt}}, \gamma_k^{\text{opt}} \geq 0$, $\forall k \in \mathcal{N}$, for guaranteeing a nice sign structure on $T(x^{\text{opt}}, r^{\text{opt}})$ does not always hold. Now, we wish to study a less conservative sufficient condition here. It follows from (2.18) that the off-diagonal entries of $T(x^{\text{opt}}, r^{\text{opt}})$ are non-positive if

$$\frac{\text{Re}\{y_{kl}\}}{2} \left(\lambda_{kl}^{\text{opt}} + \lambda_{kl}^{\text{opt}} + \lambda_{k}^{\text{opt}} + \lambda_{l}^{\text{opt}} \right) - \frac{\text{Im}\{y_{kl}\}}{2} \left(\gamma_{k}^{\text{opt}} + \gamma_{l}^{\text{opt}} \right) \geq 0 \tag{2.19}$$

for every $(k, l) \in \mathcal{L}$. This condition is satisfied for IEEE benchmark systems. The interpretation of this condition for a single $(k, l) \in \mathcal{L}$ is as follows:

- Define a modified OPF with the following active/reactive load constraints

$$\begin{aligned} P_{L_m} &= P_{D_m}, \quad Q_{L_m} = Q_{D_m}, \quad \forall m \in \mathcal{N} \backslash \{k, l\} \\ P_{L_m} &\geq P_{D_m}, \quad Q_{L_m} \geq Q_{D_m}, \quad \forall m \in \{k, l\} \end{aligned}$$

where the load over-satisfaction at buses k and l must obey the relations

$$\begin{aligned}
P_{L_k} - P_{D_k} = P_{L_l} - P_{D_l} = \tau \times \mathrm{Re}\{y_{kl}\} \\
Q_{L_k} - Q_{D_k} = Q_{L_l} - Q_{D_l} = \tau \times \mathrm{Im}\{-y_{kl}\} \\
\max\{P_{lm}, P_{ml}\} \leq P_{lm}^{\max} - \tau \times \mathrm{Re}\{y_{kl}\}
\end{aligned}$$

for some nonnegative number τ.

- The dual of the above modified OPF problem can be obtained from the Dual OPF by incorporating the extra constraint (2.19).
- If optimal τ becomes zero, then the OPF and modified OPF problems will have the same solution, meaning that the (k, l) entry of $T(x^{\mathrm{opt}}, r^{\mathrm{opt}})$ is non-positive.

Notice that the modified OPF problem defined above allows the reactive load at bus k to be over-satisfied, but enforces extra consumption of both active and reactive loads at buses k, l and reduces the maximum flow limit on line (k, l). Therefore, it is very likely to obtain $\tau^{\mathrm{opt}} = 0$ due to these penalties for load over-satisfaction (note that the imposed over-satisfaction of active load often leads to more power loss). The above modified OPF problem is defined to ensure the non-positivity of only the (k, l) entry of $T(x^{\mathrm{opt}}, r^{\mathrm{opt}})$. A similar modified OPF can be defined corresponding to all off-diagonal entries of $T(x^{\mathrm{opt}}, r^{\mathrm{opt}})$.

So far, the reason why the off-diagonal entries of $T(x^{\mathrm{opt}}, r^{\mathrm{opt}})$ are expected to be non-positive is investigated. Having assumed the presence of this sign structure on $T(x^{\mathrm{opt}}, r^{\mathrm{opt}})$, consider the matrix

$$\begin{bmatrix} T(x^{\mathrm{opt}}, r^{\mathrm{opt}}) & \bar{T}(x^{\mathrm{opt}}, r^{\mathrm{opt}}) \times \omega \\ -\bar{T}(x^{\mathrm{opt}}, r^{\mathrm{opt}}) \times \omega & T(x^{\mathrm{opt}}, r^{\mathrm{opt}}) \end{bmatrix} \tag{2.20}$$

for a given real number ω. As argued in the proof of Theorem 3, the smallest eigenvalue of the above matrix is repeated twice when $\omega = 0$. Hence, there exists an interval $[0, \omega^{\max}]$ (where $\omega^{\max} > 0$) such that the smallest eigenvalue of the matrix (2.20) is repeated twice for every ω belonging to this interval. Now, note that if $\omega^{\max} > 1$, then the zero-duality-gap condition given in Theorem 2 is satisfied. This happens whenever $\bar{T}(x^{\mathrm{opt}}, r^{\mathrm{opt}})$ is sufficiently smaller than $T(x^{\mathrm{opt}}, r^{\mathrm{opt}})$ with respect to a suitable measure on their entries. As can be justified intuitively and verified in simulations, this is the case for practical systems operating at a normal condition, including the IEEE test systems.

It is noteworthy that Theorem 5 can be generalized to a general network (with arbitrary constraints) to deduce that there exists an unbounded open set for Y such that the ε-modified OPF problem has zero duality gap with respect to all network topologies Y in that region.

2.4.4 Power Loss Minimization

In this subsection, we consider the loss minimization problem, as an important special case of the OPF problem. This corresponds to the assumption $f_k(P_{G_k}) = P_{G_k}$ for every $k \in \mathcal{G}$. Most of the results to be presented here can be extended to a general OPF problem. With no loss of generality, assume that $\text{Re}\{Y\}$ has exactly one zero eigenvalue, implying that (i) the graph associated with the resistive part of the network is strongly connected [31], and (ii) every load modeled as a shunt admittance has no resistive part. Notice that the power loss in a power system can be reduced by either increasing the voltage limits or decreasing the resistance of transmission lines. The next lemma investigates an ideal case where the power loss is zero.

Theorem 6 *If the active power losses in the transmission lines were zero at optimality, then there would exist an optimal dual point (x^{opt}, r^{opt}) satisfying the relations*

$$r^{opt} = 0, \quad \lambda_k^{opt} = 1, \quad \gamma_k^{opt} = \mu_k^{opt} = \lambda_{lm}^{opt} = \mu_{lm}^{opt} = 0$$

for every $k \in \mathcal{N}$ and $(l,m) \in \mathcal{L}$. Moreover, this dual solution satisfies the zero-duality-gap condition (ii) given in Theorem 2.

Proof: Consider a specific point (x, r) defined as $r = 0$ and

$$\begin{aligned} &\underline{\lambda}_k = \underline{\gamma}_k = \bar{\gamma}_k = \underline{\mu}_k = \bar{\mu}_k = \lambda_{lm} = \mu_{lm} = 0 \\ &\bar{\lambda}_k := \begin{cases} 0 & \text{if } k \in \mathcal{G} \\ 1 & \text{otherwise} \end{cases} \end{aligned}$$

for all $k \in \mathcal{N}$ and $(l,m) \in \mathcal{L}$. It is straightforward to verify that $h(x,r) = \sum_{k\in\mathcal{N}} P_{D_k}$. On the other hand, since the OPF problem is feasible and the total power loss is zero, the optimal objective value of the OPF problem is equal to the total demand. This shows that the objective value of the dual problem at (x, r) is identical to the optimal value of the OPF problem. Hence, to prove that (x, r) is a dual solution, it suffices to show that (x, r) is a feasible point of this optimization problem. To this end, it can be verified that

$$\lambda_k = 1, \quad \gamma_k = 0, \quad \mu_k = 0, \quad \forall k \in \mathcal{N}$$

and hence

$$A(x, r) = \begin{bmatrix} \mathrm{Re}\{Y\} & 0 \\ 0 & \mathrm{Re}\{Y\} \end{bmatrix}.$$

Therefore, $A(x, r)$ is positive semidefinite and has a zero eigenvalue of multiplicity 2. This means that $(x^{\mathrm{opt}}, r^{\mathrm{opt}}) = (x, r)$ is indeed a maximizer of Optimization 4 for which the sufficient zero-duality-gap condition (ii) given in Theorem 2 holds. ■

Theorem 6 studies a special type of the OPF problem in an ideal case of no power loss, and presents an optimal dual solution explicitly from which it can be seen that the duality gap is zero. However, active power loss is nonzero, but small, in practice. In that case, if the Lagrange multipliers λ_k^{opt}, γ_k^{opt} and μ_k^{opt} are treated as nodal prices for active and reactive powers as well as voltage levels, it can be argued that the optimal point in a lossy case is likely to be close enough to the dual solution given in Theorem 6 so that the matrix $A(x^{\mathrm{opt}}, r^{\mathrm{opt}})$ will still have two zero eigenvalues. In other words, it is expected that a small power loss in transmission lines does not create a duality gap.

2.5 Power System Examples

This section illustrates our results through two examples. Example 1 uses the IEEE benchmark systems archived at [109] to show the practicality of our result. Since the systems analyzed in Example 1 are so large that the specific values of the optimal solution cannot be provided here, some smaller examples are analyzed in Example 2 with more details.

The results of this section are attained using the following software tools:

- The MATLAB-based toolbox “YALMIP” (together with the solver “SEDUMI”) is used to solve the Dual OPF problem (i.e., Optimization 4), which is an SDP problem [68].

- The software toolbox “MATPOWER” is used to solve the OPF problem in Example 1 for the sake of comparison. The data for the IEEE benchmark systems analyzed in this example is extracted from the library of this toolbox [118].

- The software toolbox "PSAT" is used to draw and analyze the power networks given in Example 2 [71].

2.5.1 Example 1: IEEE Benchmark Systems

Consider the OPF problems associated with IEEE systems with 14, 30, 57, 118, and 300 buses, where

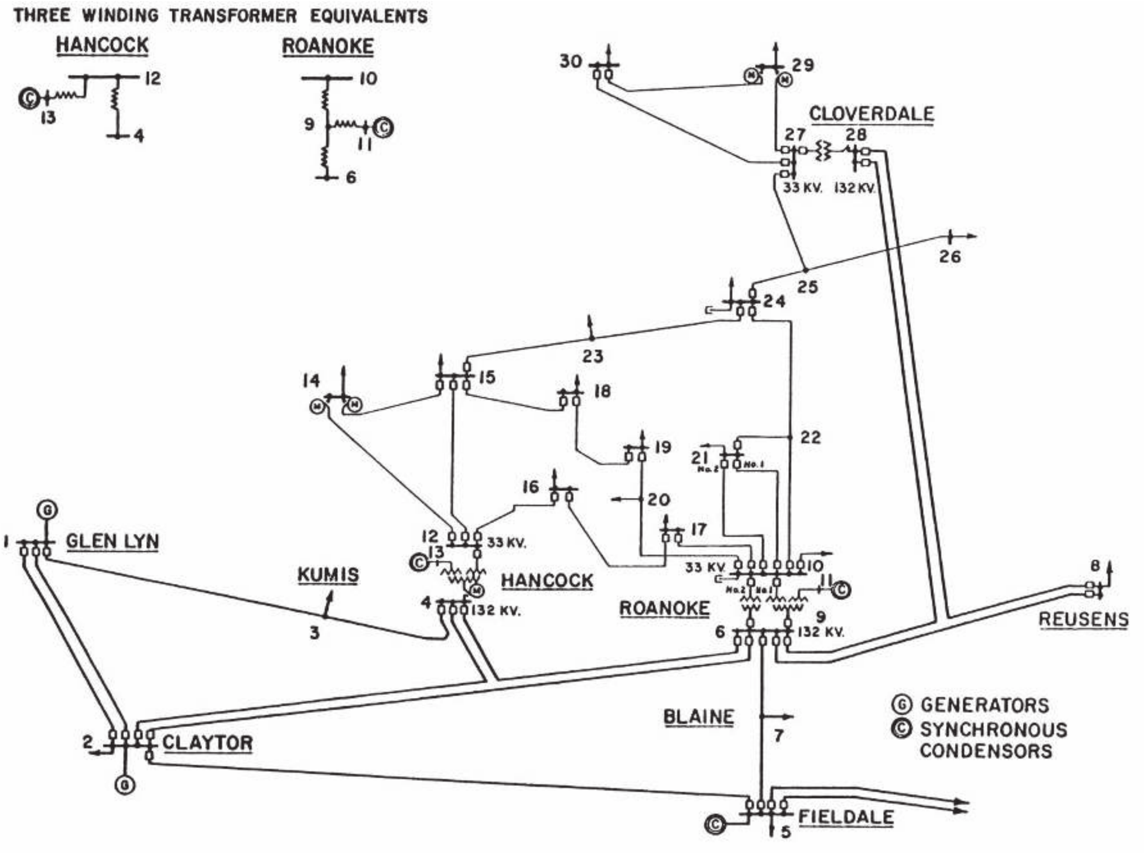
- There are constraints on the voltage magnitude, active power and reactive power at every bus as well as the apparent power at every line.
- The objective function is either the total generation cost or the power loss.

In simulations, we observed that the necessary and sufficient zero-duality-gap condition (i) given in Theorem 2 is always satisfied for all these systems. However, since the main algorithm proposed here is based on the sufficient zero-duality-gap condition (ii) delineated in Theorem 2, we studied this condition for IEEE systems and noticed that the condition is always satisfied after a small perturbation of Y, as discussed below. Due to space restrictions, the details will be provided only in one case: the loss minimization for the IEEE 30-bus system.

Consider the OPF problem for the IEEE 30-bus system, where the objective is to minimize the total power generated by the generators. When Optimization 4 is solved, the four smallest eigenvalues of the matrix

$$A(x^{\text{opt}}, r^{\text{opt}}) = \begin{bmatrix} T(x^{\text{opt}}, r^{\text{opt}}) & \bar{T}(x^{\text{opt}}, r^{\text{opt}}) \\ -\bar{T}(x^{\text{opt}}, r^{\text{opt}}) & T(x^{\text{opt}}, r^{\text{opt}}) \end{bmatrix}$$

would be obtained as $0, 0, 0, 0$. Since the number of zero eigenvalues is 4, condition (ii) in Theorem 2 is violated. To explore the underlying reason, consider the circuit of this power system that is depicted in Figure 2.2. The circuit is composed of three regions connected to each other via some transformers. This implies that if each line of the circuit is replaced by its resistive part, the resulting resistive graph will not be connected (since the lines with transformers are assumed to have no resistive parts). Thus, the graph induced by $\text{Re}\{Y\}$ is not strongly connected. Although this does not create a nonzero duality gap,



(a)

Figure 2.2: The circuit of the IEEE 30-bus system taken from [109]

it causes our sufficient duality-gap condition to be violated (see Corollary 2). This is an issue with all the IEEE benchmark systems. This can be easily fixed by adding a little resistance to each transformer, say on the order of 10^{-5} (per unit). After this modification to the real part of Y, the four smallest eigenvalues of the matrix $A(x^{\text{opt}}, r^{\text{opt}})$ turn out to be $0, 0, 0.0053, 0.0053$; i.e., the zero eigenvalues resulting from the non-connectivity of the resistive graph have disappeared. Now, condition (ii) in Theorem 2 is satisfied and therefore the vector of optimal voltages can be recovered using the algorithm described after Theorem 2.

To illustrate the discussions made in Section 2.4, we note that (for every $k \in \mathcal{N}$)

$$\lambda_k^{\text{opt}} \in [1, 1.1466], \quad \gamma_k^{\text{opt}} \in [-0.0062, 0.1443], \quad \mu_k^{\text{opt}} \in [-0.0216, 0].$$

Hence

- λ_k^{opt}'s are all positive and around 1.
- γ_k^{opt}'s are all but one nonnegative, and are around 0.
- μ_k^{opt}'s are all very close to 0.

Moreover, the maximum absolute values of the entries of $\bar{T}(x^{\text{opt}}, r^{\text{opt}})$ is 0.1844, whereas the average absolute values of the nonzero entries of $T(x^{\text{opt}}, r^{\text{opt}})$ is 4.2583. This confirms the claim in Section 2.4.3 that the matrix $\bar{T}(x^{\text{opt}}, r^{\text{opt}})$ is expected to be negligible compared to $T(x^{\text{opt}}, r^{\text{opt}})$.

The computation on the IEEE benchmark examples were all finished in a few seconds and the number of iterations for each example was between 5 and 20. Note that although Optimization 4 is convex and there is no convergence problem regardless of what initial point is used, the number of iterations needed to converge mainly depends on the choice of starting point. It is worth mentioning that when different algorithms implemented in Matpower were applied to these systems, some of the constraints are violated at the optimal point probably due to the relatively large-scale and non-convex nature of the OPF problem. However, no constraint violation has occurred by solving the dual of the OPF problem due to its convexity.

Table 2.1: Parameters of the systems given in Figure 2.3

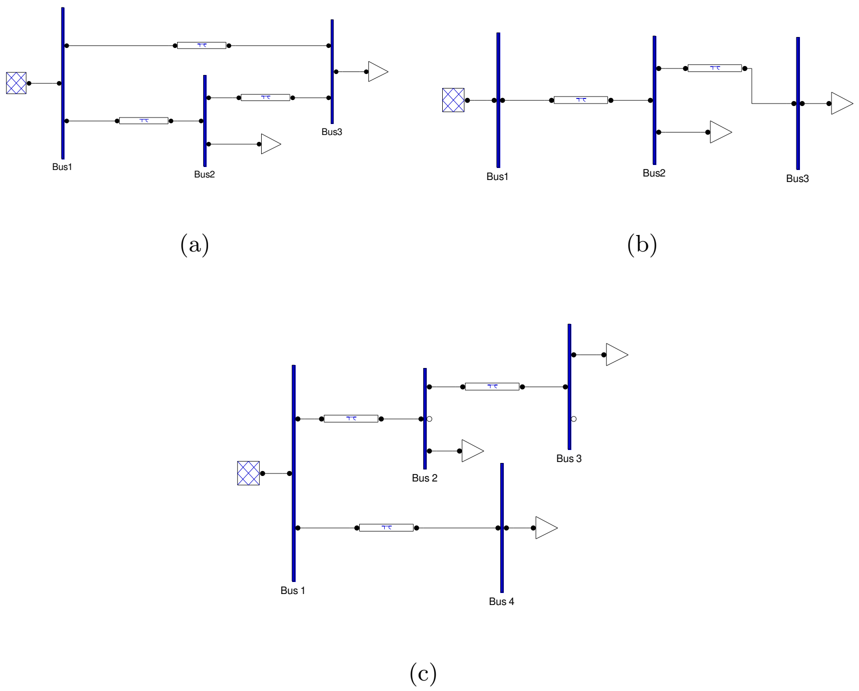
Parameters	System 1	System 2	System 3
$\bar{z}_{12}$	$0.05 + 0.25$i	$0.1 + 0.5$i	$0.10 + 0.1$i
$\bar{z}_{13}$	$0.04 + 0.40$i	None	None
$\bar{z}_{23}$	$0.02 + 0.10$i	$0.02 + 0.20$i	$0.01 + 0.1$i
$\bar{z}_{14}$	None	None	$0.01 + 0.2$i
$\bar{y}_{12}$	0.12i	0.04i	0.12i
$\bar{y}_{13}$	0.10i	None	None
$\bar{y}_{23}$	0.04i	0.04i	0.04i
$\bar{y}_{14}$	None	None	0.04i

2.5.2 Example 2: Small Systems

The IEEE test systems in the previous example operate in a normal condition when the optimal bus voltages are close to each other in both magnitude and phase. This example illustrates that the sufficient zero-duality-gap condition (ii) given in Theorem 2 is satisfied even in the absence of such a normal operation. Consider three distributed power systems, referred to as Systems 1, 2, and 3, depicted in Figure 2.3. Note that Systems 2 and 3 are radial, while System 1 has a loop. The detailed specifications of these systems are provided in Table 2.1 in per unit for the voltage rating 400 kV and the power rating 100 MVA, in which $\bar{z}_{lm}$ and $\bar{y}_{lm}$ denote the series impedance and the shunt admittance of the Π model of the transmission line connecting buses $l, m \in \{1, 2, 3, 4\}$. The goal is to minimize the active power injected at slack bus 1 while satisfying the constraints given in Table 2.2.

Optimization 4 is solved for each of these systems, and it is observed that the zero-duality-gap condition derived in this work always holds. A globally optimal solution of the OPF problem recovered from the solution of Optimization 4 is provided in Table 2.3 (P_{loss} and Q_{loss} in the table represent the total active and reactive power losses, respectively). It is interesting to note that although different buses have very disparate voltage magnitudes and phases, the duality gap is still zero. The optimal solution of Optimization 4 is summarized in Table 2.4 to demonstrate that the Lagrange multipliers corresponding to active and reactive power constraints are positive.

As another scenario, let the desired voltage magnitude at the slack bus of System 1 be changed from 1.05 to 1. It can be verified that the optimal value of Optimization 4 becomes $+\infty$, which simply implies that the corresponding OPF problem is infeasible.



(a)

(b)

(c)

Figure 2.3: Figures (a), (b), and (c) depict Systems 1, 2 and 3 studied in Example 2, respectively.

Table 2.2: Constraints to be satisfied for the systems given in Figure 2.3

Constraints	System 1	System 2	System 3
$P_{D_2} + Q_{D_2}\mathrm{i}$	$0.95 + 0.4\mathrm{i}$	$0.7 + 0.02\mathrm{i}$	$0.9 + 0.02\mathrm{i}$
$P_{D_3} + Q_{D_3}\mathrm{i}$	$0.9 + 0.6\mathrm{i}$	$0.65 + 0.02\mathrm{i}$	$0.6 + 0.02\mathrm{i}$
$P_{D_4} + Q_{D_4}\mathrm{i}$	None	None	$0.9 + 0.02\mathrm{i}$
$V_1^{\max}$	1.05	1.4	1

Table 2.3: Parameters of the OPF problem recovered from the solution of Optimization 2

Recovered Parameters	System 1	System 2	System 3
V_1^{opt}	$1.05\angle 0°$	$1.4\angle 0°$	$1\angle 0°$
V_2^{opt}	$0.71\angle{-20.11}°$	$1.10\angle{-25.73}°$	$0.78\angle{-10.58}°$
V_3^{opt}	$0.68\angle{-21.94}°$	$1.08\angle{-31.96}°$	$0.76\angle{-16.31}°$
V_4^{opt}	None	None	$0.95\angle{-10.82}°$
$P_{\text{loss}}^{\text{opt}}$	0.2193	0.1588	0.3877
$Q_{\text{loss}}^{\text{opt}}$	1.2944	0.7744	0.5343

Table 2.4: Lagrange multipliers obtained by solving Optimization 2 for the systems given in Figure 2.3.

Lagrange Multipliers	System 1	System 2	System 3
$\lambda_2^{\rm opt}$	1.3809	1.4028	1.7176
$\lambda_3^{\rm opt}$	1.4155	1.4917	1.7900
$\lambda_4^{\rm opt}$	None	None	1.0207
$\gamma_2^{\rm opt}$	0.4391	0.2508	0.1764
$\gamma_3^{\rm opt}$	0.4955	0.2633	0.1858
$\gamma_4^{\rm opt}$	None	None	0.0061
$\mu_1^{\rm opt}$	0.0005	0.0001	0.0005

We repeated several hundred times this example by randomly choosing the parameters of the systems given in Figure 2.3 over a wide range of values. In all these trials, the algorithm prescribed in Section 2.3 always found a globally optimal solution of the OPF problem or detected its infeasibility.

2.6 Summary

This chapter is concerned with the optimal power flow (OPF) problem that has been studied for about half a century and is notorious for its high nonconvexity. We have derived the dual of a reformulated OPF problem as a convex (SDP) optimization, which can be solved efficiently in polynomial time. We have provided a necessary and sufficient condition under which the duality gap is zero and hence a globally optimal solution to the OPF problem can be recovered from a dual optimal solution. This condition is satisfied for the IEEE benchmark systems with 14, 30, 57, 118, and 300 buses. Since this condition is hard to study, a sufficient zero-duality-gap condition is also proposed. We justify why this sufficient condition might hold widely in practice. The main underlying reasons for zero duality gap are (i) the particular modeling of transmission lines and transformers, and (ii) the non-negativity of physical quantities such as resistance and inductance.

As expected and already reported in [4], local-search algorithms converge faster than SDP algorithms for solving an OPF problem. However, the SDP problem derived here can be useful for addressing many problems such as: (i) finding a globally optimal solution, (ii) verifying whether a locally optimal solution is globally optimal, (iii) solving emerging optimization problems in smart grids where the existing local-search algorithms may not work well [60], and (iv) identifying the number of solutions of a power flow problem. Note that the current SDP solvers cannot handle OPF problems with several thousand buses efficiently. However, we have observed that those SDP problems can be reduced to second-order-cone programs, which can be solved in less than a minute for OPF problems with as many as 10,000 buses. The details of this result and some other by-products of the convexification of the OPF problem are currently under study.

2.7 Appendix

2.7.1 LMI and SDP Optimization Problems

The area of convex optimization has seen remarkable progress in the past two decades, particularly in linear matrix inequalities (LMIs) and semidefinite programming (SDP) where the goal is to minimize a linear function subject to some LMIs [13, 24]. The book [10] describes several difficult control problems that can be cast as LMI/SDP problems and then solved efficiently. The recent advances in this field have been successfully applied to different problems in other areas, e.g., circuit and communications [110, 62]. A powerful property in semidefinite programming is that the dual of an SDP optimization problem is again an SDP problem and, moreover, strong duality often holds [24].

Given the scalar variables $x_1, ..., x_n$, consider the problem of minimizing

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{2.21}$$

subject to the LMI constraint

$$A_0 + A_1x_1 + \cdots + A_nx_n \preceq 0. \tag{2.22}$$

where $a_1, ..., a_n$ are given real numbers and $A_0, ..., A_n$ are given symmetric matrices in $\mathbf{R}^{n_0 \times n_0}$, for some natural number n_0. Notice that the objective of the above optimization problem is a linear scalar function, and its constraint is an LMI. The above optimization problem is referred to as *an SDP problem*, which belongs to the category of convex optimization problems that can be solved efficiently. To write the Lagrangian for the above optimization problem, a Lagrange multiplier should be introduced for the inequality (2.22). In light of the generalized Lagrangian theory, the multiplier associated with the inequality (2.22) is a symmetric matrix W in $\mathbf{R}^{n_0 \times n_0}$ that must be positive semidefinite. The corresponding Lagrangian will be as follows:

$$\sum_{k=1}^{n} a_k x_k + \mathrm{Tr}\left\{ W \left(A_0 + \sum_{k=1}^{n} A_k x_k \right) \right\}$$

Note that the trace operator performs the multiplication between the expression in the constraint (2.22) and its associated Lagrange multiplier. Minimizing the above Lagrangian over $x_1, ..., x_n$ and then maximizing the resulting term over $W \succeq 0$ lead to the optimization problem of maximizing

$$\mathrm{Tr}\{WA_0\}$$

subject to the constraints

$$\mathrm{Tr}\{WA_k\} + a_k = 0, \quad k = 1, 2, ..., n$$

for a symmetric matrix variable $W \succeq 0$. This optimization problem is the dual of the initial optimization problem formulated in (2.21) and (2.22). If some mild conditions (such as Slater's conditions) hold, then the duality gap between the solutions of these two optimization problems becomes zero, meaning that the optimal objective values obtained by these problems will be identical. In this case, it is said that "strong duality" holds; otherwise, only "weak duality" holds in which case the optimal value of the dual problem is only a lower bound on the optimal value of the original problem. One can refer to [13] and [24] for detailed discussions on LMI and SDP problems.

2.7.2 NP-Hardness of the OPF Problem

Consider two extremely special (artificial) instances of the OPF problem in the sequel:

- **Case 1:** This case corresponds to the situation where $\mathcal{G} = \mathcal{N}$ and

$$\begin{aligned} & f_k(P_{G_k}) = P_{G_k}, && \forall k \in \mathcal{G} \\ & V_k^{\min} = V_k^{\max} = 1, && \forall k \in \mathcal{N} \\ & P_k^{\min} = Q_k^{\min} = -\infty, && \forall k \in \mathcal{G} \\ & P_k^{\max} = Q_k^{\max} = +\infty, && \forall k \in \mathcal{G} \\ & S_{lm}^{\max} = P_{lm}^{\max} = \Delta V_{lm}^{\max} = \infty, && \forall (l,m) \in \mathcal{L}. \end{aligned}$$

The above setting makes the power balance equations together with the constraints (2.1a), (2.1b), (2.1d), (2.1e) and (2.1f) all disappear. It is straightforward to verify that the OPF problem reduces to

$$\begin{aligned} & \min_{\mathbf{V}} \left(\operatorname{Re}\{\mathbf{V}^* Y \mathbf{V}\} + \sum_{k \in \mathcal{N}} P_{D_k} \right) \\ & s.t. \quad |V_k| = 1, \quad \forall k \in \mathcal{N}. \end{aligned} \tag{2.23}$$

Note that if the lower limit $P_k^{\min}$ chosen as $-\infty$ is not allowed to be less than zero, one can choose P_{D_k} sufficiently large so that the OPF problem again turns into the above optimization problem. Observe that the feasibility region of this OPF problem in the space of $\mathbf{V}$ is a connected, but nonconvex, set (the nonconvexity comes from the fact that this region encloses the origin but does not contain it).

- **Case 2:** This case is obtained from Case 1 by including the extra assumption $\operatorname{Im}\{Y\} = 0$ and changing the limits $Q_k^{\min} = -\infty$ and $Q_k^{\max} = +\infty$ to $Q_k^{\min} = Q_k^{\max} = 0$ for every $k \in \mathcal{G}$. With no loss of generality, suppose that the voltage angle at bus 1 is equal to 0. Then, the OPF problem can be written as

$$\begin{aligned} & \min_{\mathbf{V}} \left(\mathbf{V}^* Y \mathbf{V} + \sum_{k \in \mathcal{N}} P_{D_k} \right) \\ & s.t. \quad V_k \in \{-1, 1\}, \quad \forall k \in \mathcal{N}. \end{aligned} \tag{2.24}$$

The feasibility region of this problem is a discrete set with an exponential number of points in terms of n.

The optimization problems given in (2.23) and (2.24) are both NP-hard [115]. Hence, the OPF problem is NP-hard as well, due to its special (artificial) Cases 1 and 2 being NP-hard. Note that although the NP-harness of the OPF problem was proved here by focusing on the voltage constraints, one can come to the same conclusion by only considering the active or reactive constraints. Indeed, Lemma 1 presented later in this work shows that these constraints introduce indefinite quadratic constraints, which again make the OPF problem NP-hard [115].

2.7.3 Proofs

In this subsection, we prove Lemmas 1–2, Theorems 1–2 and Corollary 1.

Proof of Lemma 1: In order to prove (2.2a), one can write:

$$\begin{aligned} P_{k,\text{inj}} &= \text{Re}\{V_k I_k^*\} = \text{Re}\{\mathbf{V}^* e_k e_k^* \mathbf{I}\} = \text{Re}\{\mathbf{V}^* Y_k \mathbf{V}\} \\ &= \mathbf{X}^T \begin{bmatrix} \text{Re}\{Y_k\} & -\text{Im}\{Y_k\} \\ \text{Im}\{Y_k\} & \text{Re}\{Y_k\} \end{bmatrix} \mathbf{X} \\ &= \frac{1}{2}\mathbf{X}^T \begin{bmatrix} \text{Re}\{Y_k + Y_k^T\} & \text{Im}\{Y_k^T - Y_k\} \\ \text{Im}\{Y_k - Y_k^T\} & \text{Re}\{Y_k + Y_k^T\} \end{bmatrix} \mathbf{X} \\ &= \mathbf{X}^T \mathbf{Y}_k \mathbf{X} = \text{Tr}\left\{\mathbf{Y}_k \mathbf{X}\mathbf{X}^T\right\}. \end{aligned}$$

The inequality (2.2b) can be derived similarly. On the other hand, the technique used above can be exploited to show that

$$\begin{aligned} S_{lm}^* &= V_l^* \left(V_l \bar{y}_{lm}\right) + V_l^* \left(V_l - V_m\right) y_{lm} = \mathbf{V} Y_{lm} \mathbf{V}^* \\ &= \text{Tr}\left\{\mathbf{Y}_{lm} \mathbf{X}\mathbf{X}^T\right\} - \text{Tr}\left\{\bar{\mathbf{Y}}_{lm} \mathbf{X}\mathbf{X}^T\right\} \text{i}. \end{aligned}$$

Inequalities (2.2c) and (2.2d) follow immediately from the above equality. The remaining inequalities in (2.2) can be proved similarly. ■

Proof of Lemma 2: Assume that W^{opt} is a rank-one solution of Optimization 3. Write this matrix as $\mathbf{X}^{\text{opt}}(\mathbf{X}^{\text{opt}})^T$ for some vector $\mathbf{X}^{\text{opt}}$, and define $\mathbf{X}_1^{\text{opt}}$ and $\mathbf{X}_2^{\text{opt}}$ in such a way that $\mathbf{X}^{\text{opt}} = \begin{bmatrix} (\mathbf{X}_1^{\text{opt}})^T & (\mathbf{X}_2^{\text{opt}})^T \end{bmatrix}^T$. It can be verified that the matrix

$$\frac{1}{2}\mathbf{X}^{\text{opt}}(\mathbf{X}^{\text{opt}})^T + \frac{1}{2}\begin{bmatrix} \mathbf{X}_1^{\text{opt}}\omega_1 - \mathbf{X}_2^{\text{opt}}\omega_2 \\ \mathbf{X}_1^{\text{opt}}\omega_2 + \mathbf{X}_2^{\text{opt}}\omega_1 \end{bmatrix}\begin{bmatrix} \mathbf{X}_1^{\text{opt}}\omega_1 - \mathbf{X}_2^{\text{opt}}\omega_2 \\ \mathbf{X}_1^{\text{opt}}\omega_2 + \mathbf{X}_2^{\text{opt}}\omega_1 \end{bmatrix}^T$$

is a solution of Optimization 3 for every real numbers ω_1 and ω_2 such that $\omega_1^2 + \omega_2^2 = 1$. The proof is completed by noting that the above matrix has rank 2 for generic values of (ω_1, ω_2). ■

Proof of Part (i) of Theorem 1: Consider the Lagrange multipliers introduced before Optimization 4 with the only difference that the multiplier

$$\begin{bmatrix} 1 & r_k^1 \\ r_k^1 & r_k^2 \end{bmatrix}$$

given in (2.7) should be replaced by a general matrix

$$\begin{bmatrix} r_k^0 & r_k^1 \\ r_k^1 & r_k^2 \end{bmatrix}$$

(indeed, we do not yet know that $r_k^0 = 1$.) The Lagrangian for Optimization 1 can be written as (after some simplifications)

$$\text{Tr}\left\{A(x,r)\mathbf{X}\mathbf{X}^T\right\} + h(x,r) + \sum_{k\in\mathcal{G}}(1 - r_k^0)\alpha_k.$$

To obtain the dual of Optimization 1, the Lagrangian should first be minimized over $\mathbf{X}$ and α_k's, and then be maximized over the Lagrange multipliers. Observe that

- The minimum of $\left(1 - r_k^0\right)\alpha_k$ over the variable α_k is $-\infty$ unless $r_k^0 = 1$, in which case the minimum is zero.

- The minimum of the term
$$\text{Tr}\left\{A(x,r)\mathbf{X}\mathbf{X}^T\right\}$$
over $\mathbf{X}$ is $-\infty$ unless $A(x,r)$ is positive semidefinite, in which case the minimum is zero.

The proof follows immediately from these observations. ■

Proof of Part (ii) of Theorem 1: One can derive the dual of Optimization 3 by means of the standard procedure outlined in Appendix 2.7.1 (see [13] and [10] for more details). This leads to Optimization 4, where its variable W plays the role of the Lagrange multiplier for the matrix constraint (2.8a) in Optimization 3. The details are omitted for brevity. In what follows, we will show that strong duality holds between Optimizations 3 and 4. Since these optimizations are both semidefinite programs and hence convex, it suffices to prove that Optimization 4 has a finite optimal objective value and a strictly feasible point (Slater's condition). Since the OPF problem is feasible and equivalent to Optimization 1, Optimization 1 has a finite optimal value. Optimization 4 is its dual by Part (i) of Theorem 1, and is therefore upper bounded by the finite optimal value of Optimization 1 (weak duality). To show that Optimization 4 has a strictly feasible point, consider the point (x, r) given below

$$
\begin{aligned}
&\underline{\lambda}_k = \begin{cases} c_{k1}+1 & \text{if } k \in \mathcal{G} \\ 1 & \text{otherwise} \end{cases}, \quad \bar{\lambda}_k = 1, \quad \lambda_{lm} = \varepsilon \\
&\underline{\gamma}_k = \bar{\gamma}_k = 1, \\
&\underline{\mu}_k = 1, \quad \bar{\mu}_k = 2, \quad \mu_{lm} = 1, \\
&r_k^1 = 0, \quad r_k^2 = 1, \\
&r_{lm}^1 = r_{lm}^4 = r_{lm}^6 = 1, \quad r_{lm}^2 = r_{lm}^3 = r_{lm}^5 = 0
\end{aligned}
\tag{2.25}
$$

for $k \in \mathcal{N}$ and $(l,m) \in \mathcal{L}$, where ε is some positive number. Then $\lambda_k = \gamma_k = 0$ and $\mu_k = 1$. Now, observe that

- The variable x whose entries are specified in (2.25) is strictly positive componentwise.
- The relations
$$
\begin{bmatrix} r_{lm}^1 & r_{lm}^2 & r_{lm}^3 \\ r_{lm}^2 & r_{lm}^4 & r_{lm}^5 \\ r_{lm}^3 & r_{lm}^5 & r_{lm}^6 \end{bmatrix} = I \succ 0
$$
$$
\begin{bmatrix} 1 & r_{l1} \\ r_{l11} & r_{l2} \end{bmatrix} = I \succ 0
$$
hold.

- We have

  $$h(x,r) = I + \varepsilon \sum_{(l,m)\in\mathcal{L}} \mathbf{Y}_{lm} + \sum_{(l,m)\in\mathcal{L}} M_{lm}$$

  Since M_{lm} is positive semidefinite, $h(x,r)$ becomes strictly positive definite for sufficiently small values of ε.

In light of the above observations, (x,r) given in (2.25) is a strictly feasible point of Optimization 4 for an appropriate value of ε. Hence, strong duality holds. ■

Proof of Part (i) of Theorem 2: Recall that the following properties hold for Optimizations 1–4:

- The optimal (objective) values of Optimizations 1 and 2 are the same, due to the equivalence between these optimizations.

- The optimal values of Optimizations 3 and 4 are identical, due to strong duality.

These properties yield that the duality gap for Optimization 1 is equal to the difference between the optimal values of Optimizations 2 and 3. The proof is completed by noting that this difference is zero if and only if Optimization 3 has a rank-one solution.

Proof of Part (ii) of Theorem 2: Let W^{opt} denote a solution of Optimization 3. It follows from Part (ii) of Theorem 1 and the KKT conditions that

$$\text{Tr}\left\{A(x^{\text{opt}}, r^{\text{opt}})W^{\text{opt}}\right\} = 0. \tag{2.26}$$

Denote the nonzero eigenvalues of W^{opt} as $\rho_1, ..., \rho_f$ and their associated unit eigenvectors as $E_1, ..., E_f$ for some nonnegative integer f. By writing W^{opt} as $\sum_{l=1}^{f} \rho_l E_l E_l^T$, it can be conduced from (2.26) and the positive semi-definiteness of W^{opt} and $A(x^{\text{opt}}, r^{\text{opt}})$ that

$$A(x^{\text{opt}}, r^{\text{opt}})E_l = 0, \quad \forall l \in \{1, ..., f\}.$$

This implies that the orthogonal eigenvectors $E_1, ..., E_f$ all belong to the null space of $A(x^{\text{opt}}, r^{\text{opt}})$, which has dimension 2. Hence, f is less than or equal to 2. On the other hand, if $f = 1$, then Optimization 3 has a rank-one solution and consequently the duality gap is zero for Optimization 1 (see Part (i) of Theorem 2). Therefore, assume that f is equal to 2. It can be shown that there exist two matrices $T(x,r)$ and $\bar{T}(x,r)$ such that

$$A(x,r)=\begin{bmatrix} T(x,r) & \bar{T}(x,r) \\ -\bar{T}(x,r) & T(x,r) \end{bmatrix}. \tag{2.27}$$

Decompose E_1 as $\begin{bmatrix} E_{11}^T & E_{12}^T \end{bmatrix}^T$ for some vectors $E_{11}, E_{12} \in \mathbf{R}^n$. It can be inferred from the above equation that $\begin{bmatrix} -E_{12}^T & E_{11}^T \end{bmatrix}^T$ is in the null space of $A(x^{\text{opt}}, r^{\text{opt}})$ as well. Since this vector is orthogonal to E_1, the vector E_2 must be equal to $\pm\begin{bmatrix} -E_{12}^T & E_{11}^T \end{bmatrix}^T$. Thus, one can write

$$W^{\text{opt}}=\rho_1\begin{bmatrix} E_{11} \\ E_{12} \end{bmatrix}\begin{bmatrix} E_{11}^T & E_{12}^T \end{bmatrix}+\rho_2\begin{bmatrix} -E_{12} \\ E_{11} \end{bmatrix}\begin{bmatrix} -E_{12}^T & E_{11}^T \end{bmatrix}. \tag{2.28}$$

Consider now the rank-one matrix

$$(\rho_1+\rho_2)\begin{bmatrix} E_{11} \\ E_{12} \end{bmatrix}\begin{bmatrix} E_{11}^T & E_{12}^T \end{bmatrix}. \tag{2.29}$$

Since W^{opt} given in (2.28) satisfies the constraints of Optimization 3 and also maximizes its objective function, it is easy to verify that the rank-one matrix in (2.29) is also a solution of Optimization 3. In other words, Optimization 3 has a rank-one solution, which makes the duality gap for Optimization 1 equal to zero (in light of Part (i) of Theorem 2). ■

Proof of Corollary 1: As can be deduced from the proof of Part (ii) of Theorem 2, since $\begin{bmatrix} X_1^T & X_2^T \end{bmatrix}^T$ belongs to the null space of $A(x^{\text{opt}}, r^{\text{opt}})$, the vector $\begin{bmatrix} X_2^T & -X_1^T \end{bmatrix}^T$ is also is in the null space of the same matrix. Now, recall that Optimization 3 has a rank-one solution W^{opt} that is decomposable as $\mathbf{X}^{\text{opt}}(\mathbf{X}^{\text{opt}})^T$, where $\mathbf{X}^{\text{opt}}$ is a solution of Optimization 1. In light of the relation (2.26), $\mathbf{X}^{\text{opt}}$ belongs to the null space of $A(x^{\text{opt}}, r^{\text{opt}})$ and hence there exist two real numbers ζ_1 and ζ_2 such that

$$\mathbf{X}^{\text{opt}}=\zeta_1\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}+\zeta_2\begin{bmatrix} -X_2 \\ X_1 \end{bmatrix}$$

or equivalently

$$\mathbf{V}^{\text{opt}}=(\zeta_1+\zeta_2\mathrm{i})(X_1+X_2\mathrm{i}).$$

This completes the proof of Part (i) of Corollary 1. Part (ii) of this corollary follows immediately from the proof of Part (ii) of Theorem 2. The details are omitted for brevity. ■

Chapter 3

Convexification of Fundamental Nonlinear Power Problems

Most of the fundamental optimization problems for power systems are highly non-convex and NP-hard (in the worst case), partially due to the nonlinearity of certain physical quantities, e.g., active power, reactive power, and magnitude of voltage. The classical optimal power flow (OPF) problem is one of such problems, which has been studied for half a century. In the previous chapter, we obtained a condition under which the duality gap is zero for the classical OPF problem and hence a globally optimal solution to this problem can be found efficiently by solving a semidefinite program. We showed that this zero-duality-gap condition is satisfied for IEEE benchmark systems and is likely to hold widely in practice due to the physical properties of transmission lines. This chapter studies the case when there are other common sources of non-convexity, such as variable shunt elements, variable transformer ratios and contingency constraints. It is shown that zero duality gap for the classical OPF problem implies zero duality gap for a general OPF-based problem with these extra sources of non-convexity. This result makes it possible to find globally optimal solutions to several fundamental power problems in polynomial time.

3.1 Introduction

The classical optimal power flow (OPF) problem aims to find a steady-state operating point of a power system that minimizes a desirable cost function, e.g., power loss or generation cost, and satisfies network and physical constraints on loads, powers, voltages, and line flows [72]. The OPF problem is not only non-convex but also NP-hard, because of its possible reduction in a special case to the $(0,1)$-quadratic optimization. Started by the work [21] in 1962, many of the existing optimization techniques have been adapted to solve the OPF problem, leading to algorithms based on linear programming, Newton Raphson, quadratic programming, nonlinear programming, Lagrange relaxation, interior point method, artificial intelligence, artificial neural network, fuzzy logic, genetic algorithm, evolutionary programming, and particle swarm optimization [106, 73, 74, 83, 48]. Due to the non-convexity of the OPF problem, these algorithms are not robust, lack performance guarantees, and may not be able to find a global optimum.

By exploiting the physical properties of transmission lines, we showed in the previous chapter that the classical OPF problem corresponding to a practical power system can be convexified naturally and then solved efficiently. More precisely, we considered some equivalent form of the OPF problem whose dual can be cast as a semidefinite program [13]. Although this dual problem is solvable in polynomial time, its solution may not help solve the OPF problem in light of the duality gap being possibly nonzero. We derived a zero-duality-gap condition for the OPF problem in [63] and [64] under which a globally optimal solution to the OPF problem can be recovered from a solution to its dual. This condition is satisfied for all IEEE benchmark systems with 14, 30, 57, 118, and 300 buses, and is expected to hold for every practical power system (for more details, see the algebraic and geometric studies provided in [63, 64]).

Many of the fundamental optimization problems arising in power systems are based on a single or a set of classical OPF problems with more constraints and variables. A question arises as whether these problems can also be convexified. This chapter aims to address this question. The objective is to show that zero duality gap for the classical OPF problem implies zero duality gap for the following important problems (and any combinations of them) as well:

- The OPF problem with extra variables associated with unknown shunt elements [72].

- The OPF problem with extra variables associated with unknown transformer ratios [72].

- The security-constraint OPF problem (known also as contingency-constrained OPF), which corresponds to a set of coupled OPF problems [19].

The technique developed in this chapter can be used to generalize the above-mentioned zero-duality-gap result to several other OPF-based problems such as the dynamic OPF problem or the power system planning with renewable resources [117].

Notations: The following notations will be used throughout this chapter:

- i : The imaginary unit.
- $\mathbf{R}$: The set of real numbers.
- $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$: The operators returning the real and imaginary parts of a complex matrix.
- T : The transpose operator.
- $*$: The conjugate transpose operator.
- $\succeq$: The matrix inequality sign in the positive semidefinite sense [13].

3.2 Preliminaries and Problem Formulation

Given two natural numbers m and n such that $m \leq n$, consider a power network with n buses, labeled as $1, 2, ..., n$, and m generators connected to buses $1, 2, ..., m$. Assume that each bus $k \in \{1, 2, ..., n\}$ is connected to a load with the given apparent power $P_{D_k} + Q_{D_k}\mathrm{i}$ (this number is zero whenever a bus is not connected to any load). For every $l \in \{1, 2, ..., m\}$, let P_{G_l} and Q_{G_l} denote the unknown active and reactive powers supplied by generator l, respectively, and $f_l(P_{G_l}) = c_{l2}P_{G_l}^2 + c_{l1}P_{G_l} + c_{l0}$ denote a cost function associated with this generator, for some nonnegative numbers c_{l0}, c_{l1}, c_{l2}. Define V_k as the unknown complex voltage at bus $k \in \{1, 2, ..., n\}$ and $\mathbf{V}$ as the vector of all bus voltages. The classical OPF problem aims to minimize $\sum_{l=1}^{m} f_l(P_{G_l})$ over the unknown parameters $\mathbf{V}, P_{G_1}, ..., P_{G_m}$, $Q_{G_1}, ..., Q_{G_m}$ subject to the constraints that every bus $k \in \{1, 2, ..., n\}$ must be able to deliver the power $P_{D_k} + Q_{D_k}\mathrm{i}$ to its load and that

$$
\begin{aligned}
&P_{k,\min} \leq P_{G_k} \leq P_{k,\max}, && k \in \{1, ..., m\} \\
&Q_{k,\min} \leq Q_{G_k} \leq Q_{k,\max}, && k \in \{1, ..., m\} \\
&V_{k,\min} \leq |V_k| \leq V_{k,\max}, && k \in \{1, ..., n\} \\
&|S_{kl}| \leq S_{kl,\max}, && k, l \in \{1, ..., n\}
\end{aligned}
$$

for some given limits $P_{k,\min}, P_{k,\max}, Q_{k,\min}, Q_{k,\max}, V_{k,\min}, V_{k,\max}, S_{kl,\max}$, where S_{kl} denotes the apparent power transferred from bus k to the rest of the network through the line (k,l) (note that S_{kl} is zero if the line (k,l) does not exist).

In order to mathematically formulate the problem, the first step is to find an equivalent circuit model of the network with only three types of lumped elements: resistors, capacitors, and inductors. This model can be obtained by replacing every transmission line and transformer with their equivalent Π models [72]. In the derived equivalent circuit, let y_{kl} denote the mutual admittance between buses k and l, and y_{kk} denote the admittance-to-ground at bus k, for every $k,l \in \{1,2,...,n\}$. Define the admittance matrix Y of the network as an $n \times n$ complex-valued matrix whose (k,l) entry is equal to $-y_{kl}$ if $k \neq l$ and $y_{kk} + \sum_{r \in \mathcal{N}(k)} y_{kr}$ otherwise, where $\mathcal{N}(k)$ is the set of those buses that are connected to bus k. It is worth mentioning that Y plays the role of a complex-valued (generalized) Laplacian matrix for a weighted graph associated with the power system. Define the current vector $\mathbf{I}$ as $Y\mathbf{V}$ and represent its k^{th} element with I_k, for every $k \in \{1,2,...,n\}$. Note that I_k is indeed the net current injected to bus k.

Let $e_1, e_2, ..., e_n$ denote the standard basis vectors in $\mathbf{R}^n$. Define the following matrices for every $k,l \in \{1,2,...,n\}$:

$$\begin{aligned}
Y_k &:= e_k e_k^T Y \\
Y_{kl} &:= \frac{1}{2} e_k (b_{kl}\mathrm{i}) e_k^T + e_k y_{kl} e_k^T - e_k y_{kl} e_l^T \\
\mathbf{Y}_k &:= \frac{1}{2} \begin{bmatrix} \mathrm{Re}\{Y_k + Y_k^T\} & \mathrm{Im}\{Y_k^T - Y_k\} \\ \mathrm{Im}\{Y_k - Y_k^T\} & \mathrm{Re}\{Y_k + Y_k^T\} \end{bmatrix} \\
\mathbf{Y}_{kl} &:= \frac{1}{2} \begin{bmatrix} \mathrm{Re}\{Y_{kl} + Y_{kl}^T\} & \mathrm{Im}\{Y_{kl}^T - Y_{kl}\} \\ \mathrm{Im}\{Y_{kl} - Y_{kl}^T\} & \mathrm{Re}\{Y_{kl} + Y_{kl}^T\} \end{bmatrix} \\
\bar{\mathbf{Y}}_k &:= \frac{-1}{2} \begin{bmatrix} \mathrm{Im}\{Y_k + Y_k^T\} & \mathrm{Re}\{Y_k - Y_k^T\} \\ \mathrm{Re}\{Y_k^T - Y_k\} & \mathrm{Im}\{Y_k + Y_k^T\} \end{bmatrix} \\
\bar{\mathbf{Y}}_{kl} &:= \frac{-1}{2} \begin{bmatrix} \mathrm{Im}\{Y_{kl} + Y_{kl}^T\} & \mathrm{Re}\{Y_{kl} - Y_{kl}^T\} \\ \mathrm{Re}\{Y_{kl}^T - Y_{kl}\} & \mathrm{Im}\{Y_{kl} + Y_{kl}^T\} \end{bmatrix} \\
\mathbf{X} &:= \begin{bmatrix} \mathrm{Re}\{\mathbf{V}\}^T & \mathrm{Im}\{\mathbf{V}\}^T \end{bmatrix}^T
\end{aligned}$$

where b_{kl} denotes the capacitance of the transmission line (k,l) (note that Y_{kl}, $\mathbf{Y}_{kl}$, $\bar{\mathbf{Y}}_{kl}$

are all zero if $k = l$ or the line (k, l) does not exist). For every $k \in \{1, 2, ..., n\}$, denote the net active and reactive powers injected to bus k as $P_{k,\text{inj}}$ and $Q_{k,\text{inj}}$, respectively. Given $l \in \{1, ..., m\}$, $l' \in \{m+1, ..., n\}$, and $k, k' \in \{1, ..., n\}$, it can be shown that (see [63])

$$\begin{aligned}
P_{l,\text{inj}} &= P_{G_l} - P_{D_l}, \\
Q_{l,\text{inj}} &= Q_{G_l} - Q_{D_l}, \\
P_{l',\text{inj}} &= -P_{D_l'}, \\
Q_{l',\text{inj}} &= -Q_{D_l'}, \\
|V_k|^2 &= \text{trace}\left\{M_k \mathbf{X}\mathbf{X}^T\right\} \\
P_{k,\text{inj}} &= \text{trace}\left\{\mathbf{Y}_k \mathbf{X}\mathbf{X}^T\right\}, \quad Q_{k,\text{inj}} = \text{trace}\left\{\bar{\mathbf{Y}}_k \mathbf{X}\mathbf{X}^T\right\}, \\
|S_{kk'}|^2 &= \left(\text{trace}\left\{\mathbf{Y}_{kk'} \mathbf{X}\mathbf{X}^T\right\}\right)^2 + \left(\text{trace}\left\{\bar{\mathbf{Y}}_{kk'} \mathbf{X}\mathbf{X}^T\right\}\right)^2
\end{aligned}$$

where $M_k \in \mathbf{R}^{2n \times 2n}$ is a diagonal matrix whose entries are all equal to zero, except for its (k, k) and $(n+k, n+k)$ entries that are equal to 1. To simplify the presentation, assume that $f_l(P_{G_l}) = P_{G_l}$ for every $l \in \{1, 2, ..., m\}$, implying that the cost to be minimized is simply the total power generation (the results being developed here are valid for the general case as well). Hence, the classical OPF problem corresponds to the minimization of

$$\sum_{l=1}^{m} \left(\text{trace}\left\{\mathbf{Y}_l \mathbf{X}\mathbf{X}^T\right\} + P_{D_l}\right) \tag{3.1}$$

over the variable $\mathbf{X} \in \mathbf{R}^{2n}$ subject to the constraints

$$P_{k,\min} - P_{D_k} \leq \text{trace}\left\{\mathbf{Y}_k \mathbf{X}\mathbf{X}^T\right\} \leq P_{k,\max} - P_{D_k} \tag{3.2a}$$

$$Q_{k,\min} - Q_{D_k} \leq \text{trace}\left\{\bar{\mathbf{Y}}_k \mathbf{X}\mathbf{X}^T\right\} \leq Q_{k,\max} - Q_{D_k} \tag{3.2b}$$

$$(V_{k,\min})^2 \leq \text{trace}\left\{M_k \mathbf{X}\mathbf{X}^T\right\} \leq (V_{k,\max})^2 \tag{3.2c}$$

$$\text{trace}\left\{\mathbf{Y}_{kl} \mathbf{X}\mathbf{X}^T\right\}^2 + \text{trace}\left\{\bar{\mathbf{Y}}_{kl} \mathbf{X}\mathbf{X}^T\right\}^2 \leq (S_{kl,\max})^2 \tag{3.2d}$$

for all $k, l \in \{1, 2, ..., n\}$, where $P_{k,\min}$, $P_{k,\max}$, $Q_{k,\min}$, and $Q_{k,\max}$ are considered as zero (by convention) if $k > m$. In order to avoid triviality, assume that $\mathbf{X} = 0$ (or equivalently $\mathbf{V} = 0$) is not a solution to the OPF problem. We introduce four optimization problems in the sequel whose interrelation and relation to the OPF problem are illustrated in the diagram given in Figure 3.1.

Optimization 1: This optimization is obtained from the OPF problem formulated in (3.1) and (3.2) by replacing its constraint (3.2d) with the equivalent condition of the positive semi-definiteness of the matrix

$$\begin{bmatrix} (S_{kl,\max})^2 & \operatorname{trace}\left\{\mathbf{Y}_{kl}\mathbf{X}\mathbf{X}^T\right\} & \operatorname{trace}\left\{\bar{\mathbf{Y}}_{kl}\mathbf{X}\mathbf{X}^T\right\} \\ \operatorname{trace}\left\{\mathbf{Y}_{kl}\mathbf{X}\mathbf{X}^T\right\} & 1 & 0 \\ \operatorname{trace}\left\{\bar{\mathbf{Y}}_{kl}\mathbf{X}\mathbf{X}^T\right\} & 0 & 1 \end{bmatrix}.$$

Optimization 2: This optimization is defined as the dual of Optimization 1, which indeed minimizes

$$\sum_{k=1}^{n}\Bigg\{\boldsymbol{\lambda}_k P_{D_k} + \bar{\boldsymbol{\lambda}}_k Q_{D_k} + \lambda_{k,\min}P_{k,\min} - \lambda_{k,\max}P_{k,\max} + \bar{\lambda}_{k,\min}Q_{k,\min} - \bar{\lambda}_{k,\max}Q_{k,\max} + \mu_{k,\min}\left(V_{k,\min}\right)^2 - \mu_{k,\max}\left(V_{k,\max}\right)^2 - \sum_{l=1}^{n}\left(\left(S_{kl,\max}\right)^2 h_{kl}^{11} + h_{kl}^{22} + h_{kl}^{33}\right)\Bigg\}$$

over the nonnegative scalar variables $\lambda_{k,\min}, \lambda_{k,\max}, \bar{\lambda}_{k,\min}, \bar{\lambda}_{k,\max}, \mu_{k,\min}, \mu_{k,\max}$, and the positive semidefinite matrices $H_{kl} \in \mathbf{R}^{3\times3}$, $\forall k, l \in \{1, 2, ..., n\}$, subject to

$$A(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}, \boldsymbol{\mu}, \mathbf{H}) := \sum_{k=1}^{n}\left\{\boldsymbol{\lambda}_k \mathbf{Y}_k + \bar{\boldsymbol{\lambda}}_k \bar{\mathbf{Y}}_k + \boldsymbol{\mu}_k M_k + 2\sum_{l=1}^{n}\left(h_{kl}^{12}\mathbf{Y}_{kl} + h_{kl}^{13}\bar{\mathbf{Y}}_{kl}\right)\right\} \succeq 0$$

where h_{kl}^{ij} denotes the (i, j) entry of H_{kl} for every $i, j \in \{1, 2, 3\}$, and

$$\boldsymbol{\lambda}_k := \begin{cases} -\lambda_{k,\min} + \lambda_{k,\max} + 1 & \text{if } k = 1, ..., m \\ -\lambda_{k,\min} + \lambda_{k,\max} & \text{otherwise} \end{cases},$$

$$\bar{\boldsymbol{\lambda}}_k := -\bar{\lambda}_{k,\min} + \bar{\lambda}_{k,\max}, \quad \boldsymbol{\mu}_k := -\mu_{k,\min} + \mu_{k,\max},$$

$$\boldsymbol{\lambda} := \left\{\lambda_{k,\min}, \lambda_{k,\max}\right\}_{k=1}^{n}, \quad \bar{\boldsymbol{\lambda}} := \left\{\bar{\lambda}_{k,\min}, \bar{\lambda}_{k,\max}\right\}_{k=1}^{n},$$

$$\boldsymbol{\mu} := \left\{\mu_{k,\min}, \mu_{k,\max}\right\}_{k=1}^{n}, \quad \mathbf{H} = \left\{H_{kl}\right\}_{k,l=1}^{n}$$

(note that H_{kl} can be taken as zero if the line (k, l) does not exist in the power system).

Optimization 3: This optimization is obtained from Optimization 1 by first replacing every term $\mathbf{X}\mathbf{X}^T$ with a symmetric matrix variable $W \in \mathbf{R}^{2n\times2n}$ and then adding the constraint $W \succeq 0$ (thus, the variable has changed from $\mathbf{X}$ to W).

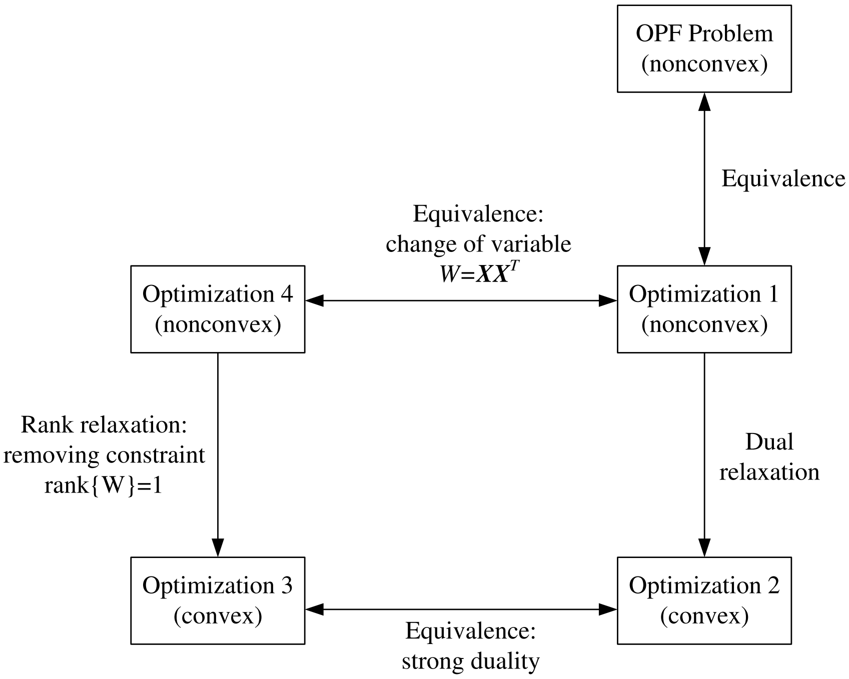


Figure 3.1: This diagram demonstrates how Optimizations 1–4 are interrelated and also related to the OPF problem.

Optimization 4: This optimization is obtained from Optimization 3 by including the additional constraint $\text{rank}\{W\} = 1$.

3.2.1 Previous Results

As illustrated in Figure 3.1 and proven in our recent work [63, 64] (see the previous chapter for more details), Optimization 1 is naturally equivalent to the OPF problem, Optimization 2 is the dual of Optimization 1, Optimization 3 is the dual of Optimization 2 (strongly duality holds), Optimization 4 is different from Optimization 3 by an extra rank constraint, and finally Optimization 1 is equivalent to Optimization 4 via the change of variable $W = \mathbf{X}\mathbf{X}^T$. Due to the natural equivalence between Optimization 1 and the OPF problem, the names *OPF problem*, *dual of the OPF problem*, and *dual of the dual of the OPF problem* will be used interchangeably for Optimizations 1, 2, and 3, respectively. The dual of the OPF problem is always feasible, but its optimal objective value can be: (i) infinite or (ii) finite. In Case (i), the OPF problem must be infeasible. In Case (ii), since the OPF problem is nonconvex, the optimal objective values of the OPF problem and its dual might not be identical. Whenever Case (i) happens (which detects the infeasibility of the OPF problem) or the optimal objective values of the OPF problem and its dual are the same, it is said that *the duality gap is zero for the OPF problem.* We proved the following important result in the preceding chapter.

Theorem 1 *The duality gap is zero for the OPF problem if its dual has a solution* $(\boldsymbol{\lambda}_{opt}, \bar{\boldsymbol{\lambda}}_{opt}, \boldsymbol{\mu}_{opt}, \mathbf{H}_{opt})$ *such that the matrix* $A(\boldsymbol{\lambda}_{opt}, \bar{\boldsymbol{\lambda}}_{opt}, \boldsymbol{\mu}_{opt}, \mathbf{H}_{opt})$ *has rank at least* $2n-2$. *In this case, two properties hold:*

- *The dual of the dual of the OPF problem has a rank-one solution* W_{opt}.
- *Given any nonzero vector* $\begin{bmatrix} U_1^T & U_2^T \end{bmatrix}^T$ *in the null space of* $A(\boldsymbol{\lambda}_{opt}, \bar{\boldsymbol{\lambda}}_{opt}, \boldsymbol{\mu}_{opt}, \mathbf{H}_{opt})$, *there exist two scalars* ζ_1 *and* ζ_2 *such that* $\mathbf{V}_{opt} = (\zeta_1 + \zeta_2 i)(U_1 + U_2 i)$ *is a solution to the OPF problem.*

Consider a special case of the OPF problem formulated in (3.1) and (3.2) where Y is a real-valued matrix, the constraints given in (3.2a), (3.2b), and (3.2d) are removed (by setting the corresponding lower and upper bounds as $-\infty$ and $+\infty$), all reactive loads are zero, and finally $V_{k,\min} = V_{k,\max}$ for every $k \in \{1, 2, ..., n\}$. In this case, the feasibility region for the real-valued vector $(V_1, ..., V_n)$ consists of 2^n points in the form of $(\pm V_{1,\min}, ..., \pm V_{n,\min})$. This substantiates that the OPF problem may have a complicated feasibility region, which can make it NP-complete for an arbitrary Y and hence create a nonzero duality gap [63]. However, the admittance matrix Y corresponding to a power system is structured in light of the physical properties of transmission lines. Using this fact, we showed in the preceding chapter that the zero-duality-gap condition stated in Theorem 1 is satisfied for all IEEE test systems with 14, 30, 57, 118, and 300 buses and, moreover, this condition is likely to hold for every practical power system.

3.2.2 Problem Statement

Define $\mathcal{D}$ as the set of every admittance matrix Y whose associated OPF problem has no duality gap for all possible values of the limits $P_{k,\min}, P_{k,\max}, Q_{k,\min}, Q_{k,\max}, V_{k,\min}, V_{k,\max}, S_{kl,\max}$, $k, l \in \{1, 2, ..., n\}$. The set $\mathcal{D}$ characterizes every network topology (including those corresponding to practical power systems) for which a globally optimal solution of the OPF problem can be found efficiently (by solving its dual). Recall that the classical OPF problem was non-convex partially due to the nonlinearity of *active power*, *reactive power*, and *magnitude of voltage* with respect to the state variable $\mathbf{V}$. This source of non-convexity appears in almost all fundamental optimization-based power problems. These problems are often based on a single or a set of coupled classical OPF problems with more sources of non-convexity, e.g., variable transformer ratios, variable shunt elements, stability constraints, and security constraints. The objective of this chapter is to prove the following statement: *zero duality gap for the classical OPF problem implies zero duality gap for harder power problems with more sources of non-convexity.* In other words, it is intended to show that if Y belongs to $\mathcal{D}$, fundamental power problems (based on OPF) can be convexified naturally via the duality theory.

3.3 Main Results

Different generalizations to the classical OPF problem will be studied in the sequel.

3.3.1 Security-Constrained Optimal Power Flow

As far as the steady-state operation of a power system is concerned, there are two types of parameters: (i) a state vector $\mathbf{X}$ containing the real and imaginary parts of the bus voltages, (ii) a control vector $\mathbf{U}$ containing the controllable parameters of the power system. Note that every power system has certain controllable parameters (depending on its control strategy) such as active powers and voltage magnitudes at generator buses, sizes of capacitor banks, and transformer tap ratios. A general OPF-based problem can be formulated as:

$$\min_{\mathbf{X},\mathbf{U}} f(\mathbf{X},\mathbf{U}) \tag{3.4a}$$

$$\text{s.t.} \quad g(\mathbf{X},\mathbf{U}) = 0 \tag{3.4b}$$

$$h(\mathbf{X},\mathbf{U}) \geq 0, \tag{3.4c}$$

where

- $f(\mathbf{X},\mathbf{U})$ is an appropriate cost to be minimized (such as power loss or total generation cost).
- The relation (3.4b) describes the set of all equality constraints resulting from the power flow equations.

- The relation (3.4c) describes the set of all inequality constraints resulting from the physical limits imposed on the parameters of the system.

Assume that the power system is subject to c different contingencies, where each contingency corresponds to a new configuration in which certain transmission lines and generators are disconnected. The security-constrained optimal power flow (SCOPF) problem aims to optimize the performance of the power system under the normal condition such that the load and physical constraints are still satisfied after every pre-specified contingency. This problem can be formulated as:

$$\min_{\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(c)},\mathbf{U}^{(0)},\ldots,\mathbf{U}^{(c)}} \; f\left(\mathbf{X}^{(0)},\mathbf{U}^{(0)}\right) \tag{3.5a}$$

$$\text{s.t.} \quad g_t\left(\mathbf{X}^{(t)},\mathbf{U}^{(t)}\right) = 0, \quad t = 0,\ldots,c \tag{3.5b}$$

$$h_t\left(\mathbf{X}^{(t)},\mathbf{U}^{(t)}\right) \geq 0, \quad t = 0,\ldots,c \tag{3.5c}$$

$$\left|\mathbf{U}^{(r)} - \mathbf{U}^{(0)}\right| \leq \Delta\mathbf{U}^{(r)}_{\max}, \quad r = 1,\ldots,c \tag{3.5d}$$

where

- $\mathbf{X}^{(t)}$ and $\mathbf{U}^{(t)}$ denote the state and control vectors for the t^{th} configuration ($t = 0$ is the normal configuration and $t > 0$ is a contingency case).

- The equality and inequality constraints for the t^{th} configuration are given by (3.5b) and (3.5c).

- Given the constant vector $\Delta\mathbf{U}^{(r)}_{\max}$, the constraint (3.5d) accounts for the fact that the controllable parameters of a power system may not be able to change arbitrarily fast after a reconfiguration (this is partially due to physical ramp-up and ramp-down constraints).

It is worth mentioning that if $\Delta\mathbf{U}^{(r)}_{\max}$ is zero, the corresponding control strategy is said to be *preventive* in light of taking no control action after a contingency; otherwise, it is said to be *corrective*. Note that some of the entries of $\Delta\mathbf{U}^{(r)}_{\max}$ can be infinity, implying that the corresponding controllable parameters can change arbitrarily after a reconfiguration.

The objective of this part is to prove the following statement: *zero duality gap for the OPF problem implies zero duality gap for the SCOPF problem.* To this end, for the sake of simplifying the presentation, assume that every controllable parameter can only be an active power, a reactive power or a voltage magnitude. Using the techniques being developed in the next subsections, the results can be generalized to incorporate loads, shunt elements, and transformer ratios into the control vector $\mathbf{U}$. Moreover, with no loss of generality, suppose that the cost function $f_0(\mathbf{X}^{(0)}, \mathbf{U}^{(0)})$ is the total power generation. As before, we use the superscript (t) for every parameter of the power system in the t^{th} configuration, $t = 0, 1, ..., c$. For instance, $Y^{(0)}$ is equal to Y, and $Y^{(r)}$, $r = 1, 2, ..., c$, is the admittance matrix of the power system under the r^{th} contingency. The SCOPF problem can be expressed as the minimization of

$$\sum_{l=1}^{m} P_{G_l}^{(0)} \tag{3.6}$$

subject to

$$P_{k,\min}^{(t)} \leq P_{G_k}^{(t)} \leq P_{k,\max}^{(t)}, \qquad k \in \{1, ..., m\} \tag{3.7a}$$

$$Q_{k,\min}^{(t)} \leq Q_{G_k}^{(t)} \leq Q_{k,\max}^{(t)}, \qquad k \in \{1, ..., m\} \tag{3.7b}$$

$$V_{k,\min}^{(t)} \leq \left|V_k^{(t)}\right| \leq V_{k,\max}^{(t)}, \qquad k \in \{1, ..., n\} \tag{3.7c}$$

$$\left|S_{kl}^{(t)}\right| \leq S_{kl,\max}^{(t)}, \qquad k, l \in \{1, ..., n\} \tag{3.7d}$$

$$\left|P_{G_k}^{(r)} - P_{G_k}^{(0)}\right| \leq \Delta P_{k,\max}^{(r)}, \qquad k \in \{1, ..., m\} \tag{3.7e}$$

$$\left|Q_{G_k}^{(r)} - Q_{G_k}^{(0)}\right| \leq \Delta Q_{k,\max}^{(r)}, \qquad k \in \{1, ..., m\} \tag{3.7f}$$

$$\left|\left|V_k^{(r)}\right|^2 - \left|V_k^{(0)}\right|^2\right| \leq \left(\Delta V_{k,\max}^{(r)}\right)^2, \quad k \in \{1, ..., n\} \tag{3.7g}$$

for every $t \in \{0, ..., c\}$ and $r \in \{1, ..., c\}$, where $\Delta P_{k,\max}^{(r)}$, $\Delta Q_{k,\max}^{(r)}$, and $\Delta V_{k,\max}^{(r)}$ are some given nonnegative numbers. Note that

- If $\Delta P_{k,\max}^{(r)}$ is zero, it implies that no corrective action is taken for the controllable parameter P_{G_k}. Furthermore, if $\Delta P_{k,\max}^{(r)}$ is infinity (so that the corresponding inequality can be removed from the SCOPF problem) implies that P_{G_k} is either a non-controllable parameter or a controllable parameter with no ramp constraint. A similar remark can be made about Q_{G_k} and V_k.

- The formulation given in (3.6) and (3.7) is capable of modeling faults in both the transmission network and the generators. For instance, $Y^{(1)} \neq Y$ implies that some of the transmission lines are disconnected under the first contingency, while $Y^{(1)} = Y$ and $P^{(1)}_{1,\min} = P^{(1)}_{1,\max} = 0$ imply that the generator of bus 1 is removed under the first contingency.

The dual of the SCOPF problem can be derived based on the method presented in [63], which turns out to be a semidefinite program similar to Optimization 2. A question arises as whether a globally optimal solution of the non-convex SCOPF problem can be found by solving this semidefinite program. This question is answered in the next theorem.

Theorem 2 *Assume that the duality gap is zero for every classical OPF problem associated with each of the configurations $Y^{(0)}, Y^{(1)}, ..., Y^{(c)}$ (i.e., $Y^{(0)}, ..., Y^{(c)} \in \mathcal{D}$). Then, the duality gap is zero for the SCOPF problem as well so that a globally optimal solution of this problem can be recovered from an optimal solution of its convex dual problem.*

Proof: Consider the optimization problem of minimizing

$$\sum_{l=1}^{m} \left(\text{trace}\left\{ \mathbf{Y}_l^{(0)} W^{(0)} \right\} + P_{D_l} \right) \tag{3.8}$$

over the positive semidefinite matrices $W^{(0)}, W^{(1)}, ..., W^{(c)}$ subject to

$$P^{(t)}_{k,\min} - P_{D_k} \leq \text{trace}\left\{ \mathbf{Y}_k^{(t)} W^{(t)} \right\} \leq P^{(t)}_{k,\min} - P_{D_k} \tag{3.9a}$$

$$Q^{(t)}_{k,\min} - Q_{D_k} \leq \text{trace}\left\{ \bar{\mathbf{Y}}_k^{(t)} W^{(t)} \right\} \leq Q^{(t)}_{k,\min} - Q_{D_k} \tag{3.9b}$$

$$\text{trace}\left\{ \mathbf{Y}_{kl}^{(t)} W^{(t)} \right\}^2 + \text{trace}\left\{ \bar{\mathbf{Y}}_{kl}^{(t)} W^{(t)} \right\}^2 \leq \left(S^{(t)}_{kl,\max} \right)^2 \tag{3.9c}$$

$$\left(V^{(t)}_{k,\min} \right)^2 \leq \text{trace}\left\{ M_k W^{(t)} \right\} \leq \left(V^{(t)}_{k,\max} \right)^2 \tag{3.9d}$$

$$\left| \text{trace}\left\{ \mathbf{Y}_k^{(r)} W^{(r)} \right\} - \text{trace}\left\{ \mathbf{Y}_k^{(0)} W^{(0)} \right\} \right| \leq \Delta P^{(r)}_{k,\max} \tag{3.9e}$$

$$\left| \text{trace}\left\{ \bar{\mathbf{Y}}_k^{(r)} W^{(r)} \right\} - \text{trace}\left\{ \bar{\mathbf{Y}}_k^{(0)} W^{(0)} \right\} \right| \leq \Delta Q^{(r)}_{k,\max} \tag{3.9f}$$

$$\left| \text{trace}\left\{ M_k W^{(r)} \right\} - \text{trace}\left\{ M_k W^{(0)} \right\} \right| \leq \left(\Delta V^{(r)}_{k,\max} \right)^2 \tag{3.9g}$$

for every $k, l \in \{1, 2, ..., n\}$, $r \in \{1, 2, ..., c\}$, and $t \in \{0, 1, ..., c\}$ (as before, $P^{(t)}_{k,\min}$, $P^{(t)}_{k,\max}$, $Q^{(t)}_{k,\min}$, and $Q^{(t)}_{k,\max}$ are set to zero by convention if $k > m$). A diagram similar to the one despited in Figure 3.1 for the OPF problem can be derived to deduce the following two properties:

i) The convex optimization given in (3.8) and (3.9) is the dual of the dual of the SCOPF problem specified in (3.6) and (3.7).

ii) The optimization given in (3.8) and (3.9) under the additional non-convex constraints $\text{rank}\{W^{(0)}\} = \cdots = \text{rank}\{W^{(c)}\} = 1$ can be equivalently converted to the SCOPF problem via the change of variables $W^{(t)} = \mathbf{X}^{(t)}\mathbf{X}^{(t)^T}$, $t = 0, 1, ..., c$.

It can be concluded from Property (ii) that the SCOPF problem is infeasible if the optimization problem (3.8) and (3.9) is infeasible. Therefore, assume that the latter optimization problem is feasible. Using the above properties and in line with the argument made in [63], one can infer that the duality gap between the SCOPF problem and its dual is zero, provided the optimization problem given in (3.8) and (3.9) has a minimizer $(W^{(0)}_{\text{opt}}, W^{(1)}_{\text{opt}} ..., W^{(c)}_{\text{opt}})$ such that

$$\text{rank}\left\{W^{(0)}_{\text{opt}}\right\} = \text{rank}\left\{W^{(1)}_{\text{opt}}\right\} = \cdots = \text{rank}\left\{W^{(c)}_{\text{opt}}\right\} = 1.$$

To prove the existence of such a solution to the dual of the dual of the SCOPF problem, let $(W^{(0)}_{\text{opt}}, W^{(1)}_{\text{opt}} ..., W^{(c)}_{\text{opt}})$ be an arbitrary minimizer of this optimization problem. Consider a feasibility problem with the variable $W^{(c)}$ and the constraints $(\forall k, l \in \{1, ..., n\})$

$$\text{trace}\left\{\mathbf{Y}^{(c)}_k W^{(c)}\right\} = \text{trace}\left\{\mathbf{Y}^{(c)}_k W^{(c)}_{\text{opt}}\right\} \tag{3.10a}$$

$$\text{trace}\left\{\bar{\mathbf{Y}}^{(c)}_k W^{(c)}\right\} = \text{trace}\left\{\bar{\mathbf{Y}}^{(c)}_k W^{(c)}_{\text{opt}}\right\} \tag{3.10b}$$

$$\text{trace}\left\{\mathbf{Y}^{(c)}_{kl} W^{(c)}\right\}^2 + \text{trace}\left\{\bar{\mathbf{Y}}^{(c)}_{kl} W^{(c)}\right\}^2 \leq \left(S^{(c)}_{kl,\max}\right)^2 \tag{3.10c}$$

$$\text{trace}\left\{M_k W^{(c)}\right\} = \text{trace}\left\{M_k W^{(c)}_{\text{opt}}\right\}. \tag{3.10d}$$

Obviously, $W^{(c)} = W^{(c)}_{\text{opt}}$ is a solution to this feasibility problem (i.e., it satisfies the above constraints). In addition, it can be shown that $(W^{(0)}_{\text{opt}}, ..., W^{(c-1)}_{\text{opt}}, W_f)$ is a solution to the optimization given in (3.8) and (3.9) for some matrix W_f if $W^{(c)} = W_f$ is a solution to the feasibility problem (3.10). Now, the goal is to show that this feasibility problem has a rank-one solution W_f. To this end, convert this feasibility problem into an optimization problem by minimizing $\sum_{l=1}^{m}(\text{trace}\{\mathbf{Y}^{(c)}_l W^{(c)}\} + P_{D_l})$. The diagram given in Figure 3.1

yields that this optimization problem is the dual of the dual of the following OPF problem:

$$\begin{aligned}
&\min \sum_{l=1}^{m} P_{G_l}^{(c)} \\
\text{s.t.}\quad & P_{k,\text{inj}}^{(c)} = \text{trace}\left\{\mathbf{Y}_k^{(c)} W_{\text{opt}}^{(c)}\right\}, \quad k \in \{1,...,n\} \\
& Q_{k,\text{inj}}^{(c)} = \text{trace}\left\{\bar{\mathbf{Y}}_k^{(c)} W_{\text{opt}}^{(c)}\right\}, \quad k \in \{1,...,n\} \\
& \left|V_k^{(c)}\right|^2 = \text{trace}\left\{M_k W_{\text{opt}}^{(c)}\right\}, \quad k \in \{1,...,n\} \\
& \left|S_{kl}^{(c)}\right|^2 \leq \left(S_{kl,\max}^{(c)}\right)^2, \quad k,l \in \{1,...,n\}.
\end{aligned}$$

Since the duality gap is zero for this OPF problem (due to the assumption $Y^{(c)} \in \mathcal{D}$), the dual of its dual has a rank-one solution (see Theorem 1 and the diagram given in Figure 3.1). In other words, there exists a rank-one matrix W_f such that $W^{(c)} = W_f$ satisfies the constraints given in (3.10). Following the discussion made earlier, this result simply implies that $W_{\text{opt}}^{(c)}$ can be taken as the rank-one matrix W_f. The same argument can be continued for other matrices $W_{\text{opt}}^{(0)},...,W_{\text{opt}}^{(c-1)}$ to conclude that the dual of the dual of the SCOPF problem has a solution $(W_{\text{opt}}^{(0)}, W_{\text{opt}}^{(1)}...,W_{\text{opt}}^{(c)})$, where each of the matrices $W_{\text{opt}}^{(0)},...,W_{\text{opt}}^{(c)}$ has rank one. This completes the proof. ■

3.3.2 Optimization of Shunt Elements in Power Systems

A popular method towards a better steady-state control of a power system is to exploit variable reactive/capacitive shunt elements (e.g., capacitor banks or static VAR compensators) at some designated buses. To optimize these shunt parameters, they should be incorporated into the classical OPF problem. In order to formulate the underlying problem, assume that each bus $k \in \{1,2,...,n\}$ is equipped with a variable shunt element with the admittance $b_k\text{i}$, where b_k must lie between two given lower and upper bounds $b_{k,\min}$ and $b_{k,\max}$ (these bounds can take both positive and negative values). Note that if some bus k does not have such a shunt element, the bounds $b_{k,\min}$ and $b_{k,\max}$ are set to zero. As before, with no loss of generality, assume that the objective function to be minimized is the total generation. The elements $b_1,...,b_n$ can be directly incorporated into the admittance matrix of the power system, which makes some of the elements of this matrix unknown and therefore adds another source of non-convexity to the OPF problem. Alternatively, one can use the fact that the shunt element of bus $k \in \{1,2,...,n\}$ injects no active power but the reactive power $b_k|V_k|^2$

to its corresponding bus. Hence, the resulting OPF problem with variable shunt elements can be obtained from the classical OPF problem by replacing the constraints

$$Q_{k,\min} - Q_{D_k} \leq Q_{k,\text{inj}} \leq Q_{k,\max} - Q_{D_k}, \quad k = 1, ..., n$$

with the new constraints

$$\begin{aligned} & Q_{k,\min} - Q_{D_k} + b_k |V_k|^2 \leq Q_{k,\text{inj}} && \text{(3.11a)} \\ & Q_{k,\text{inj}} \leq Q_{k,\max} - Q_{D_k} + b_k |V_k|^2 && \text{(3.11b)} \\ & b_{k,\min} \leq b_k \leq b_{k,\max} && \text{(3.11c)} \end{aligned}$$

where $b_1, ..., b_n$ are a part of the variables of the new optimization problem. Since $|V_k|$ is a nonnegative number, the change of variable $Q_{k,b} := b_k |V_k|^2$ equivalently converts the OPF problem with variable shunt elements into the following:

$$\begin{aligned} & \min \quad \sum_{l=1}^{m} P_{G_l} \\ \text{s.t.} \quad & P_{k,\min} - P_{D_k} \leq P_{k,\text{inj}} \\ & P_{k,\text{inj}} \leq P_{k,\max} - P_{D_k} \\ & Q_{k,\min} - Q_{D_k} + Q_{k,b} \leq Q_{k,\text{inj}} \\ & Q_{k,\text{inj}} \leq Q_{k,\max} - Q_{D_k} + Q_{k,b} \\ & V_{k,\min} \leq |V_k| \leq V_{k,\max} \\ & |S_{kl}| \leq S_{kl,\max} \\ & b_{k,\min} |V_k|^2 \leq Q_{k,b} \leq b_{k,\max} |V_k|^2 \end{aligned} \tag{3.12}$$

$\forall k, l \in \{1, 2, ..., n\}$, where $Q_{1,b}, ..., Q_{n,b}$ are the extra variables of the optimization problem. In this subsection, we study this variant of the OPF problem with unknown shunt elements. The dual of this problem can be expressed as a semidefinite program. The next theorem proves that the solution of this dual problem can be used to find a solution to the original primal problem.

Theorem 3 *Assume that the duality gap is zero for every classical OPF problem associated with the configuration Y (i.e., $Y \in \mathcal{D}$). Then, the duality gap is zero for the OPF problem* *with variable shunt elements as well.*

Sketch of Proof: The dual of the dual of the OPF problem with variable shunt elements minimizes $\sum_{l=1}^{m}(\text{trace}\{\mathbf{Y}_l W\}+P_{D_l})$ over the positive semidefinite matrix W and the scalars $Q_{1,b},...,Q_{n,b}$ subject to

$$\begin{aligned}
&P_{k,\min} - P_{D_k} \leq \text{trace}\left\{\mathbf{Y}_k W\right\} \leq P_{k,\min} - P_{D_k}\\
&Q_{k,\min} - Q_{D_k} + Q_{k,b} \leq \text{trace}\left\{\bar{\mathbf{Y}}_k W\right\}\\
&\text{trace}\left\{\bar{\mathbf{Y}}_k W\right\} \leq Q_{k,\min} - Q_{D_k} + Q_{k,b}\\
&\text{trace}\left\{\mathbf{Y}_{kl} W\right\}^2 + \text{trace}\left\{\bar{\mathbf{Y}}_{kl} W\right\}^2 \leq (S_{kl,\max})^2\\
&(V_{k,\min})^2 \leq \text{trace}\left\{M_k W\right\} \leq (V_{k,\max})^2\\
&b_{k,\min}\text{trace}\left\{M_k W\right\} \leq Q_{k,b} \leq b_{k,\max}\text{trace}\left\{M_k W\right\}
\end{aligned}$$

$\forall k,l \in \{1,2,...,n\}$. Similar to the technique used in the proof of Theorem 2, it suffices to show that this optimization problem has a solution $(W_{\text{opt}},Q_{1,b}^{\text{opt}},...,Q_{n,b}^{\text{opt}})$ such that $\text{rank}\{W_{\text{opt}}\}=1$. To this end, given an arbitrary solution $(W_{\text{opt}},Q_{1,b}^{\text{opt}},...,Q_{n,b}^{\text{opt}})$ to the above problem, consider the following optimization:

$$\begin{aligned}
\min_{W} \quad & \sum_{l=1}^{m}\text{trace}\left\{\mathbf{Y}_l W\right\}\\
\text{s.t.} \quad & \text{trace}\left\{\mathbf{Y}_k W\right\} = \text{trace}\left\{\mathbf{Y}_k W_{\text{opt}}\right\}\\
& \text{trace}\left\{\bar{\mathbf{Y}}_k W\right\} = \text{trace}\left\{\bar{\mathbf{Y}}_k W_{\text{opt}}\right\}\\
& \text{trace}\left\{\mathbf{Y}_{kl} W\right\}^2 + \text{trace}\left\{\bar{\mathbf{Y}}_{kl} W\right\}^2 \leq (S_{kl,\max})^2\\
& \text{trace}\left\{M_k W\right\} = \text{trace}\left\{M_k W_{\text{opt}}\right\}
\end{aligned} \tag{3.13}$$

$\forall k,l \in \{1,2,...,n\}$. It can be observed that

i) $W = W_{\text{opt}}$ is a solution to the optimization (3.13).

ii) The feasibility region of the optimization (3.13) is a subset of the feasibility region of the dual of the dual of the OPF problem with variable shunt elements after fixing $Q_{k,b}$ as $Q_{k,b}^{\text{opt}}$, $k=1,...,n$.

These two properties imply that $(W_f,Q_{1,b}^{\text{opt}},...,Q_{n,b}^{\text{opt}})$ is a solution to the dual of the dual of the OPF problem with variable shunt elements for any arbitrary minimizer W_f of the optimization (3.13). On the other hand, the optimization (3.13) is the dual of the dual of some classical OPF problem with respect to the configuration Y. Hence, this optimization problem has a rank-one solution W_f. As a result, the minimizer W_{opt} can be taken as W_f. This completes the proof. ■

3.3.3 Optimization of Transformer Ratios in Power Systems

Every practical power system is normally accompanied by a number of transformers whose (tap) ratios are controllable within certain limits. To optimize the performance of a power system, these ratios are often considered as some controllable parameters in the corresponding OPF problem. This subsection aims to study how the OPF problem with variable transformer ratios can be convexified using the duality theory. To this end, consider a transformer installed on some transmission line of the system. The most common method is to replace the transformer with a two-port Π block in order to be able to have an equivalent circuit model for the power system with only resistors, capacitors, and inductors. However, if the transformer ratio is an unknown parameter, it appears in a nonlinear way in the admittance matrix of the equivalent circuit model.

To bypass the foregoing issue, we exploit a different modeling method here. First, we replace every transformer with an ideal transformer and some lumped elements (accounting for the leakage reactance, series resistance, etc.). Then, we add some virtual buses to the set of the real (existing) buses in such a way that every ideal transformer is connected directly to two real/virtual buses (this may need defining a virtual bus for every transformer). For the sake of simplifying the presentation, we present the ideas for the case when there is only one tap-changing transformer in the system that connects bus 1 to bus 2. The generalization to multi-transformer case is straightforward.

Assume that bus 1 is connected to bus 2 via an ideal transformer. Let $P_{12}+Q_{12}\mathrm{i}$ denote the power transferred from bus 1 to the rest of the network through the transformer and η denote the transformer ratio bounded by the given nonnegative numbers $\eta_{\min}$ and $\eta_{\max}$. With a slight abuse of notation, define Y as the admittance matrix of the power system after removing the transformer (i.e., after disconnecting the line $(1,2)$). By virtue of having no power loss in the transformer, one can write the power flow equations at buses 1 and 2 as follows:

$$\begin{aligned}\text{trace}\left\{\mathbf{Y}_1\mathbf{X}\mathbf{X}^T\right\} &= P_{1,\text{inj}} - P_{12}\\ \text{trace}\left\{\mathbf{Y}_2\mathbf{X}\mathbf{X}^T\right\} &= P_{2,\text{inj}} + P_{12}\\ \text{trace}\left\{\bar{\mathbf{Y}}_1\mathbf{X}\mathbf{X}^T\right\} &= Q_{1,\text{inj}} - Q_{12}\\ \text{trace}\left\{\bar{\mathbf{Y}}_2\mathbf{X}\mathbf{X}^T\right\} &= Q_{2,\text{inj}} + Q_{12}.\end{aligned} \tag{3.14}$$

On the other hand, the voltages at the two ports of the transformer are related as

$$\text{Re}\{V_1\} = \eta \times \text{Re}\{V_2\} \tag{3.15a}$$

$$\text{Im}\{V_1\} = \eta \times \text{Im}\{V_2\} \tag{3.15b}$$

$$\eta_{\text{min}} \leq \eta \leq \eta_{\text{max}}. \tag{3.15c}$$

In order to remove the nonlinearity caused by the product of η and the components of V_2, we eliminate the variable η. For this purpose, consider the relations

$$\eta_{\text{min}}^2|V_2|^2 \leq |V_1|^2 \leq \eta_{\text{max}}^2|V_2|^2 \tag{3.16a}$$

$$\text{Re}\{V_1\} \times \text{Im}\{V_2\} = \text{Re}\{V_2\} \times \text{Im}\{V_1\} \tag{3.16b}$$

$$\text{Re}\{V_1\} \times \text{Re}\{V_2\} \geq 0 \tag{3.16c}$$

$$\text{Im}\{V_1\} \times \text{Im}\{V_2\} \geq 0. \tag{3.16d}$$

It can be shown that the relations in (3.16) are satisfied if and only if there exists a nonnegative number η satisfying the relations in (3.15). Notice that all of the constraints given in (3.16) are quadratic in $\mathbf{V}$, which is a useful property for studying the duality gap. To formulate the OPF problem with the variable tap ratio η, the following actions should be taken:

- Write the power flow equations and physical limit constraints for every bus $k \in \{3, 4, ..., n\}$.

- Write the line flow constraints for all lines except for the line $(1, 2)$.

- Add the extra constraints given in (3.14) and (3.16), where P_{12} and Q_{12} are considered as scalar variables.

- Add the condition $P_{12}^2 + Q_{12}^2 \leq (S_{12,\text{max}})^2$ associated with the flow constraint of the line $(1,2)$.

It can be verified that the dual of this problem is a semidefinite program with the same structure as the dual of the classical OPF problem (partially due to the quadratic nature of the constraints in (3.16)). Now, one can write the dual of the dual of the OPF problem with the variable tap ratio η in terms of the matrix variable W and the scalar variables P_{12}, Q_{12}. In this optimization problem, the constraints corresponding to the ones given in (3.16) are

$$\begin{aligned} &\eta_{\min}^2 \text{trace}\{M_2 W\} \leq \text{trace}\{M_1 W\} \\ &\text{trace}\{M_1 W\} \leq \eta_{\max}^2 \text{trace}\{M_2 W\} \end{aligned}$$

and

$$\begin{aligned} &W_{1,n+2} = W_{2,n+1} \\ &W_{1,2} \geq 0, \quad W_{n+1,n+2} \geq 0 \end{aligned} \tag{3.17}$$

where $W_{i,j}$ denotes the (i,j) entry of W for every $i,j \in \{1,2,...,2n\}$. Now, it can be observed that the constraints corresponding to the unknown transformer ratio have appeared linearly in terms of the entries of W. If the conditions in (3.17) are removed from the dual of the dual of the OPF problem with the extra variable η, the technique used in the proof of Theorem 3 can be simply applied to this problem to show the existence of no duality gap. The removal of these two constraints corresponds to designing a complex-valued transformer ratio η such that $\eta_{\min} \leq |\eta| \leq \eta_{max}$. However, for the case when the ratio η is a real number (as considered in this work), the above-mentioned technique is not sufficient and, indeed, the long proof developed in [63] for the classical OPF problem should be followed closely. The details are omitted here for brevity.

3.3.4 Further Generalizations

It was shown in the preceding subsections that zero duality gap for the OPF problem implies zero duality gap for a general OPF-based problem with variable shunt elements, variable transformer ratios and contingency constraints. The technique used in the proofs of Theorems 2 and 3 can be exploited to extend the results to several other cases. Two of these generalizations are given below:

- The aforementioned optimizations are all *static*, corresponding to the steady-state operation of the power system. However, one can define a discrete-time dynamic

OPF (associated with the optimal control of the power system), where an optimal dynamic equation (rather than an optimal value) should be found for each controllable parameter. This problem can be tackled similarly to the SCOPF problem to prove the existence of no duality gap.

- Assume that a part of the power generated in the network is supplied by renewable resources. Then, the power coming from these resources must not exceed a certain portion of the total generated power in order to maintain the dynamic stability of the system. Adding a stability constraint to guarantee this does not create a duality gap.

3.4 Simulation Results

Let the results of this chapter be applied to the IEEE test systems with 14 and 30 buses. The specifications of these benchmark systems can be found in the library of the toolbox [118] and the online database [109].

The IEEE 30-bus system has 6 generators at buses 1, 2, 13, 22, 23, and 27. Assume that the controllable parameters of the system are the active powers supplied by the generators and the voltage magnitudes at the generator buses. If the classical OPF problem is solved to minimize the total generation (or equivalently the active power loss), the optimal values of the controllable parameters will be obtained as

$$\begin{aligned}
&P_{G_1} = 7.69, \quad P_{G_2} = 48.57, \quad P_{G_{13}} = 40.00,\\
&P_{G_{22}} = 32.17, \quad P_{G_{23}} = 16.66, \quad P_{G_{27}} = 45.99,\\
&|V_1| = 1.028, \quad |V_2| = 1.027, \quad |V_{13}| = 1.090,\\
&|V_{22}| = 1.032, \quad |V_{23}| = 1.048, \quad |V_{27}| = 1.069.
\end{aligned} \tag{3.18}$$

Suppose that while the controllable parameters of the power system are controlled continuously in order to be kept at their optimal values, a fault happens in the transmission line $(2,6)$ leading to its disconnection. It can be shown that some of the line flow constraints will be violated in this case. To avoid this issue, one can solve an SCOPF problem to optimize the controllable parameters in such a way that the total generation is minimized and that the power flow and physical constraints are satisfied in the normal and contingency states. Due to the non-convexity of the SCOPF problem, this work suggests solving the dual of the

SCOPF problem, which is a semidefinite problem. The duality gap is zero for this problem and, therefore, a globally optimal solution to the SCOPF problem can be obtained as

$$\begin{aligned}
&P_{G_1} = 12.66, \quad P_{G_2} = 43.06, \quad P_{G_{13}} = 40.00,\\
&P_{G_{22}} = 31.16, \quad P_{G_{23}} = 18.89, \quad P_{G_{27}} = 45.50,\\
&|V_1| = 1.031, \quad |V_2| = 1.030, \quad |V_{13}| = 1.094,\\
&|V_{22}| = 1.021, \quad |V_{23}| = 1.048, \quad |V_{27}| = 1.068.
\end{aligned} \tag{3.19}$$

Now, consider the problem of the loss minimization for the IEEE 14-bus system, where the tap ratios of the transformers in the lines $(4, 7)$ and $(4, 9)$ are to be optimized as well. Assume that these unknown tap ratios must lie in the range $(0.8, 1.2)$. The duality gap for the OPF problem with these two variable tap ratios turns out to be zero, which makes it possible to globally optimize the parameters of the system. The optimal tap ratios for the transformers $(4, 7)$ and $(4, 9)$ are both equal to 0.9157. If the transformers are equipped with phase shifters, the optimal complex ratios of these transformers will be obtained as $0.9158 + 0.0066\text{i}$ and $0.9157 - 0.0146\text{i}$.

As another example, assume that two reactive shunt elements are installed at buses 10 and 15 of the IEEE 30-bus system, where the reactance of each of them can change continuously in the interval $[-0.1, 0.1]$. The duality gap is zero for the OPF problem with these variable shunt elements. The globally optimal values of the shunt elements at buses 10 and 15 can be obtained as 0.0992 and 0.0701, which correspond to the reactive powers 10.5040 and 7.6369, respectively. Now, suppose that the sum of the reactive powers generated by these shunt elements cannot be more than 10. The duality gap is again zero under this new constraint. The new optimal values of the shunt elements are 0.0992 and -0.0046, corresponding to the reactive powers 10.4944 and -0.4944, respectively.

3.5 Summary

The classical optimal power flow (OPF) problem is one of the most fundamental optimization problems in power systems. This problem has been extensively studied in the past several years to deal with its non-convexity. Although the dual of the OPF problem is a semidefinite program that can be solved efficiently, the lack of strong duality might not allow recovery of a solution to the OPF problem. However, we showed in the previous chapter that the duality gap is zero for IEEE test systems and, more importantly, this gap is very likely to be zero for every practical power system due to the physical properties of a power network. In this chapter, it is shown that this duality-gap result can be generalized to a great extent. More precisely, it is proved that zero duality gap for the classical OPF problem implies zero duality gap for more complicated power problems with other sources of non-convexity, such as variable shunt elements, variable transformer ratios and security constraints.

Part II

Circuit and Systems

Chapter 4

Synthesis of Large-Scale Linear Circuits

Motivated by different applications in circuits, electromagnetics, and optics, this chapter is concerned with the synthesis of a particular type of linear circuit, where the circuit is associated with a control unit. The objective is to design a controller for this control unit such that certain specifications on the parameters of the circuit are satisfied. It is shown that designing a control unit in the form of a switching network is an NP-complete problem that can be formulated as a rank-minimization problem. It is then proven that the underlying design problem can be cast as a semidefinite optimization if a passive network is designed instead of a switching network. Since the implementation of a passive network may need too many components, the design of a decoupled (sparse) passive network is studied subsequently. This chapter introduces a trade-off between the design simplicity and the implementation complexity for an important class of linear circuits. The superiority of the developed techniques is demonstrated by different simulations. In particular, for the first time in the literature, a wavelength-size passive antenna is designed that has an excellent beamforming capability and can make a null in at least 8 directions concurrently.

4.1 Introduction

Many important problems in circuits, electromagnetics, and optics can be reduced to the analysis and synthesis of some linear systems in the frequency domain. These systems, in the circuit theory, consist of passive elements including resistors, inductors, capacitors, ideal transformers, and ideal gyrators [76]. Since the seminal work [15], there has been remarkable progress in characterizing such passive (dissipative) systems using the concept of positive real functions. This notion plays a vital role not only in circuit design but also in various control problems [76, 77, 5].

The application of control theory in circuit and communication areas evidently goes beyond the passivity concept. Indeed, the emerging optimization tools developed by control theorists, such as linear matrix inequalities (LMIs) [13] and sum-of-squares (SOS) [84], have been successfully applied to a number of fundamental problems in these fields. The work [110] is one of the earliest papers connecting the convex optimization theory to circuit design, whose objective is to optimize the dominant time constant of a linear resistor-capacitor circuit using semidefinite programming. The recent paper [36] proposes an LMI optimization to check whether a given multi-port network can be realized using a pre-specified set of linear time-invariant components (namely an inductor and small-signal model of a transistor). Moreover, the work [38] formulates the pattern synthesis of large arrays with bound constraints on the sidelobe and mainlobe levels as a semidefinite programming problem.

Different problems in circuits, electromagnetics, and optics may be formulated as an optimization over the parameters of a multi-port passive network that is obtained, for instance, via an electromagnetic (EM) simulation. As an example, it is shown in [2] that a strikingly efficient and practical way to deal with certain complex antenna problems is to extract a circuit model and then search for appropriate values of its parameters. The circuit model proposed in [2] is indeed a simple, general model that could be considered the abstract model of different types of problems. A question arises as to whether there exists a systematic method to study such circuit problems by means of efficient algorithms. This chapter basically aims to address this question using the available techniques developed in the control theory, especially the LMI and passivity concepts.

Motivated by the papers [2] and [3] on the design of on-chip antennas, a linear multi-port network is considered in this chapter for which certain design specifications on its input admittance and output voltages must be satisfied at a desired frequency. To achieve this, some of the output ports of the network, referred to as *controllable ports*, are connected to a control unit. It is shown that designing a control unit in the form of a switching network that makes the circuit meet the design specifications is an NP-complete problem. Instead, the design of a passive network for the control unit can be cast as a semidefinite

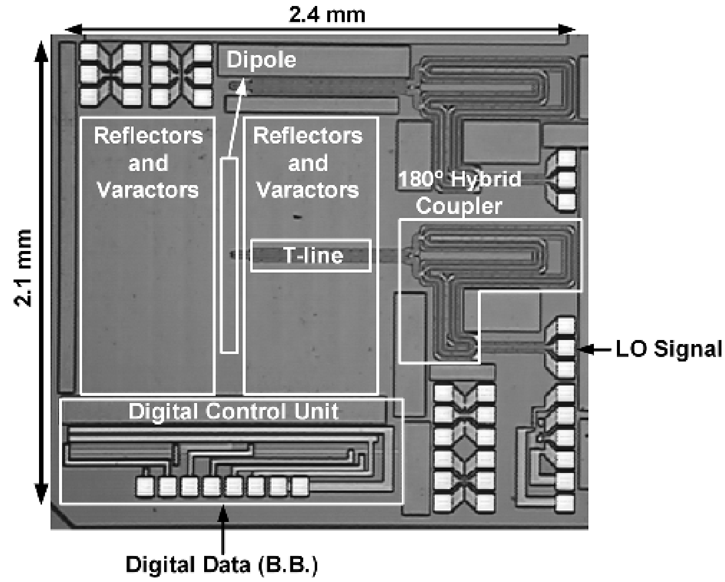


Figure 4.1: This is an implementation of an important antenna configuration whose optimal synthesis can be cast as the problem studied here (see [2] for more details on this chip micrograph).

optimization. Since a passive network may require many components (elements) for its implementation, the design of a sparse (decoupled) passive controller is also studied. To this end, a rank-minimization problem is obtained that can be handled using the convex-based heuristic method proposed in [27] (and further studied in [91]). This heuristic method is able to solve the rank minimization problem correctly in some cases. Note that the main assumption required in this work is the linearity of the given network at the desired frequency, and hence the developed technique is not applicable to nonlinear circuits that cannot be linearized satisfactorily at the frequency of interest.

The techniques developed here are applied to two antenna design problems to demonstrate how optimal antenna configurations with a superior performance can be engineered. In particular, an on-chip wavelength-size passive antenna is designed that can steer the beam to an arbitrary direction and make a null in at least 8 directions simultaneously. This is the first antenna system reported in the literature with such properties that has a significant beamforming capability. Note that the type of the antenna designed here is practically implementable; in particular, we have already implemented a non-optimal antenna with the same structure in [2] leading to the chip micrograph given in Figure 4.1.

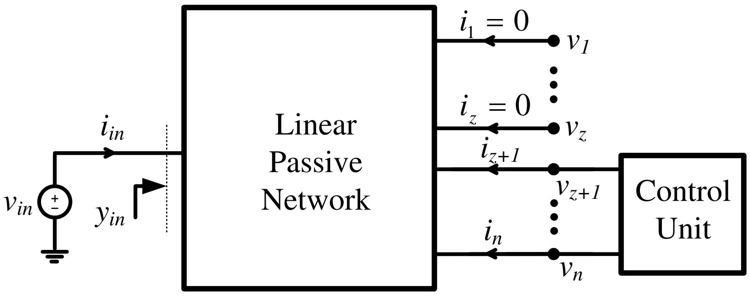


Figure 4.2: Circuit 1 studied in this work.

4.2 Problem Formulation and Motivation

Given a natural number n, consider a linear passive $(n+1)$-port (reciprocal) network, where ports $1, 2, ..., n$ play the role of the outputs of the network and port $n+1$ is the input of the network that is connected to a voltage source with the fixed voltage v_{in}. The output ports of this network are divided into two groups, for a number $z \in \{1, 2, ..., n-1\}$, as follows:

- *Output ports* $1, 2, ..., z$*:* These ports are the output ports of interest, i.e., the ones whose voltages must satisfy some design specifications (linear constraints).

- *Output ports* $z+1, z+2, ..., n$*:* These ports are the *controllable* output ports, i.e., the ones that are connected to a control unit and must be controlled in such a way that the output voltages at ports $1, 2..., z$ as well as the input admittance of the network at port $n+1$ satisfy certain linear specifications.

Since the output ports $1, 2, ..., z$ will not be connected to any device/controller and are used to only measure their voltages, the current through each of these ports must be zero.

The circuit corresponding to the above configuration is shown in Figure 4.2, which will be referred to as *Circuit 1* throughout this chapter. As shown in the figure, let v_p and i_p denote the voltage and current of port p, respectively, for every $p \in \{1, 2, ..., n\}$; moreover, let i_{in} be the current at port $n+1$ and y_{in} be the input admittance of the linear network. To be more specific about the objective of the present work, consider a desired frequency ω_0. The goal is to design a controller for the control unit so that the parameters of Circuit

1 at the frequency ω_0 satisfy the design specifications

$$\left|\mathrm{Re}\left\{v_j - v_j^d\right\}\right| \leq \varepsilon_j, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.1a}$$

$$\left|\mathrm{Im}\left\{v_j - v_j^d\right\}\right| \leq \bar{\varepsilon}_j, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.1b}$$

$$\left|\mathrm{Re}\left\{y_{\mathrm{in}} - y_{\mathrm{in}}^d\right\}\right| \leq \varepsilon, \tag{4.1c}$$

$$\left|\mathrm{Im}\left\{y_{\mathrm{in}} - y_{\mathrm{in}}^d\right\}\right| \leq \bar{\varepsilon}, \tag{4.1d}$$

$$i_j = 0, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.1e}$$

where the operators Re$\{\cdot\}$ and Im$\{\cdot\}$ return the real and imaginary parts of a complex number and

- $v_1^d, ..., v_z^d$ are the given desired voltages for output ports $1, ..., z$, respectively.
- y_{in}^d is the desired input admittance.
- $\varepsilon_j, \bar{\varepsilon}_j$, $\forall j = 1, 2, ..., z$, and $\varepsilon, \bar{\varepsilon}$ are arbitrary nonnegative numbers.

The primary objective of this work is to study the design of different types of control units—such as switching, passive, and decoupled passive controllers—for Circuit 1 and then investigate the trade-off between the design simplicity and the implementation complexity for each of these types.

Note that the circuit being studied here is assumed to be passive and connected to only one voltage source. However, the results of this work can be generalized to the case when there are more than one voltage (current) source and, besides, certain active elements exist in the circuit.

4.2.1 Simple Illustrative Example

Although the main motivation of the present work is the synthesis of circuits derived from electromagnetic structures, it is helpful to illustrate how some generic circuit problems may be modeled as Circuit 1. To this end, consider the simple filter drawn in Figure 4.3. Assume that the goal is to find the numerical values of the impedances Z_1 to Z_5 in such a way that the input-output gain of the filter is maximized at a pre-specified frequency ω_0. To this end, one can re-organize the elements of this filter to obtain the equivalent model given in Figure 4.4, where the known elements are clustered in the block “linear passive network”

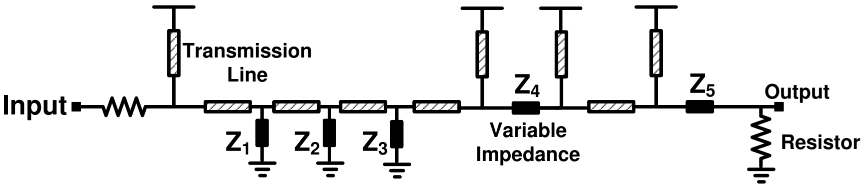


Figure 4.3: A distributed circuit with variable impedances

and the unknown components are grouped in the block "control unit". Under this setting, the objective reduces to designing the control unit in Figure 4.4 so that the magnitude of the observed output of the circuit is maximized. Three points can be made here as follows:

- The control unit in Figure 4.4 is highly structured in the sense that its seven terminal ports are connected to each other in a particular way by the elements Z_1 to Z_5. It will be later explained in Section 4.3.6 how to design a control unit with a prescribed structure.

- The linear passive network in Figure 4.4 has some distributed elements, namely transmission lines. However, they can be replaced by their lumped models at the frequency ω_0.

- As a generalization to the feasibility problem defined earlier by (4.1), one can also maximize some quantity of interest in addition to imposing the constraints given in (4.1). For example, the magnitude of the observed output of the circuit can be maximized (this is explained in Section 4.3.6).

4.2.2 Motivation

Numerical methods and efficient optimization techniques, enabled by increasing computational power, have been markedly instrumental in advancing the field of modern electrodynamics. The progress in this field that was limited to the development of analytical models for antenna characteristics such as pattern, efficiency, and impedance, has been greatly influenced by novel numerical techniques in time or frequency domains. Frequency domain techniques such as finite element method [50] and method of moments [35], as well as time domain algorithms such as finite difference technique [56], have been extensively used in designing electromagnetic structures. These numerical methods combined with optimization techniques such as genetic algorithm [90] and particle swarm optimization [93] provide

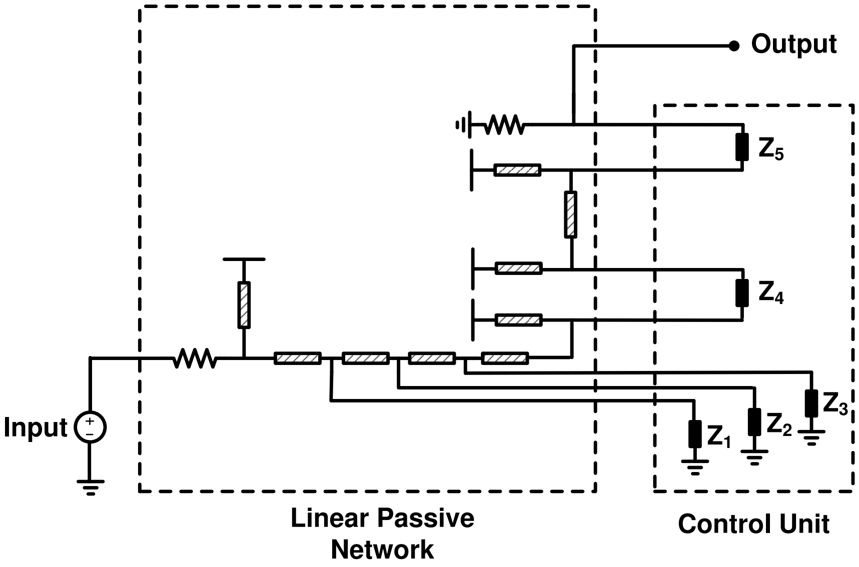


Figure 4.4: The filter given in Figure 4.3 is redrawn in the from of Circuit 1.

a valuable, but inefficient, tool in designing large-scale electromagnetic structures where thousands of passive elements are involved. Indeed, the available numerical techniques iteratively search for a sub-optimal solution. Since a new time-consuming EM simulation needs to be run at each iteration, this approach could be really prohibitive, due to the exponential number of iterations.

In the recent paper [2], this crucial issue is partially resolved by introducing a novel method, which requires performing the EM simulation only once to extract the equivalent circuit model of the system at a single frequency of interest. The electromagnetic problem then reduces to solving a non-iterative optimization problem over the parameters of this circuit model. It is noteworthy that this circuit model is in the form of Circuit 1, in which ports $1, 2, ..., z$ correspond to receiving antennas at the far field, and ports $z + 1, ..., n$ correspond to the controllable ports on the transmitting antenna. Now, the ports $z + 1, ..., n$ on the transmitting antenna should be controlled in such a way that desired voltages are received in the far field at the receiving antennas $1, 2, ..., z$. Roughly speaking, many problems governed by Maxwell's differential equations seeking optimal values of the termination impedances/voltages can be converted to the circuit problem introduced above.

4.2.3 Related Work

The work [110] studies a linear resistor-capacitor (RC) circuit described by the differential equation

$$C\frac{d\mathbf{v}(t)}{dt} = -G(\mathbf{v}(t) - \mathbf{u}(t)), \tag{4.2}$$

where $C \in \mathbf{R}^{n\times n}$ and $G \in \mathbf{R}^{n\times n}$ are symmetric positive-definite capacitance and conductance matrices to be designed, $\mathbf{v}(t) \in \mathbf{R}^n$ is a vector of node voltages and $\mathbf{u}(t) \in \mathbf{R}^n$ is a vector of independent voltage sources. Let $\mathbf{x} := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ be a vector of unknown design parameters and assume that the matrices C and G being sought are required to depend affinely on $\mathbf{x}$, i.e.,

$$C = C_0 + C_1x_2 + \cdots + C_nx_n,$$
$$G = G_0 + G_1x_2 + \cdots + G_nx_n,$$

where $C_0, ..., C_n, G_0, .., G_n$ are some given matrices. It is shown in [110] that the problem of finding the parameter vector $\mathbf{x}$ in such a way that the dominant time constant of the circuit (4.2) is optimized can be cast as a semidefinite programming problem. The present work deals with another type of circuit problem, which is more complicated than the one tackled in [110]. The reason is that the control unit to be designed for Circuit 1 may not be characterizable as an affine function of the design parameters $v_1, v_2, ..., v_z$ and y_{in}. However, it will be shown here that the underlying problem can also be cast as a semidefinite programming problem.

4.3 Main Results

Different types of control units will be designed for Circuit 1 in the following subsections.

4.3.1 Switching Control Unit

Motivated by the antenna application [2] discussed earlier, the most desirable (and simplest) type of control unit is likely a switching controller, which connects every port $p \in \{z+1, ..., n\}$ to an ideal switch that is either on or off (the switch connected to port p is called *switch* p). This is shown in Figure 4.5 and the corresponding circuit is referred to as *Circuit 2*. The problem being addressed here is formalized next.

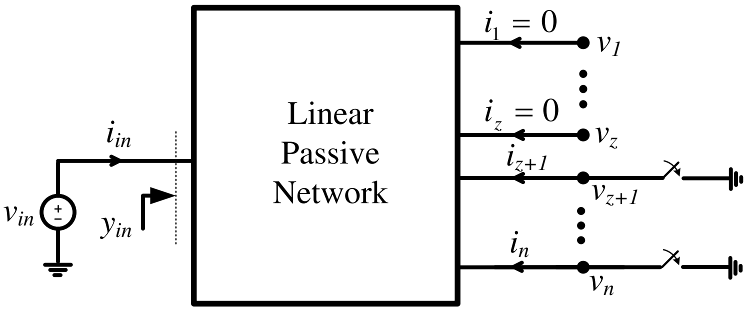


Figure 4.5: Circuit 2 obtained from Circuit 1 by using a switching control unit

Problem 1: Find whether it is possible to turn on a subset of switches $z+1, z+2, ..., n$ in Circuit 2 so that the design specifications given in (4.1) are all satisfied

To analyze Circuit 2, introduce the shorthand notations

$$\mathbf{i} = \begin{bmatrix} i_1 & i_2 & \cdots & i_n \end{bmatrix},$$
$$\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}.$$

One can write a number of equations as

$$\begin{bmatrix} \mathbf{i} & i_{\text{in}} \end{bmatrix} = \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix} Y_s, \tag{4.3a}$$

$$i_{\text{in}} = y_{\text{in}} v_{\text{in}}, \tag{4.3b}$$

$$i_j = 0, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.3c}$$

where Y_s denotes the admittance transfer function of the linear, passive $(n+1)$-port network (the middle block in the circuit) at the given frequency ω_0, or equivalently the Y-parameter matrix of the network at the frequency ω_0. Note that Y_s is a complex-valued matrix whose real and imaginary parts are both symmetric.

Denote the set of complex numbers with $\mathbf{C}$. Define $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ and $\{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, ..., \tilde{\mathbf{e}}_{n+1}\}$ to be the sets of standard basis vectors of $\mathbf{R}^n$ and $\mathbf{R}^{n+1}$, respectively. Throughout this chapter, the notation $\succ$ is used to show matrix inequalities in the positive definite sense. The symbol "*" is also used to denote the conjugate transpose of a matrix. The following theorem recasts Problem 1 as an optimization problem.

Theorem 1 *Minimize the rank of the matrix*

$$\begin{bmatrix} X & \begin{bmatrix} \mathbf{u}^* \\ v_{in} \end{bmatrix} \\ \begin{bmatrix} \mathbf{u} & v_{in} \end{bmatrix} & 1 \end{bmatrix} \tag{4.4}$$

for the variables $X \in \mathbf{C}^{(n+1)\times(n+1)}$ and $\mathbf{u} \in \mathbf{C}^{1\times n}$ subject to the constraints

$$\left|Re\left\{\mathbf{u}\mathbf{e}_j - v_j^d\right\}\right| \leq \varepsilon_j, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.5a}$$

$$\left|Im\left\{\mathbf{u}\mathbf{e}_j - v_j^d\right\}\right| \leq \bar{\varepsilon}_j, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.5b}$$

$$\left|Re\left\{\left[v_{in}^{-1}\mathbf{u} \ \ 1\right] Y_s \tilde{\mathbf{e}}_{n+1} - y_{in}^d\right\}\right| \leq \varepsilon, \tag{4.5c}$$

$$\left|Im\left\{\left[v_{in}^{-1}\mathbf{u} \ \ 1\right] Y_s \tilde{\mathbf{e}}_{n+1} - y_{in}^d\right\}\right| \leq \bar{\varepsilon}, \tag{4.5d}$$

$$\begin{bmatrix} \mathbf{u} & v_{in} \end{bmatrix} Y_s \tilde{\mathbf{e}}_j = 0, \quad \forall j \in \{1, 2, ..., z\}, \tag{4.5e}$$

$$\mathbf{x}_j Y_s \tilde{\mathbf{e}}_j = 0, \qquad \forall j \in \{z+1, ..., n\}, \tag{4.5f}$$

$$X = X^*, \tag{4.5g}$$

where $\mathbf{x}_j$ denotes the j^{th} row of the matrix X. Problem 1 is feasible if and only if the value of the minimum rank is equal to 1, in which case a feasible solution can be extracted as follows: for every $j \in \{z+1, ..., n\}$, turn on switch j if and only if the j^{th} entry of $\mathbf{u}$ is zero.

Proof of necessity: Assume that Problem 1 has a feasible solution. Let k denote the number of switches whose connection makes the design specifications given in (4.1) be satisfied. Denote the set of such switches with $\{p_1, p_2, ..., p_k\} \subseteq \{z+1, ..., n\}$. The goal is to construct a matrix $X \in \mathbf{C}^{(n+1)\times(n+1)}$ and a vector $\mathbf{u} \in \mathbf{C}^{1\times n}$ for which the rank of the matrix (4.4) is 1 and, in addition, the constraints in (4.5) are all satisfied. To this end, consider Circuit 2 with switches $p_1, p_2, ..., p_k$ turned on (and the remaining switches turned off). One can write

$$\begin{aligned} &v_j = 0, \quad \forall j \in \{p_1, p_2, ..., p_k\}, \\ &i_j = 0, \quad \forall j \in \{z+1, z+2, ..., n\} \backslash \{p_1, p_2, ..., p_k\}. \end{aligned}$$

This implies that

$$v_j^* i_j = 0, \quad \forall j \in \{z+1, z+2, ..., n\}. \tag{4.6}$$

On the other hand, it follows from (4.3a) that

$$i_j = \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix} Y_s \tilde{\mathbf{e}}_j, \quad \forall j \in \{1, 2, ..., n\}. \tag{4.7}$$

The equation (4.7) can be substituted into (4.6) to obtain

$$v_j^* \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix} Y_s \tilde{\mathbf{e}}_j = 0, \quad \forall j \in \{z+1, z+2, ..., n\}. \tag{4.8}$$

Define

$$\mathbf{u} := \mathbf{v}, \quad X := \begin{bmatrix} \mathbf{v}^* \\ v_{\text{in}} \end{bmatrix} \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix}. \tag{4.9}$$

The constraints given in (4.5) are all satisfied for this particular choice of X and $\mathbf{u}$, because of the following observations:

- In light of the relations

$$v_j = \mathbf{v}\mathbf{e}_j, \quad \forall j \in \{1, 2, ..., z\},$$
$$y_{\text{in}} = v_{\text{in}}^{-1} i_{\text{in}} = v_{\text{in}}^{-1} \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix} Y_s \tilde{\mathbf{e}}_{n+1},$$

  the constraints (4.5a), (4.5b), (4.5c), and (4.5d) in Theorem 1 correspond to the design specifications (4.1a), (4.1b), (4.1c), and (4.1d), respectively, which are already assumed to hold when switches $p_1, p_2, ..., p_k$ are turned on.

- The constraint (4.5e) corresponds to the design specification (4.1e) (due to the equality (4.7)).

- The constraint (4.5f) corresponds to the relation (4.8) on noting that

$$\mathbf{x}_j = v_j^* \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix}, \quad \forall j \in \{1, 2, ...n\}.$$

- The condition $X = X^*$ given in (4.5g) holds due to the definition of the matrix X in (4.9) as

$$X = \begin{bmatrix} \mathbf{v}^* \\ v_{\text{in}} \end{bmatrix} \begin{bmatrix} \mathbf{v} & v_{\text{in}} \end{bmatrix}.$$

- The rank of the matrix provided in (4.4) is equal to 1 in light of the vector decomposition

$$\begin{bmatrix} X & \begin{bmatrix} \mathbf{u}^* \\ v_{\text{in}} \end{bmatrix} \\ \begin{bmatrix} \mathbf{u} & v_{\text{in}} \end{bmatrix} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{v}^* \\ v_{\text{in}} \\ 1 \end{bmatrix} \begin{bmatrix} \mathbf{v} & v_{\text{in}} & 1 \end{bmatrix}.$$

Proof of sufficiency: Assume that there exist a matrix $X \in \mathbf{C}^{(n+1)\times(n+1)}$ and a vector $\mathbf{u} \in \mathbf{C}^{1\times n}$ such that the rank of the matrix (4.4) is equal to 1 and that the constraints in (4.5) are all satisfied. Identify every index $j \in \{z+1, ..., n\}$ for which the j^{th} entry of $\mathbf{u}$ is zero, and denote the set of all such indices as $\{p_1, p_2, ..., p_k\}$. The intent is to prove that Problem 1 is feasible, and indeed the design specifications (4.1) are satisfied for Circuit 2 when switches $p_1, p_2, ..., p_k$ are turned on. To this end, consider the matrix

$$\begin{bmatrix} X & \begin{bmatrix} \mathbf{u}^* \\ v_{\text{in}} \end{bmatrix} \\ \begin{bmatrix} \mathbf{u} & v_{\text{in}} \end{bmatrix} & 1 \end{bmatrix} \tag{4.10}$$

whose rank is assumed to be 1. Since X satisfies the constraint (4.5g), this matrix is Hermitian. Due to the above matrix being both Hermitian and rank 1, one can apply the singular-value-decomposition theorem to this matrix to infer that there exists a vector $\boldsymbol{\alpha} \in \mathbf{C}^{n+2}$ such that this matrix is equal to either $\boldsymbol{\alpha}\boldsymbol{\alpha}^*$ or $-\boldsymbol{\alpha}\boldsymbol{\alpha}^*$. However, the last diagonal entry of the matrix (4.10) being equal to 1 does not allow this matrix to be equal to the negative semidefinite matrix $-\boldsymbol{\alpha}\boldsymbol{\alpha}^*$. Hence

$$\begin{bmatrix} X & \begin{bmatrix} \mathbf{u}^* \\ v_{\text{in}} \end{bmatrix} \\ \begin{bmatrix} \mathbf{u} & v_{\text{in}} \end{bmatrix} & 1 \end{bmatrix} = \boldsymbol{\alpha}\boldsymbol{\alpha}^*.$$

This relation can be simplified to obtain

$$\boldsymbol{\alpha}^* = \pm \begin{bmatrix} \mathbf{u} & v_{\text{in}} & 1 \end{bmatrix}.$$

As a result, X satisfies the equation

$$X = \begin{bmatrix} \mathbf{u}^* \\ v_{\text{in}} \end{bmatrix} \begin{bmatrix} \mathbf{u} & v_{\text{in}} \end{bmatrix}. \tag{4.11}$$

Define now

$$\tilde{\mathbf{u}} := \begin{bmatrix} \mathbf{u} & v_{\text{in}} \end{bmatrix} Y_s \tag{4.12}$$

and denote the j^{th} entries of $\mathbf{u}$ and $\tilde{\mathbf{u}}$ with u_j and $\tilde{u}_j$, respectively, for every $j \in \{1, 2, ..., n\}$. The equality (4.5e) yields

$$\tilde{u}_j = 0, \quad \forall j \in \{1, 2, ..., z\}.$$

Likewise, the equations (4.5f), (4.11), and (4.12) lead to

$$u_j^* \tilde{u}_j = 0, \quad \forall j \in \{z+1, z+2, ..., n\}$$

or

$$\tilde{u}_j = 0, \quad \forall j \in \{z+1, z+2, ..., n\} \backslash \{p_1, p_2, ..., p_k\}$$

(because u_j is assumed to be nonzero if $j \in \{z+1, z+2, ..., n\} \backslash \{p_1, p_2, ..., p_k\}$). So far, it is shown that there are two vectors $\mathbf{u} \in \mathbf{C}^{1\times n}$ and $\tilde{\mathbf{u}} \in \mathbf{C}^{1\times(n+1)}$ such that

- The relation $\tilde{\mathbf{u}} = \begin{bmatrix} \mathbf{u} & v_{\text{in}} \end{bmatrix} Y_s$ holds.
- $\tilde{u}_j$ is equal to 0 for every $j \in \{1, 2, ..., z\}$.
- u_j is equal to 0 for every $j \in \{p_1, p_2, ..., p_k\}$.
- $\tilde{u}_j$ is equal to 0 for every $j \in \{z+1, z+2, ..., n\} \backslash \{p_1,\ p_2, ..., p_k\}$.

It can be concluded from these properties and the set of equations in (4.3) that

$$\mathbf{u} = \mathbf{v}, \quad \tilde{\mathbf{u}} = \begin{bmatrix} \mathbf{i} & i_{\text{in}} \end{bmatrix},$$

where $\mathbf{v}$, $\mathbf{i}$, and i_{in} are the parameters of Circuit 2 when switches $p_1, p_2, ..., p_k$ are turned on. Now, notice that the design specifications (4.1a), (4.1b), (4.1c), (4.1d), and (4.1e) are equivalent to (4.5a), (4.5b), (4.5c), (4.5d), and (4.5e) in Theorem 1, respectively (see the proof of necessity for an explanation of this equivalency). Hence, the design specifications are satisfied for this particular switching in Circuit 2. ■

Theorem 1 states that Problem 1 is tantamount to an optimization problem whose constraints are all linear. However, the rank of a Hermitian matrix is to be minimized, which makes the problem non-convex. Since a rank-minimization problem is NP-hard in

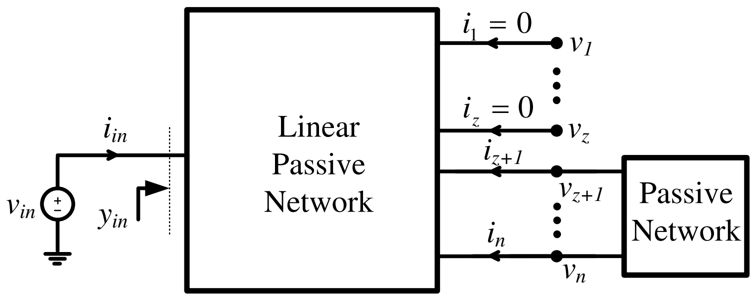


Figure 4.6: Circuit 3 obtained from Circuit 1 by using a linear, passive control unit

general, there may not be an efficient algorithm to solve it exactly. The possibility of using a heuristic method to solve this problem will be later discussed in Section 4.3.4.

A question arises as to whether it is possible to convert Problem 1 to another optimization problem that can be solved efficiently using deterministic algorithms (rather than randomized or heuristic algorithms). This question is tackled in the appendix, where it is shown that Problem 1 is NP-complete, which makes it one of the hardest problems from the computational point of view. An intuitive argument for the NP-completeness of Problem 1 is as follows: *the constraint that each controllable port must be connected to an ideal switch can be interpreted as the input power of each port must be exactly zero. Since the power is a nonconvex fucntion of the voltage and current parameters, deciding whether there are appropriate voltage and current values to make several power terms be precisely equal to zero becomes a hard problem.*

We wish to study how Problem 1 can be modified slightly so that it becomes convex. This is the crux of the next subsection.

4.3.2 Passive Control Unit

The non-convexity of Problem 1 originates from the fact that the output ports $z+1, z+2, ..., n$ are controlled by ideal switches. In this part, let the control unit in Circuit 1 be a general linear, strictly passive network, as opposed to a switching network. This leads to Circuit 3 shown in Figure 4.6. Henceforth, assume that the network corresponding to the admittance Y_s is *strictly* passive (rather than being only passive). The objective of this subsection is formalized in the following.

Problem 2: Find whether it is possible to design a control unit in the form of a linear, strictly passive (reciprocal) network such that the design specifications given in (4.1) are met for Circuit 3.

Let Y denote the admittance of the linear, strictly passive network being designed at the given frequency ω_0. Note that the reciprocity condition in the above problem can be translated as the real and imaginary parts of Y are both symmetric. It is aimed to show that Problem 2 can be turned into a convex optimization problem with a simple form. In what follows, a lemma is presented that will be used later to prove this important result.

Lemma 1 *Given symmetric matrices $M, N \in \mathbf{R}^{n\times n}$, if M is nonsingular, then the following statements are equivalent:*

i) M is a positive definite matrix.

ii) $M + NM^{-1}N$ is a positive definite matrix.

Proof: First, assume that M is a positive define matrix. Thus, M^{-1} is positive definite and so is $NM^{-1}N$. This implies that $M + NM^{-1}N$ is a positive definite matrix. So far, it is shown that (i) implies (ii). To complete the proof, it remains to show that the converse statement is also true. To this end, assume that $M + NM^{-1}N$ is a positive definite matrix. Define the matrices

$$
\begin{aligned}
P &:= \begin{bmatrix} M & N \\ N & -M \end{bmatrix}, \\
T &:= \begin{bmatrix} I & -NM^{-1} \\ 0 & I \end{bmatrix}, \\
Q &:= \begin{bmatrix} M + NM^{-1}N & 0 \\ 0 & -M \end{bmatrix}.
\end{aligned} \tag{4.13}
$$

It is easy to verify that $P = TQT^*$. Denote the number of positive, negative, and zero eigenvalues of the symmetric matrix P with η_1, η_2, η_3, respectively. Analogously, denote the same quantities of the matrix Q with the triple $(\bar{\eta}_1, \bar{\eta}_2, \bar{\eta}_3)$. Since the matrix T is nonsingular, applying Sylvester's Law of Inertia to the relation $P = TQT^*$ yields

$$(\eta_1, \eta_2, \eta_3) = (\bar{\eta}_1, \bar{\eta}_2, \bar{\eta}_3). \tag{4.14}$$

On the other hand, it can be concluded from the Hamiltonian structure of the matrix P that

$$\eta_1 = \eta_2. \tag{4.15}$$

Furthermore, since every eigenvalue of $M+NM^{-1}N$ is an eigenvalue of Q and all eigenvalues of $M+NM^{-1}N$ are positive, the quantity $\bar{\eta}_1$ is at least equal to n. In light of the equalities (4.14) and (4.15), the relation $\bar{\eta}_1 \geq n$ is possible only if $\eta_1 = \eta_2 = \bar{\eta}_1 = \bar{\eta}_2 = n$. Thus, the matrix Q has n negative eigenvalues. Nonetheless, the negative eigenvalues of this matrix are the same as those of the matrix $-M$; hence, $-M \in \mathbf{R}^{n\times n}$ has the maximum number of negative eigenvalues. This simply proves that the eigenvalues of M are all positive, which completes the proof. ■

Decompose the matrix Y_s in a block form as

$$Y_s = \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix},$$

where $W_{11} \in \mathbf{C}^{z\times z}$, $W_{22} \in \mathbf{C}^{(n-z)\times(n-z)}$ and $W_{33} \in \mathbf{C}$. For given symmetric square matrices A and B of the same dimension with $\det(A) \neq 0$, it can be verified that

$$(A + B\mathrm{i})^{-1} = (A + BA^{-1}B)^{-1} - (A + BA^{-1}B)^{-1}BA^{-1}\mathrm{i}, \tag{4.16}$$

where "i" stands for the imaginary unit. This identity will be exploited in the next theorem.

Theorem 2 *Problem 2 is feasible if and only if there exist symmetric matrices $M, N \in \mathbf{R}^{(n-z)\times(n-z)}$ and vectors $\mathbf{u}_1 \in \mathbf{C}^{1\times z}$, $\mathbf{u}_2 \in \mathbf{C}^{1\times(n-z)}$ such that*

$$\begin{bmatrix} \left(Re\left\{W_{22} - W_{21}W_{11}^{-1}W_{12}\right\}\right)^{-1} - M & N \\ N & M \end{bmatrix} \succ 0 \tag{4.17}$$

and that

$$\begin{aligned}
&\left|Re\left\{\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}e_j - v_j^d\right\}\right| \le \varepsilon_j,\ \forall j \in \{1,2,...,z\}, && (4.18a)\\
&\left|Im\left\{\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}e_j - v_j^d\right\}\right| \le \bar{\varepsilon}_j,\ \forall j \in \{1,2,...,z\}, && (4.18b)\\
&\left|Re\left\{v_{in}^{-1}\mathbf{u}_1 W_{13} + v_{in}^{-1}\mathbf{u}_2 W_{23} + W_{33} - y_{in}^d\right\}\right| \le \varepsilon, && (4.18c)\\
&\left|Im\left\{v_{in}^{-1}\mathbf{u}_1 W_{13} + v_{in}^{-1}\mathbf{u}_2 W_{23} + W_{33} - y_{in}^d\right\}\right| \le \bar{\varepsilon}, && (4.18d)\\
&\mathbf{u}_1 = -\mathbf{u}_2 W_{21} W_{11}^{-1} - v_{in} W_{31} W_{11}^{-1}, && (4.18e)\\
&\mathbf{u}_2 = v_{in}(W_{31} W_{11}^{-1} W_{12} - W_{32})(M + Ni). && (4.18f)
\end{aligned}$$

Moreover, if there exist such matrices M, N satisfying the above constraints, then one candidate for the admittance matrix Y is

$$Y = (M + Ni)^{-1} - W_{22} + W_{21} W_{11}^{-1} W_{12}. \tag{4.19}$$

Proof of necessity: Assume that there exists a linear, passive controller (control unit) with an admittance Y at the frequency ω_0 such that the design specifications listed in (4.1) are satisfied for Circuit 3 under this controller. The objective is to prove that there exist symmetric matrices $M, N \in \mathbf{R}^{(n-z)\times(n-z)}$ and vectors $\mathbf{u}_1 \in \mathbf{C}^{1\times z}$, $\mathbf{u}_2 \in \mathbf{C}^{1\times(n-z)}$ for which the constraints given in (4.17) and (4.18) are satisfied. For this purpose, consider Circuit 3 under the passive network Y and define the vectors

$$\begin{aligned}
\mathbf{v}_1 &= \begin{bmatrix} v_1 & v_2 & \cdots & v_z \end{bmatrix},\\
\mathbf{v}_2 &= \begin{bmatrix} v_{z+1} & v_{z+2} & \cdots & v_n \end{bmatrix}.
\end{aligned}$$

Two equations can be written for Circuit 3 as follows:

$$\begin{aligned}
\begin{bmatrix}\mathbf{i} & i_{\text{in}}\end{bmatrix} &= \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & v_{\text{in}}\end{bmatrix} Y_s,\\
\begin{bmatrix}\mathbf{i} & i_{\text{in}}\end{bmatrix} &= -\begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & v_{\text{in}}\end{bmatrix} \begin{bmatrix} 0 & 0 & 0\\ 0 & Y & 0\\ 0 & 0 & -y_{\text{in}} \end{bmatrix}.
\end{aligned} \tag{4.20}$$

These equations can be combined to obtain

$$\begin{aligned}\mathbf{v}_1 W_{11} + \mathbf{v}_2 W_{21} + v_{\text{in}} W_{31} &= 0,\\ \mathbf{v}_1 W_{12} + \mathbf{v}_2 (W_{22} + Y) + v_{\text{in}} W_{32} &= 0,\\ \mathbf{v}_1 W_{13} + \mathbf{v}_2 W_{23} + v_{\text{in}} (W_{33} - y_{\text{in}}) &= 0.\end{aligned}$$

The above relations can be manipulated to arrive at

$$\mathbf{v}_1 = -\mathbf{v}_2 W_{21} W_{11}^{-1} - v_{\text{in}} W_{31} W_{11}^{-1}, \tag{4.21a}$$

$$\mathbf{v}_2 = v_{\text{in}} \big(W_{31} W_{11}^{-1} W_{12} - W_{32} \big) \tilde{Y}, \tag{4.21b}$$

$$y_{\text{in}} = v_{\text{in}}^{-1} \mathbf{v}_1 W_{13} + v_{\text{in}}^{-1} \mathbf{v}_2 W_{23} + W_{33}, \tag{4.21c}$$

where

$$\tilde{Y} := \big(W_{22} - W_{21} W_{11}^{-1} W_{12} + Y \big)^{-1}. \tag{4.22}$$

Note that the invertibility of the term $W_{22} - W_{21} W_{11}^{-1} W_{12} + Y$ follows from the strict passivity of Y and Y_s. It is desired to show that the constraints (4.17) and (4.18) in Theorem 2 hold if M, N, $\mathbf{u}_1$, and $\mathbf{u}_2$ are defined as

$$M := \text{Re}\{\tilde{Y}\}, \quad N := \text{Im}\{\tilde{Y}\}, \quad \mathbf{u}_1 := \mathbf{v}_1, \quad \mathbf{u}_2 := \mathbf{v}_2.$$

To this end, first observe that M and N are symmetric matrices due to the reciprocity of Y and Y_s. Besides, it can be concluded from the above definitions and (4.21c) that the constraints (4.18a), (4.18b), (4.18c), and (4.18d) correspond to the design specifications (4.1a), (4.1b), (4.1c), and (4.1d), respectively, which are assumed to hold for Circuit 3. The constraints (4.18e) and (4.18f), on the other hand, are satisfied in light of the relations (4.21a) and (4.21b). The only challenging part is to show that the inequality (4.17) in Theorem 2 holds. For this purpose, notice that the strict passivity of the network associated with Y implies the relation $\text{Re}\{Y\} \succ 0$ [76]. On applying the identity (4.16) to the equation (4.22) and using this fact, one can write

$$\begin{aligned}\mathrm{Re}\{Y\} &= \mathrm{Re}\{\tilde{Y}^{-1} - W_{22} + W_{21}W_{11}^{-1}W_{12}\} \\ &= \big(M + NM^{-1}N\big)^{-1} \\ &- \mathrm{Re}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\} \succ 0.\end{aligned} \tag{4.23}$$

Since the term $(W_{22} - W_{21}W_{11}^{-1}W_{12})^{-1}$ is the $(1,1)$ block entry of the inverse of the matrix

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

that is a principal submatrix of Y_s, it follows from the strictly passivity of the admittance matrix Y_s that

$$\mathrm{Re}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\} \succ 0. \tag{4.24}$$

The inequalities (4.23) and (4.24) lead to

$$\big(M + NM^{-1}N\big)^{-1} \succ \mathrm{Re}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\} \succ 0. \tag{4.25}$$

Two properties can be deduced from this relation as follows:

- First, Lemma 1 yields

  $$M \succ 0. \tag{4.26}$$

- Second, the inequality (4.25) can be re-arranged to obtain

  $$\Big(\mathrm{Re}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\}\Big)^{-1} \succ M + NM^{-1}N$$

  or equivalently

  $$\Big(\mathrm{Re}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\}^{-1} - M\Big) - NM^{-1}N \succ 0. \tag{4.27}$$

Schur's complement formula can be used to conclude that the inequalities (4.26) and (4.27) are equivalent to (4.17). This completes the proof of necessity.

Proof of sufficiency: Since the proof can be carried out in line with the approach taken at the proof of necessity, only a sketch of the proof will be provided here. Assume that the constraints given in (4.17) and (4.18) are satisfied for some symmetric matrices $M, N \in \mathbf{R}^{(n-z)\times(n-z)}$ and vectors $\mathbf{u}_1 \in \mathbf{C}^{1\times z}$, $\mathbf{u}_2 \in \mathbf{C}^{1\times(n-z)}$. The goal is to show that the design specifications listed in (4.1) are met for Circuit 3 if the admittance Y of the passive controller (at the frequency ω_0) is considered as

$$Y = (M + N\mathrm{i})^{-1} - W_{22} + W_{21}W_{11}^{-1}W_{12}.$$

For this choice of the matrix Y, it follows from (4.18f) and (4.21b) that $\mathbf{u}_2$ is equal to $\mathbf{v}_2$. Then, it can be concluded from (4.18e) and (4.21a) that $\mathbf{u}_1 = \mathbf{v}_1$. Now, one can easily verify that the design specifications (4.1a), (4.1b), (4.1c), and (4.1d) correspond to the inequalities (4.18a), (4.18b), (4.18c), and (4.18d), respectively, which are assumed to hold. On the other hand, the design specification (4.1e) is satisfied in light of the relation (4.18e) and the equality

$$\begin{bmatrix} i_1 & i_2 \cdots & i_z \end{bmatrix} = v_1W_{11} + v_2W_{21} + v_{\mathrm{in}}W_{31}$$

(see (4.20)). Hence, it only remains to show that the matrix Y introduced above corresponds to a strictly passive network. This can be shown using Lemma 1 and Schur's complement formula in line with the argument pursued in the proof of necessity. The details are omitted for brevity. ■

Regarding the optimization problem proposed in Theorem 2, it is easy to observe that the constraints are all linear. Therefore, Theorem 2 states that Problem 2 is equivalent to a linear matrix inequality (LMI) feasibility problem, which can be handled efficiently using a proper software tool such as YALMIP or SOSTOOLS [68, 88]. This signifies that replacing switches with a passive network facilitates the circuit design, at the cost of complicating its implementation in practice. In the case when it is strictly required to design a collection of switches, Theorem 2 is still useful. Indeed, since Circuit 2 is a special form of Circuit 3, the infeasibility of Problem 2 implies the infeasibility of Problem 1. As a result, one can regard the LMI problem proposed in Theorem 2 as a sanity test for checking the feasibility of Problem 1.

Assume that Problem 2 is feasible and, therefore, an admittance matrix Y (at the frequency ω_0) can be obtained by solving the feasibility problem given in Theorem 2. The next step is to design a reciprocal passive network whose corresponding admittance transfer function at the frequency ω_0 is equal to Y. To find such a network, note that the real part of Y is a positive definite matrix and that its imaginary part is symmetric. As a result, the matrix Y can be expressed as

$$Y = T_1 + T_2\mathrm{i},$$

where $T_1, T_2 \in \mathbf{R}^{(n-z)\times(n-z)}$ are both symmetric and T_1 is positive definite. Define an admittance transfer function $Y(s)$ as

$$Y(s) = T_1 + \frac{1}{\omega_0}T_2 s, \quad \forall s \in \mathbf{C}.$$

It is evident that $Y(\mathrm{i}\omega_0) = Y$. On the other hand, $Y(s)$ can be implemented by the parallel connection of two $(n-z)$-port networks: (i) a resistive network with the conductance matrix T_1 and (ii) a reactive network with the susceptance matrix $\frac{1}{\omega_0}T_2$. Note that some ideal transformers might also be needed to realize $Y(s)$ due to the multi-port nature of the network. One can refer to [76] and [15] for detailed discussions on the realization of a given admittance matrix by passive elements.

4.3.3 Decoupled Passive Control Unit

The main issue with the admittance matrix Y obtained in Theorem 2 is that its corresponding passive network could potentially have several components (electrical elements), which may complicate its implementation. To circumvent this drawback, one can impose a sparsity constraint on Y to make it diagonal. Note that Circuit 3 under a passive control unit with a diagonal admittance transfer function is equivalent to *Circuit 4* depicted in Figure 4.7. Alternatively, one can reason that Circuit 4 is obtained from Circuit 2 (as opposed to Circuit 3) by replacing ideal switches with varactors. Define Problem 3 to be the same as Problem 2, but under the additional constraint of the diagonality of Y. It will be shown in the sequel that Problem 3 is non-convex; however, there is a good heuristic method for this problem, as tested on several practical examples.

Theorem 3 *Minimize the rank of the matrix*

$$\begin{bmatrix} \bar{P} & I \\ I & P \end{bmatrix} \tag{4.28}$$

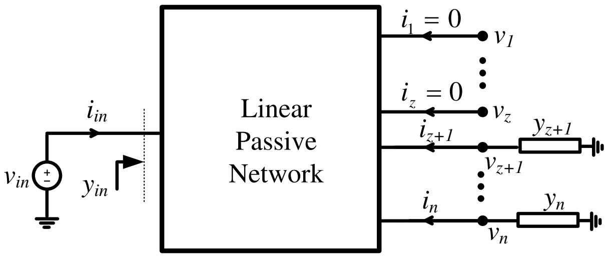


Figure 4.7: Circuit 4 obtained from Circuit 1 by using a decoupled, linear, passive control unit.

for vectors $\mathbf{u}_1 \in \mathbf{C}^{1\times z}$, $\mathbf{u}_2 \in \mathbf{C}^{1\times(n-z)}$, *symmetric matrices* $M, N \in \mathbf{R}^{(n-z)\times(n-z)}$, *and diagonal matrices* $D_1, D_2 \in \mathbf{R}^{(n-z)\times(n-z)}$ *subject to the constraints given in (4.18) and*

$$D_1 > 0,$$

$$M \succ 0,$$

where P *is provided in (4.13) and*

$$\bar{P} := \begin{bmatrix} D_1 + Re\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} & -D_2 - Im\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} \\ -D_2 - Im\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} & -D_1 - Re\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} \end{bmatrix}. \tag{4.29}$$

Problem 3 is feasible if and only if the value of the minimum rank is less than or equal to $2(n-z)$*, in which case a feasible solution for the diagonal admittance matrix* Y *is as follows:*

$$Y = D_1 + D_2 i.$$

Proof: When there is no diagonality constraint on the matrix Y, a necessary and sufficient condition for the existence of a desirable network is provided in Theorem 2. Hence, it suffices to incorporate this extra constraint into the above-mentioned condition. To this end, write Y as $D_1 + D_2\mathrm{i}$, where D_1 and D_2 are required to be diagonal. It results from the equation (4.19) that

$$D_1 + D_2\mathrm{i} + W_{22} - W_{21}W_{11}^{-1}W_{12} = (M + N\mathrm{i})^{-1}. \tag{4.30}$$

Applying the identity (4.16) to the above equation yields

$$D_1 + \mathrm{Re}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\} = \big(M + NM^{-1}N\big)^{-1}$$

and

$$D_2 + \mathrm{Im}\Big\{W_{22} - W_{21}W_{11}^{-1}W_{12}\Big\} = -\big(M + NM^{-1}N\big)^{-1}NM^{-1}.$$

These equations can be written as follows:

$$\begin{bmatrix} D_1 + \mathrm{Re}\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} & -D_2 - \mathrm{Im}\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} \\ -D_2 - \mathrm{Im}\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} & -D_1 - \mathrm{Re}\{W_{22} - W_{21}W_{11}^{-1}W_{12}\} \end{bmatrix} = \begin{bmatrix} M & N \\ N & -M \end{bmatrix}^{-1} \tag{4.31}$$

or equivalently $\bar{P} = P^{-1}$. On the other hand

$$\begin{bmatrix} \bar{P} & I \\ I & P \end{bmatrix} = \begin{bmatrix} I & P^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \bar{P} - P^{-1} & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} I & 0 \\ P^{-1} & I \end{bmatrix}. \tag{4.32}$$

In light of the equality $\bar{P} - P^{-1} = 0$ and the non-singularity of P, it follows from the above equation that the rank of the matrix given in (4.28) is exactly equal to $2(n - z)$. So far, it is shown that the diagonality of the matrix Y implies the afore-mentioned rank constraint. To prove the converse statement, notice that the condition $M \succ 0$ makes the Hamiltonian matrix P nonsingular (see the proof of Lemma 1). This, together with the identity (4.32), implies that if the rank of the matrix in (4.28) is less than or equal to $2(n - z)$, then the matrix $P - \bar{P}^{-1}$ must be zero. This result leads to the equation (4.30), which is indeed a diagonality constraint on the matrix Y. Moreover, one can easily replace the passivity constraint (4.17) given in Theorem 2 with the condition $D_1 > 0$, because the real part of the matrix Y is equal to D_1. ■

Remark 1: Unlike Problem 2 that had a convex formulation, Problem 3 turned out to be a rank-minimization problem that is not convex. A question arises as to what makes Problem 3 hard. To answer this question, consider the special case when the circuit is resistive and the controller to be designed needs to be resistive as well. This particular case makes all complex variables real-valued, which will later reveal the design difficulties. Notice that since each controllable port must be connected to a resistor, the following power constraints should be satisfied:

$$v_j i_j \leq 0, \quad \forall j \in \{z+1, ..., n\}. \tag{4.33}$$

Given $j \in \{z+1, ..., n\}$, this means that one of the cases $v_j \leq 0, i_j \geq 0$ or $v_j \geq 0, i_j \leq 0$ must occur, which implies that there are two possibilities for the parameters (v_j, i_j). Hence, it follows from (4.33) that there are 2^{n-z} possibilities for the parameters $(v_{z+1}, ..., v_n, i_{z+1}, ..., i_n)$. As a result, the above power constraints correspond to a non-convex feasibility region that is composed of an exponential number (2^{n-z}) of convex parts attached to each other at the origin. This highly non-convex feasibility region is the source of difficulty in tackling Problem 3.

4.3.4 Heuristic Method for Rank Minimization

Since the optimization problems given in Theorems 1 and 3 are associated with rank constraints, the objective of this part is to study rank-minimization problems. Consider a standard rank optimization problem in the form of

$$\begin{aligned} \text{minimize} \quad & \text{Rank}(X) \\ \text{subject to} \quad & \mathcal{A}(X) = b, \end{aligned} \tag{4.34}$$

where

- X is an $n_1 \times n_2$ matrix decision variable (n_1 and n_2 are given numbers).
- $\mathcal{A} : \mathbf{R}^{n_1 \times n_2} \to \mathbf{R}^m$ is a known linear map.
- b is a given vector in $\mathbf{R}^m$.

It is known that the optimization problem (4.34) is NP-hard, in general. However, several heuristic methods have been proposed in the literature to relax the problem to a convex one, whose solution may be identical or near to that of the original problem [27, 34]. The heuristic method developed in the papers [27] and [91] has been widely used in the literature. This method suggests solving the following optimization problem instead of (4.34):

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & \mathcal{A}(X) = b, \end{aligned} \tag{4.35}$$

where $\|X\|_*$ denotes the nuclear norm of the matrix X (defined as the sum of the singular values of X). The main advantage of this heuristic method is that the optimization problem (4.35) is convex and, thus, its global solution can be found efficiently.

An important question arises as to when the solutions of optimizations (4.34) and (4.35) coincide. To answer the raised question probabilistically, represent the linear map $\mathcal{A}(X)$ in the matrix form as $A \times \mathrm{vec}(X)$ where $A \in \mathbf{R}^{m \times n_1 n_2}$ is a matrix and $\mathrm{vec}(X)$ is a vector obtained from X by stacking up the columns of X. It is shown in [92] that as m goes to infinity, the probability that the optimizations (4.34) and (4.35) have the same solution tends to 1 if the entries of the matrix A are sampled independently from a zero-mean, unit-variance Gaussian distribution. In other words, the above-mentioned heuristic method works almost always correctly for a standard rank-minimization problem whose linear constraints are randomly generated using a Gaussian probability distribution.

The rank-minimization problems given in Theorems 1 and 3 can be handled using the heuristic method discussed above. As a result, the nuclear norm of the matrices (4.4) and (4.28) should be minimized in the related optimization problems instead of their ranks. In the case when this heuristic method leads to a rank greater than 1 for the optimization problem in Theorem 1 or $2(n-z)$ for the optimization problem in Theorem 3, there are two possibilities: (i) Problem 1 (or Problem 3) is infeasible, (ii) the heuristic method fails to find a solution with the minimum rank. One can use the necessary and sufficient condition derived in [92] to see if case (i) takes place, although this may be complicated. Note that since the optimization problems in Theorems 1 and 3 are highly structured (partially due to the presence of fixed elements 1 and I as well as a Hamiltonian matrix in the constraints), these optimizations may be far from being a Gaussian random instance of a rank-minimization problem. Therefore, they may not lie into the category of problems for which the above-mentioned heuristic method almost always works correctly (note that the results developed in [92] are applicable to these optimizations, because they can be transformed into the standard form (4.34) using the technique delineated therein).

We did an extensive simulation to test the efficiency of the nuclear norm heuristic method on different antenna problems, and made some important observations as follows:

i) The heuristic method works correctly all the time for the optimization problem in Theorem 3 if there are no constraints on $v_1, v_2, ..., v_z$, i.e., if all constraints are on y_{in}.

ii) Some of the design specifications may be violated a little if the heuristic method is applied to the optimization problem of Theorem 3 with some constraints on the output voltages $v_1, v_2, ..., v_z$. For example, given an index $j \in \{1, 2, ..., z\}$, the real part of the obtained voltage v_j that is required to belong to the interval $[\text{Re}\{v_j^d\} - \varepsilon_j, \text{Re}\{v_j^d\} + \varepsilon_j]$ might lie a bit off this range.

iii) The nuclear norm heuristic method often fails to obtain a satisfactory result when applied to Theorem 1.

It is shown in the appendix that Problem 1 is NP-complete and since an NP-complete problem is well-understood to be very hard to solve, it is commonly believed that a convex heuristic method (such as the above-mentioned one) often fails to find a satisfactory solution. This might be the reason for observation (iii).

4.3.5 Design Simplicity Versus Implementation Complexity

It is desired to compare Circuits 2, 3, and 4 in terms of their design and implementation. To this end, the main properties of these circuits can be summarized as follows:

- The implementation of a control unit for Circuit 2 requires only $n - z$ switches, but finding the on/off status of every switch to satisfy the design specifications is an NP-complete problem. As a result, the synthesis of such a circuit can be extremely difficult when the number of switches, i.e., $n - z$, is greater than 30 (because the discrete space of all switching combinations has 2^{n-z} elements that is a very large set if $n - z > 30$).

- The implementation of a control unit for Circuit 3 requires about $0.5(n - z)^2$ components (e.g., resistors, capacitors, inductors). This may make the implementation of such a controller difficult for some applications. Nonetheless, the synthesis of a passive control unit can be converted to a linear matrix inequality feasibility problem, which can be handled efficiently even when $n - z$ is of the order of several thousands.

- The implementation of a control unit for Circuit 4 requires only $n - z$ components. Hence, the number of components in the controller grows linearly with respect to the number of controllable ports (i.e., $n - z$), which is a useful property for large-scale systems. Even though the synthesis of such a controller is tantamount to a rank-minimization problem, it may be solved using a heuristic method (as mentioned earlier) particularly when there are not many constraints on output voltages.

The above discussion leads to the conclusion that since the synthesis of Circuit 2 is very difficult even for moderate-sized systems, it is preferable to deploy either Circuit 3 or Circuit 4. In the case when it is desired to design a control unit online (as demanded in antenna applications due to the periodic change of the design specifications), Circuit 3 is a more suitable choice compared to Circuit 4. However, the implementation of Circuit 4 is much simpler than that of Circuit 3 for large-scale systems.

Remark 2: To reduce the implementation complexity of Circuit 3, it is preferable to use a small subset of the $n-z$ controllable ports, if possible. More specifically, it might be possible to satisfy the design specifications by only controlling a few of the controllable ports. Hence, one can take the following strategy: check whether a passive control unit can be designed for port $z+1$ to satisfy the design objectives (4.1); if not, verify the existence of a controller for ports $z+1$ and $z+2$; continue this procedure up to the point that enough number of controllable ports are found whose passive control meets the design specifications. This heuristic method can significantly reduce the implementation complexity.

4.3.6 Generalizations

Problems 1, 2, and 3 studied in this chapter target a circuit synthesis with the design specifications given in (4.1). However, the techniques developed here can be generalized to incorporate other types of design specifications. For example, assume that an output voltage v_p, $p \in \{1, 2, ..., z\}$, is required to be sufficiently weak, as demanded by antenna applications. This constraint can be formalized as $\|v_p\| \leq \tilde{\varepsilon}$, where $\|\cdot\|$ denotes the 2-norm and $\tilde{\varepsilon}$ is a given positive number. To account for this new design specification, the constraint

$$\|\mathbf{u}\mathbf{e}_p\| \leq \varepsilon$$

should be added to the optimization problem of Theorem 1; likewise, the constraint

$$\left\|\begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \mathbf{e}_p\right\| \leq \varepsilon$$

should be included in the optimization problems of Theorems 2 and 3. As another example, if one needs to design a control unit for Circuit 1 in the form of a decoupled *lossless* network, it suffices to replace the constraint $D_1 > 0$ with $D_1 = 0$ in the optimization problem of Theorem 3.

Unlike Theorems 1 and 3 that propose *minimization* problems, Theorem 2 offers a *feasibility* problem. In other words, there is no specific quantity in the feasibility problem of Theorem 2 that must be minimized (or maximized). This provides a degree of freedom in the underlying circuit synthesis. To be more precise, Theorem 2 can be employed to simultaneously solve Problem 2 and minimize (maximize) some quantity of interest such as the consumed power at a specific port. This point will be illustrated in the next section through some simulations.

As another generalization, assume that the goal is to design a passive control unit with a pre-specified structure. An example of this case is the filter given in Figure 4.4 whose control unit is structured in terms of the impedances Z_1 to Z_5. To handle this problem, it suffices to employ Theorem 3 after the following slight modifications:

- Replace the diagonality requirement of the matrix variables D_1 and D_2 with a desired pattern condition on these matrices, say certain entries of these matrices must be zero according to the desired structure of the control unit being designed.

- Replace the condition $D_1 > 0$ with the general passivity constraint (4.17).

4.4 Simulation Results

To illustrate the efficacy of the present work in the context of antenna design, note that most of the practical antenna problems deal with the optimization of the input impedance and/or the antenna gain via changing the geometry of the antenna. This is achieved in reality by means of inefficient heuristic algorithms. For instance, a particle swarm optimization technique (PSO) is deployed in [51] to optimize the antenna input impedance by varying its length, width, and feeding point. That algorithm was applied to a simple impedance matching problem with only 3 variables, which consumed more than 25 hours to obtain the solution. This clearly shows that such algorithms are dramatically time-consuming even for very small-sized antenna problems. Two important practical examples will be studied in the sequel to demonstrate that more complicated antenna design problems with 12 and 90 variables can be solved in the order of seconds rather than hours using the method developed here.

Example 1: Consider the antenna configuration depicted in Figure 4.8, which consists of a transmitting dipole antenna (blue bar), a 3x3 array of metal plates (antenna parasitic elements), and a receiving dipole antenna located at the far field (green bar). There are 14 ports in this figure as follows:

- Port 1 acts as a receiving antenna sampling the radiation pattern of the transmitting antenna at a specific angle in the far field.
- Ports 2 to 13 are intended to change the boundary condition of the transmitting antenna.
- Port 14 corresponds to the transmitting antenna.

The objective is to find optimum impedance values for the parasitic elements such that the received power and the antenna input impedance satisfy a specific set of constraints. For this purpose, the circuit model of the antenna system is extracted at the desired frequency 3.5 GHz (using localized differential lumped ports) by means of the electromagnetic software IE3D [41]. This model can be any of the circuits given in Figures 4.5, 4.6, 4.7, depending on how the impedances of the parasitic elements are designed. Note that n and z are equal to 13 and 1, respectively, in this example, and that $v_{n+1} = v_{14} = 1$.

Three important goals in a typical antenna problem are (i) received power maximization, (ii) received power maximization under an input admittance constraint, (iii) input impedance matching. Tackling these problems is central to this example, which is carried out in the sequel.

Considering the complex number v_1 as a real vector in $\mathbf{R}^2$, one can notice that the power at the receiving antenna is proportional to the 2-norm of v_1 raised to the second power. Since the maximization of the 2-norm of a quantity is normally a non-convex problem, it is desired to maximize the 1-norm of v_1, i.e., $|\text{Re}\{v_1\}| + |\text{Im}\{v_1\}|$. This suggestion is motivated by the close affinity between these two norms. Observe that the direct maximization of $|\text{Re}\{v_1\}| + |\text{Im}\{v_1\}|$ is again a non-convex optimization problem. Nevertheless, one can alternatively perform four (convex) optimizations maximizing the quantities $\text{Re}\{v_1\} + \text{Im}\{v_1\}$, $\text{Re}\{v_1\} -$

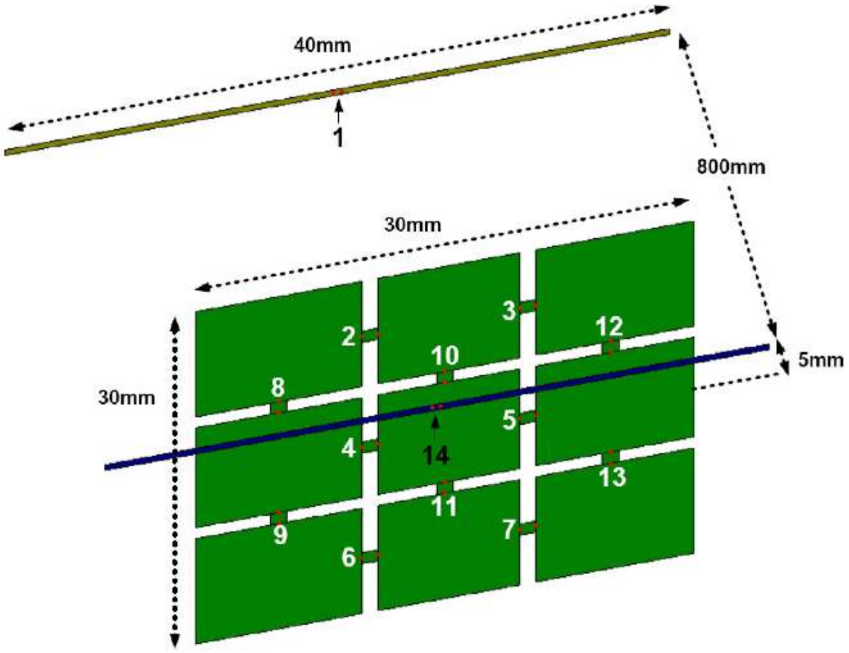


Figure 4.8: The antenna problem studied in Example 1.

$\text{Im}\{v_1\}$, $-\text{Re}\{v_1\} + \text{Im}\{v_1\}$, and $-\text{Re}\{v_1\} - \text{Im}\{v_1\}$, and then determine the maximum of the obtained solutions. Problem 2 is adopted to solve these optimization problems. The outcome of these convex optimization problems is summarized in Table 4.1, which demonstrates that the optimal value of $|\text{Re}\{v_1\}| + |\text{Im}\{v_1\}|$ is equal to 0.2833 corresponding to the antenna directivity of 8.17dBi and the radiation efficiency of 89.15%. It is interesting to note that this result is obtained by solving four convex optimization problems, each of which is handled by the software CVX [33] in a fraction of second (the simulation was run on a computer with a Pentium IV 3.0 GHz and 3.62 GB of memory).

Now, assume that the objective is to maximize the power at the receiving antenna subject to the constraint that the antenna input impedance is equal to the standard value 50Ω. As before, this power is proportional to the 2-norm of the output voltage v_1 raised to the second power. The non-convexity of the underlying problem suggests maximizing the closely related term $|\text{Re}\{v_1\}| + |\text{Im}\{v_1\}|$. Similar to the previous case, four convex optimization problems are solved, and the results are summarized accordingly in Table 4.2.

As the last scenario, the goal is to find a diagonal matrix Y such that the antenna input impedance is matched with the value 50Ω. The heuristic method given in [91] was applied to Problem 3 to find proper values for the diagonal matrices D_1 and D_2 (recall that

Table 4.1: Maximizing the received power (Example 1)

Objective function	Optimal value	Voltage at port 1	Directivity	Radiation efficiency	CPU time
$\mathrm{Re}\{v_1\}+\mathrm{Im}\{v_1\}$	0.0491	$-0.0126-0.0365\mathrm{i}$	7.62dBi	84.72%	0.34sec
$\mathrm{Re}\{v_1\}-\mathrm{Im}\{v_1\}$	0.1423	$-0.0126+0.1297\mathrm{i}$	6.81dBi	93.4%	0.55sec
$-\mathrm{Re}\{v_1\}+\mathrm{Im}\{v_1\}$	0.1902	$0.1536-0.0365\mathrm{i}$	7.97dBi	85.41%	0.56sec
$-\mathrm{Re}\{v_1\}-\mathrm{Im}\{v_1\}$	0.2833	$0.1536+0.1297\mathrm{i}$	8.17dBi	89.15%	0.63sec

Table 4.2: Maximizing the received power after imposing a constraint on the input impedance of the antenna (Example 1)

Objective function	Optimal value	Voltage at port 1	Directivity	Radiation efficiency	CPU time
$\mathrm{Re}\{v_1\}+\mathrm{Im}\{v_1\}$	0.0432	$-0.0192-0.0240\mathrm{i}$	7.34dBi	86.472%	0.78sec
$\mathrm{Re}\{v_1\}-\mathrm{Im}\{v_1\}$	0.0579	$-0.0192+0.0387\mathrm{i}$	6.30dBi	90.15%	0.87sec
$-\mathrm{Re}\{v_1\}+\mathrm{Im}\{v_1\}$	0.0674	$0.0434-0.0240\mathrm{i}$	6.56dBi	85.58%	0.86sec
$-\mathrm{Re}\{v_1\}-\mathrm{Im}\{v_1\}$	0.0821	$0.0434+0.0387\mathrm{i}$	7.75dBi	89.3%	0.82sec

$Y = D_1 + D_2\mathrm{i}$). An appropriate solution was found as

$$\begin{aligned} D_1 = \mathrm{diag}[&0, 0.0026, 0.0026, 0.0070, 0.0070, 0.0026, 0.0026,\\ &0.0138, 0.0134, 0.3252, 0.4268, 0.0136, 0.0123],\\ D_2 = \mathrm{diag}[&0, -0.0106, -0.0105, -0.0064, -0.0064, -0.0106,\\ &-0.0105, 0.0215, 0.0217, -0.0050, -0.0036, 0.0227, 0.0205], \end{aligned}$$

which corresponds to the antenna directivity of 3.55dBi.

Example 2: A general consensus in the field of antenna design is that a satisfactory beamforming with making nulls at an arbitrary number of directions is possible only when a sufficient number of (antenna) active elements are exploited in such a way that the size of the antenna array becomes several multiples of the wavelength. Many papers in the past decade, e.g., [79, 97], have concentrated on designing passive array antennas that are capable of making a null only at one direction. For instance, the paper [97] presents such a design based on a genetic algorithm whose running time is reported more than 4 weeks. Using the techniques developed here, the goal of this example is to disprove the foregoing belief. To be more precise, for the first time in the literature, we wish to design an on-chip antenna system with only one active element (antenna element) of the size equal to one wavelength such that the radiation pattern makes nulls at many undesired directions. This

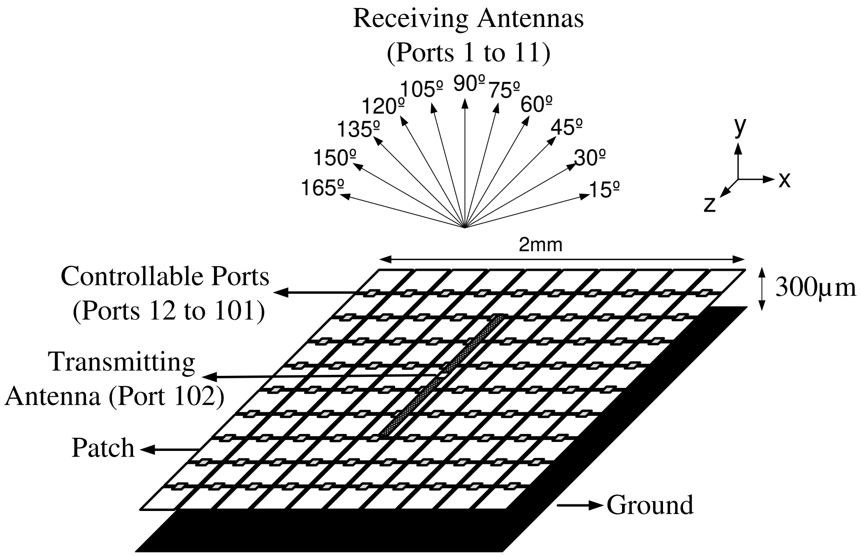


Figure 4.9: The antenna system studied in Example 2

antenna design is accomplished in a few seconds.

Consider the 2mm × 2mm antenna system depicted in Figure 4.9, consisting of a patch array with 90 controllable ports (shown by small squares) which is used for data transmission in the directions $15°, 30°, ..., 150°, 165°$. To study the programming capability of this antenna, a receiving antenna is placed at each of these directions in the far field (at the distance of 20 multiples of the wavelength from this transmitting antenna) with the length of 140 μm and the fixed terminal impedance 50 Ω. The equivalent circuit model of this antenna configuration is extracted using the electromagnetic software IE3D [41], which consists of 102 ports as follows:

- *Receiving or sensing ports (ports 1 to 11):* These ports are located in the far-field to capture the radiated power at the angles $15°, 30°, ..., 150°, 165°$.

- *Variable ports (ports 12 to 101):* Every two adjacent patches in the X-direction are connected with a port resulting in a total number of 90 ports, which are numbered from 12 to 101.

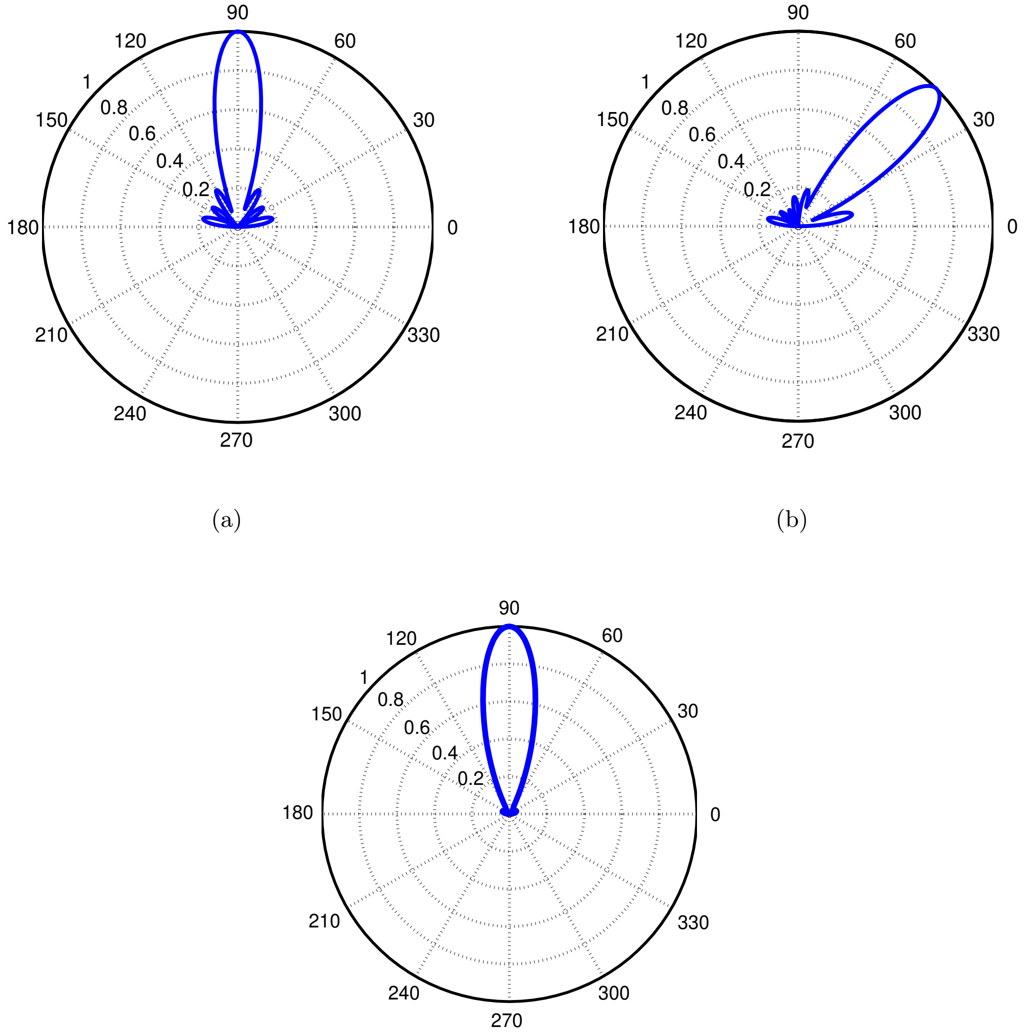
- *Transmitting port (port 102):* The transmitting port is located at the center of the transmitting antenna and is driven by a 300 GHz sinusoidal signal with a fixed amplitude of 1V.

Three objectives will be pursued in this example as follows. Assume that the first objective is to transmit data at the direction 90° with the maximum power using a passive control of the antenna. This problem reduces to finding a passive controller for the antenna configuration that generates the maximum power for the output port 6. Using Theorem 2, this leads to the voltage $v_6 = 0.00303 - 0.002274\mathrm{i}$. The corresponding radiation pattern of the antenna system is plotted in Figure 4.10(a). This figure shows that the antenna has an excellent beamforming capability; indeed, while the goal was to maximize the power in one direction, the radiating power was greatly minimized in most of the remaining directions.

As the second objective, the intent is to steer the beam towards the direction 45°. Similar to the previous case, the point $v_3 = -0.00234 - 0.0030\mathrm{i}$ is obtained with the radiation pattern drawn in Figure 4.10(b). The last objective is more interesting. We wish to transmit data to the direction 90° with the maximum power subject to the constraint that a zero signal is sent to all of the directions $15°, 30°, 45°, 60°, 120°, 135°, 150°, 165°$. Theorem 2 can be exploited to show that this highly constrained pattern shaping is possible. The optimal value $v_6 = -0.000866 + 0.000589\mathrm{i}$ is attained and the corresponding radiation pattern is depicted in Figure 4.10(c). An implication of this pattern is that the technique developed in this chapter has made it possible to design a wavelength-size antenna system with only one active element so that its proper control makes a null at many undesired directions while maximizing the power at a desired direction. The reader can contrast the patterns derived here with similar ones in the literature (such as the ones reported in [79, 97]), which radiate a high power in almost all directions and make a null in at most one direction. Note that despite the fact that the controllers designed in this example are not decoupled, one can verify that many elements of the obtained controllers are negligible, which facilitate their implementations.

4.5 Summary

This chapter studies a class of linear systems that appear in circuits, electromagnetics, optics, etc. Given such a linear system, the objective is to design a controller for the circuit (system) such that some prescribed linear constraints on the input admittance and output voltages of the circuit are satisfied. It is shown that designing a switching controller for this circuit amounts to a rank-minimization problem, and is indeed an NP-complete problem.



(a)

(b)

(c)

Figure 4.10: (a): The radiation pattern obtained by maximizing the received power at the direction 90°; (b): the radiation pattern obtained by maximizing the received power at the direction 45°; (c): the radiation pattern obtained by maximizing the received power at the direction 90° subject to the constraints $v_1 = v_2 = v_3 = v_4 = v_8 = v_9 = v_{10} = v_{11} = 0$ (Example 2)

Later on, the design of a passive controller is studied using the convex optimization theory. Since the implementation of a passive controller may be unacceptably more complicated than a switching controller, the design of a simpler type of controller, named *decoupled passive controller*, is also investigated. It is shown that this problem amounts to a rank-minimization one, which can be solved satisfactorily using a celebrated heuristic method. The results of the current work are developed based on available techniques in the control theory. As an important application, this work is exploited to design novel antenna systems with an outstanding performance.

4.6 Appendix

Consider an algorithm for a given decision problem that aims to find out whether the answer to this problem is “yes” or “no”. The notion of time complexity was introduced in the literature to evaluate the efficiency of such an algorithm. Informally speaking, the time complexity measures the number of machine instructions executed during the running time of the algorithm as a function of the size of the input. An efficient algorithm must run in polynomial time; for instance, if an algorithm needs an exponential number of iterations, then as the size of the problem increases, the running time of the algorithm grows astronomically. The class of NP-complete problems categorizes those problems that are believed to be extremely difficult to solve. Indeed, there is no known polynomial-time algorithm to solve an NP-complete problem, and moreover if an algorithm is discovered to solve an NP-complete problem in polynomial time, then the algorithm can be adapted to solve all NP problems in polynomial time [37]. It is desired to prove that the circuit switching problem posed in this chapter (i.e., Problem 1) is an NP-complete problem. This is accomplished in the sequel.

Theorem 4 *Problem 1 is NP-complete.*

Proof: Assume that n can be written as $3m + 2$, for some natural number m, and that $z = m+2$ (the technique being developed in the following can be adopted for other values of n). Recall that the present work considers output ports $\{1, 2, ..., z\}$ as the ports of interest used for specifying the design objectives and ports $\{z + 1, ..., n\}$ as the controllable ports connected to a control unit. To simplify the argument of the proof, assume with no loss of generality that ports $\{3k-2, 3k | k = 1, 2, ..., m\}$ are the controllable ports and the rest are the ports whose voltages are used for defining design specifications (a re-numbering of the output ports converts the problem to the conventional one considered here). Let the matrix Y_s have the particular form

$$Y_s = \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \end{bmatrix} & \begin{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ \begin{bmatrix} \alpha_2 & \alpha_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \alpha_m & \alpha_m & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} \begin{bmatrix} \alpha_1 & 0 & 0 \\ \alpha_1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} \alpha_2 & 0 & 0 \\ \alpha_2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \cdots & \begin{bmatrix} \alpha_m & 0 & 0 \\ \alpha_m & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{bmatrix} \mathrm{i}$$

where $\alpha_1, \alpha_2, ..., \alpha_m$ are some arbitrary integers. It is worth mentioning that this type of Y_s corresponds to a lossless network. Impose the constraints

$$\begin{gathered} v_{3k-1} = v_{\text{in}} \quad \forall k \in \{1, 2, ..., m\}, \\ v_{3m+1} = v_{\text{in}}, \quad v_{3m+2} = -v_{\text{in}} \end{gathered} \tag{4.36}$$

on the output voltages. The goal is to show that Problem 1 is NP-complete even for the special networks of the form introduced earlier under the above constraints. Given a natural number $k \in \{1, 2, ..., m\}$, the conditions in (4.36) lead to the equations

$$\begin{aligned} i_{3k-2} &= (v_{3k-2} - v_{\text{in}})\mathrm{i}, \\ i_{3k-1} &= (-v_{3k-2} + v_{3k})\mathrm{i}, \\ i_{3k} &= (-v_{3k} + v_{\text{in}})\mathrm{i}. \end{aligned}$$

Since port $3k-1$ is not a controllable port, it follows from the design specifications in (4.1) that its current must be zero. In other words, $i_{3k-1} = 0$ or equivalently $v_{3k-2} = v_{3k}$. On the other hand, the above equations yield that the switching condition $v_{3k-2}i_{3k-2} = 0$ is tantamount to the relation $v_{3k-2} \in \{0, v_{\text{in}}\}$. Thus, it can be concluded that

$$v_{3k-2} = v_{3k} \in \{0, v_{\text{in}}\} \quad \forall k \in \{1, 2, ..., m\}. \tag{4.37}$$

Moreover, since ports $3m+1$ and $3m+2$ are not controllable ports, their current must be zero, which gives rise to

$$0 = i_{3m+1} = \sum_{j=1}^{m} \alpha_j v_{3j-2}.$$

Note that the equality $i_{3m+2} = 0$ also leads to the above constraint. On using (4.37) and by letting v_{in} be equal to 1, the above equation can be interpreted as follows: given the integers $\alpha_1, \alpha_2, ..., \alpha_m$, is it possible to find a subset of these numbers with the zero sum? This problem is referred to as *subset sum problem* and is known to be NP-complete [37]. This completes the proof. ■

Chapter 5

Passively Controllable Smart Antennas

This work deals with devising a secure, power-efficient, beam-steerable and on-chip transmission system for wireless sensor networks. A *passively controllable smart (PCS) antenna system* is introduced, which can be programmed to generate different radiation patterns at the far field by adjusting its variable passive controller at every signal transmission. In particular, the PCS antenna is able to transmit data to a desired direction in such a way that no signal is sent in many undesired directions. To study the capabilities of a PCS antenna system, a number of sensor nodes are placed around a PCS antenna, where the nodes are all equipped with (short) sensing dipole antennas for signal reception. It is shown that a pre-specified set of voltages can be induced on the receiving antennas if and only if a linear matrix inequality (LMI) problem is feasible. Later on, this LMI problem is further simplified and its feasibility region is proved to be ellipsoidal. This feasibility region completely characterizes not only the values of the voltages received by different nodes but also the correlation among these voltages. Based on this ellipsoid, a transmitting node can adjust its variable passive controller to transmit data to any intended node in such a way that many of the unintended nodes all receive a zero signal (no data). Unlike the existing smart antennas whose programming leads to an NP-hard problem or are made of many active elements, the PCS antenna proposed in the present work has a low-complex programming capability and consists of only one active element. These two properties make it possible to implement a PCS antenna on a single small-sized silicon chip together with an on-chip low-power processor to satisfy the strict size and power limitations associated with the wireless sensor nodes.

5.1 Introduction

A wireless sensor network (WSN) is composed of several geographically distributed tiny sensors, where each sensor device is equipped with sensing, data processing, and communication elements. WSNs have been extensively studied for many years due to their broad range of civil and military applications such as security surveillance, object detection, target tracking, home automation, environmental monitoring, and health monitoring [18, 102, 1]. Since wireless senor nodes are small in size and have very limited computation and communication capabilities, most of the existing theories developed for a general wireless network cannot be directly applied to WSNs. Among many problems that have been investigated specifically for WSNs, one can name routing [58], localization [59, 85], security [119], joint routing and power control [86], and resource optimization [99].

Conventional antennas for wireless transmission, e.g., omni-directional antennas, radiate in almost all directions. To avoid co-channel interference and unnecessary power consumption in undesired directions, it is preferable not to deploy conventional wireless transmission systems. A great amount of effort has been made in the past several decades to design smart transmitting/receiving antenna systems, which are able to increase the capacity of wireless networks [66]. Two main types of smart antennas are switched beam and adaptive array. A switched beam smart antenna has several pre-designed fixed beam patterns, whereas an adaptive array smart antenna adaptively steers the beam to any direction of interest while simultaneously nulling interfering signals [30, 120]. Note that an array system comprises multiple active (antenna) elements for varying the relative phases and amplitudes of the respective signals in order to generate a desired radiation pattern. Other types of smart antenna systems employ only one active element surrounded by a number of passive parasitic elements, with the disadvantage that they are either non-programmable or their online programming leads to an NP-hard problem [79, 2, 3].

Different types of smart antennas, such as directional, beamforming/array, and multiple-input-multiple-output antennas, have been studied and applied to ad-hoc networks [104, 42]. The deployment of smart antennas is more crucial in WSNs than general ad-hoc networks, due to very limited resources available in WSNs. A number of papers have explored the effects of smart antennas in WSNs, e.g., the necessity to change the existing medium access control (MAC) protocols [9, 95] or the maximum flow problem under switched-beam directional antennas [39]. Nonetheless, the aforementioned smart antenna systems cannot be exploited in WSNs by virtue of the fact that these systems either need heavy computations for their programming or are very large in size due to using several active elements (antenna array) with a mutual element-to-element distance in the order of the signal wavelength. This chapter aims to build on the results developed in [2], [3] and the previous chapter to propose a new type of smart antenna system, referred to as *passively controllable smart (PCS) antenna*, which has a low-complex programming and utilizes only one active element. This PCS antenna system can be implemented on a cheap, small-sized, low-power silicon chip to comply with the strict size and power limitations in WSNs.

The PCS antenna system proposed in this chapter is composed of a main dipole (transmitting) antenna, a number of reflectors (or a patch array), and a variable (tunable) passive controller. Since changing the parameters of the passive controller modifies the radiation pattern generated at the far field, this act is regarded as *programming* of the PCS antenna. To study the programming capabilities of a PCS antenna, a number of receiving nodes are placed around the PCS antenna, which are all equipped with short dipole antennas for signal reception. It is shown that a pre-determined set of voltages can be sent to the receiving nodes if and only if the vector of voltages satisfies a linear matrix inequality (LMI) problem. Using this result, it is proved that a pre-specified radiation pattern can be generated for the receiving antennas if and only if the associated vector of voltages belongs to an ellipsoidal region. This region characterizes both the individual signals that can be sent to different antennas and the correlation among these signals. Based on the obtained properties, it is shown how the PCS antenna can be programmed to transmit data to an intended node in such a way that many of the unintended nodes receive a zero signal (no signal) simultaneously.

5.2 Problem Statement and Preliminaries

Given a natural number z, consider a wireless network with $z+1$ nodes, labeled as $0, 1, 2, ..., z$. Assume that these nodes are geographically distributed so that none of the two nodes in the set $\{1, 2, ..., z\}$ are co-linear with node 0. This assumption is made to ensure an angle diversity among nodes $1, 2, ..., z$ with respect to node 0. It is desired to devise a smart antenna system for node 0 that can be programmed to transmit data to any node $j \in \{1, 2, ..., z\}$ in such a way that many of the remaining nodes $1, ..., j-1, j+1, ..., z$ receive a zero signal. The smart antenna being contrived in this work relies on the notion of *near-field direct antenna modulation*, which has been recently introduced in the papers [2] and [3]. Let some preliminaries be provided on this notion before developing the main results of the present work.

5.2.1 Preliminaries

In a conventional wireless transmission scheme, the information is generated before the antenna and the role of the antenna is to efficiently transmit the signal. Different methods are developed to add information to a carrier signal (e.g., a sinusoidal waveform). The act of adding information to a carrier signal is referred to as *modulation*, and can be achieved by altering some properties of the carrier signal such as its frequency, amplitude, or phase. In a broad range of wireless communication systems, the information is generated in low frequencies (base-band frequency region), and then up-converted to a carrier frequency (RF) via a mixer that acts as a multiplier. The base-band data forms a series of complex numbers that can be separated into real and imaginary parts. The first set is called the in-phase signal (I) that is the real part of the complex signal, and the second set is called the quadrature-phase signal (Q) that is the imaginary part of the complex signal. A simple constellation diagram can be used to represent this complex signal, where each point of the diagram corresponds to some information symbol.

After modulating an incoming signal, a conventional antenna propagates the modulated signal in many directions. For example, consider the conventional transmitting antenna depicted in Figure 5.1, which is used to transmit four symbols A, B, C, D to a receiving antenna #1 in an environment where there exists an unwanted receiving antenna #2. As shown in the figure, assume that antenna #1 receives four points $1, 2, 3, 4$ corresponding to these symbols, whereas antenna #2 receives four other points $1', 2', 3', 4'$. It is easy to argue that the constellation diagram corresponding to antenna #2 can be obtained from the constellation diagram for antenna #1 using an appropriate rotation and re-scaling. Hence, the unintended antenna #2 can discover what symbols antenna #1 receive. This undesirable property is due to the fact that the signal being transmitted by a conventional antenna can ideally be received in different directions after some time delay and power attenuation. Therefore, a conventional antenna is associated with a number of problems as

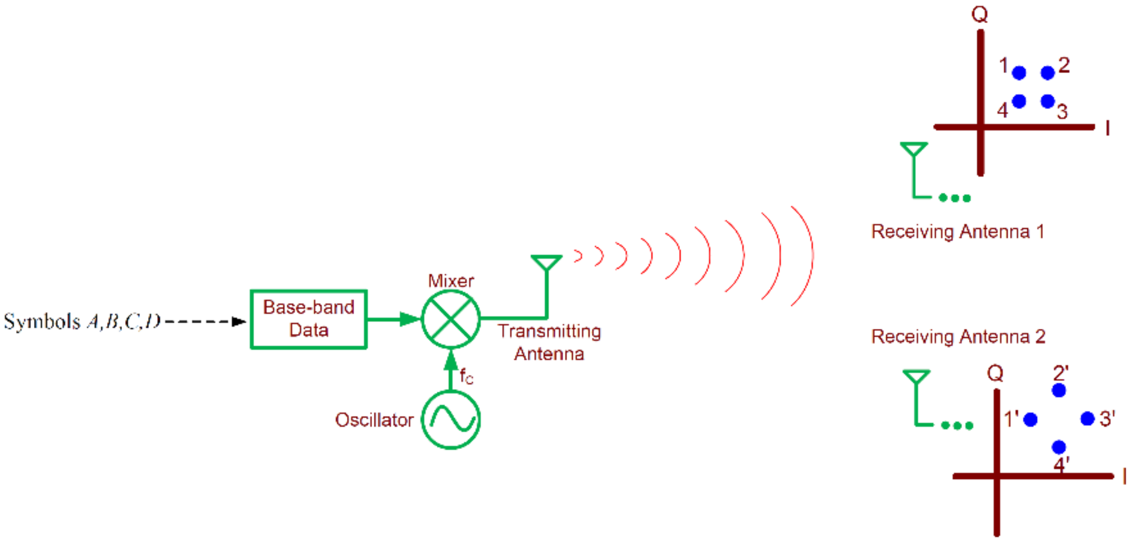


Figure 5.1: This figure illustrates the weakness of a conventional antenna in secure wireless transmission.

follows:

- In a hostile environment, the antenna does not guarantee a secure data transmission.
- The antenna causes co-channel interference in a wireless network and hence reduces the network throughput.
- The antenna wastes the transmitting power in undesired directions.

The recent papers [2] and [3] have introduced the new notion of *near-field direct antenna modulation* to design a novel type of antenna system for secure wireless transmission, which is on-chip, small-sized, and low-power consuming (due to using only one active element). The antenna system proposed therein has a main dipole transmitting antenna driven by a voltage source, a number of reflectors and several switches mounted on the reflectors. Since each switch can be turned on or off, there exist different switching strategies. Each switching combination creates a possibly unique near-field boundary condition around the antenna, which results in different radiation patterns at the far field. Therefore, each switching combination could possibly generate a new point in the constellation diagram. Figure 5.2(a) exemplifies the antenna system suggested in [2], which consists of 4 reflectors and 12 switches (shown by arrows). It can be observed that there exist 2^{12} switching combinations, which create a constellation diagram with numerous points. The antenna system introduced in [2] has not only a modulation capability, but also a direction-dependent transmission ability. Indeed, since the reflectors affect the electromagnetic field around the antenna in a non-uniform way, the constellation diagrams seen in different directions are not necessarily correlated. Figure 5.2(b) illustrates this property via an antenna system with 4 switches, which makes a 16-QAM constellation diagram in the vertical direction and a totally scrambled one in an undesired direction. Figure 5.3 shows a prototype of the antenna system suggested in [3] with 90 switches and 10 reflectors that is implemented on a silicon-based chip of the size 1.5mm × 1.3mm.

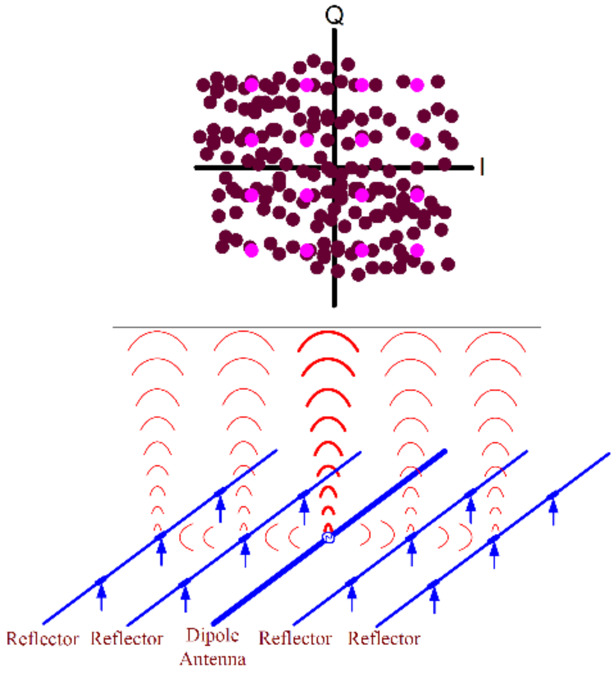
The switch-based antenna system proposed in [2] and [3] has some useful capabilities, such as secure and direction-dependent wireless transmission. However, the identification of the switching combinations that generate proper radiation patterns at the far field amounts to an NP-complete problem (because the swtiching problem can be reduced to the subset sum problem). As an alternative to the switch-based strategy, the work [2] also suggests using varactors (variable impedances) instead of switches. This chapter aims to build on this new type of antenna system to contrive a *programmable smart antenna system* for wireless sensor networks.

Before proceeding with the main results, let some necessary notations and definitions be made in the sequel.

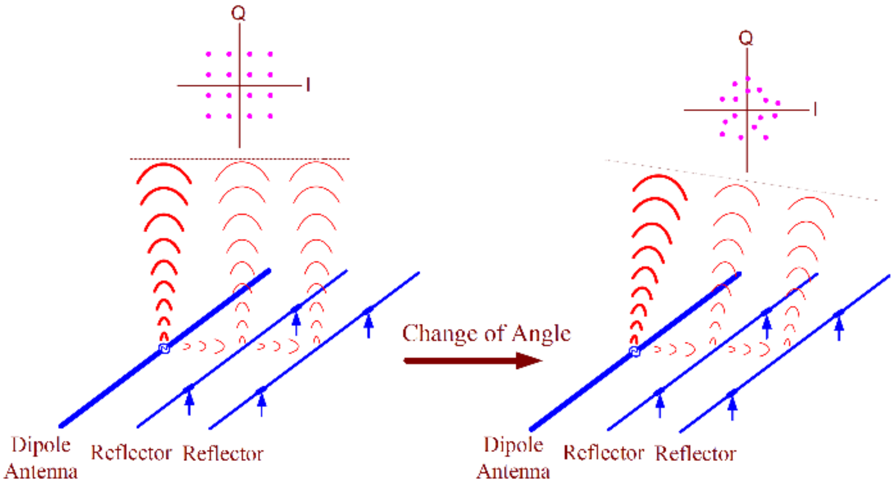
Notation 1 *Introduce the following notations:*

- i *: the imaginary unit;*
- $\mathbf{N}$*,* $\mathbf{R}$ *and* $\mathbf{C}$*: the sets of natural, real and complex numbers, respectively;*
- $\mathbf{S}^{k\times k}$*: the set of all symmetric matrices in* $\mathbf{R}^{k\times k}$ *(where* $k \in \mathbf{N}$*);*
- $Re\{\cdot\}$ *and* $Im\{\cdot\}$*: the operators returning the real and imaginary parts of a complex matrix;*
- $*$ *: the matrix operator taking the conjugate transpose of a complex-valued matrix;*
- $\succ$ *: the matrix inequality sign in the positive definite sense.*

Notation 2 *For convenience and with a slight abuse of notation, the terms circle, ellipse and ellipsoid in this chapter will refer to the interiors of the conventional circle, ellipse and ellipsoid, respectively. For instance, the unit circle here refers to the set* $\{(x, y) \in \mathbf{R}^2 \mid x^2 + y^2 < 1\}$*, as opposed to* $\{(x, y) \in \mathbf{R}^2 \mid x^2 + y^2 = 1\}$



(a)



(b)

Figure 5.2: (a): This figure illustrates the modulation capability of the switch-based antenna system proposed in [2]. (b): This figure illustrates the direction-dependent transmission capability of the switch-based antenna system proposed in [2].

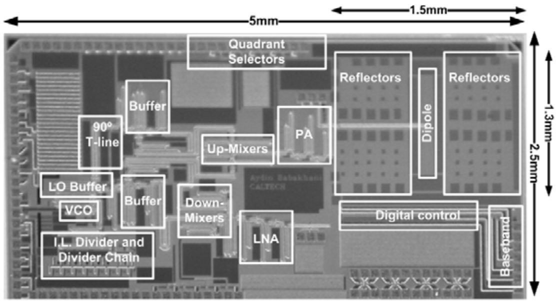


Figure 5.3: A prototype of the antenna system developed in [3]

Definition 1 *For every real-valued column vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *of the same dimension, define the 2-norm* $\|\mathbf{x}_1 + \mathbf{x}_2 i\|$ *as* $\sqrt{\mathbf{x}_1^* \mathbf{x}_1 + \mathbf{x}_2^* \mathbf{x}_2}$.

Definition 2 *Given a scalar* $k \in \mathbf{N}$ *and a set* $\mathcal{H} \subseteq \mathbf{C}^{1\times k}$*, define the real-valued representation of the set* $\mathcal{H}$ *as the set of all real vectors in the form of* $\begin{bmatrix} Re\{\boldsymbol{\alpha}\} & Im\{\boldsymbol{\alpha}\} \end{bmatrix}$ *such that* $\boldsymbol{\alpha}$ *is an element of* $\mathcal{H}$*. The operator* $\mathcal{R}(\cdot)$ *will be used henceforth to represent the real-valued representation of a set; for instance, the real-valued representation of* $\mathcal{H}$ *is denoted by* $\mathcal{R}(\mathcal{H})$.

5.3 Passively Controllable Smart Antenna

Assume for now that nodes $0, 1, 2, ..., z$ are all fixed. The case when these nodes are mobile will be later discussed in Remark 4. Define a *passively controllable smart (PCS) antenna system* as a system with the following components:

- *A dipole transmitting antenna:* This dipole antenna is the only active element of the PCS antenna system, which is driven by a sinusoidal voltage source.

- *A number of reflectors:* These reflectors surround the dipole antenna to shape the electromagnetic field in the space.

- *A number of controllable ports:* These controllable ports are mounted on the reflectors which should be controlled for every signal transmission to form a desired radiation pattern at the far field.

- *An adjustable passive network (controller):* This passive network consists of resistors, capacitors and inductors, and is connected to the controllable ports of the reflectors to control the antenna system for every signal transmission.

Note that a PCS antenna system resembles the antenna system proposed in [2] and [3], with the main difference that the switches/varactors are replaced by controllable ports that must be controlled by passive elements for every signal transmission. For simplicity, the term *PCS antenna system* will be abbreviated as *PCS antenna.* Since a PCS antenna has only one active element and its controller solely includes (variable) passive elements, it can be implemented as a low-power integrated on-chip programmable antenna. Hence, it is pragmatic to deploy PCS antennas in a wireless sensor network.

Recall the main problem of interest posed in this chapter regarding the cluster of sensor nodes $0, 1, 2, ..., z$, i.e., devising a means of wireless transmission from node 0 to a node $j \in \{1, 2, ..., z\}$, in such a way that each of the remaining nodes $1, 2, ..., j-1, j+1, ..., z$ receives a zero signal if possible. To address this problem, let nodes $1, 2, ..., z$ employ short dipole receiving antennas and node 0 exploit a PCS transmitting antenna, where each signal transmission is performed by applying an appropriate passive controller to this antenna system. For every $k \in \{1, 2, ..., z\}$, let v_k denote the voltage that the PCS transmitting antenna of node 0 induces on the receiving antenna of node j. Note that every passive controller that is applied to the controllable ports of the PCS antenna of node 0 generates a specific voltage vector $(v_1, v_2, ..., v_z)$ at the far field. The goal is to study a series of problems in the sequel, which are outlined below:

i) Given $j \in \{1, 2, ..., z\}$, what is the set of all possible voltages v_j that can be generated by the PCS antenna of node 0 (under every possible passive controller)?

ii) What is the set of all possible voltage vectors $(v_1, v_2, ..., v_z)$ that can be generated by the PCS antenna of node 0?

iii) Given $j \in \{1, 2, ..., z\}$, is it possible to passively control the PCS antenna of node 0 to transmit data to node j in such a way that many of the unintended nodes $1, 2, ..., j-1, j+1, ..., z$ receive a zero signal (voltage)?

Notice that problem (i) aims to find the constellation diagram seen at node j, problem (ii) investigates the correlation of the voltages received at different nodes, and problem (iii) studies the possibility of transmitting data mainly to one intended node. In the rest of this chapter, the act of designing and applying a passive controller to a PCS antenna will be referred to as *programming* the PCS antenna.

Let f_0 and v_{in} denote the frequency and magnitude of the sinusoidal voltage driving the dipole transmitting antenna of the PCS antenna system, respectively. Assume that nodes $1, 2, ..., z$ all lie in the far field of node 0, meaning that the distance of each of nodes $1, 2, .., z$ from node 0 is noticeably greater than the wavelength $\frac{3\times10^8}{f_0}$. For example, this assumption at the operating frequency $f_0 = 2$ GHz can be translated as all nodes $1, 2, ..., z$ are distant from node 0 by at least 0.150 meter. One can observe that this assumption normally holds in practice. To tackle problems (i), (ii), and (iii) stated earlier, one can extract the equivalent circuit model of the entire antenna configuration that consists of the transmitting antenna of node 0 and the receiving antennas of nodes $1, 2, ..., z$. This circuit model, referred to as Circuit 1, is given in Figure 5.4(a), where

- The block "Linear Passive Network" corresponds to the Y-parameter matrix of the antenna configuration (calculated from scattering parameters), which can be found using an electromagnetic simulation.

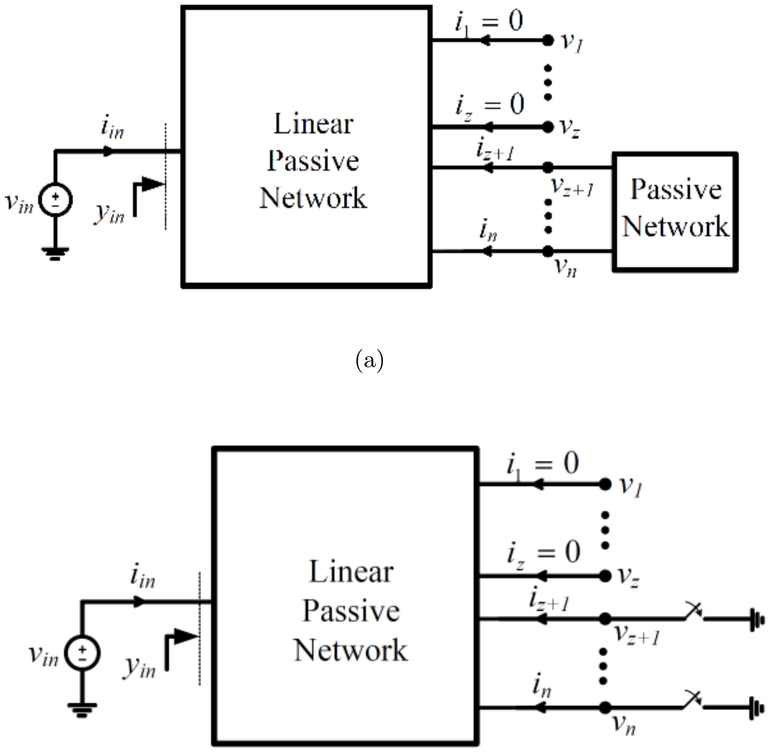
- $v_{z+1}, v_{z+2}, ..., v_n$ denote the voltages on the controllable ports of the PCS antenna of node 0 (it is assumed that there are $n - z$ controllable ports).

- The block "Passive Network" represents the adjustable passive controller applied to the controllable ports of the PCS antenna of node 0.

Note that the equivalent circuit model of the antenna system proposed in [3] can be derived from the circuit given in Figure 5.4(a) by replacing its passive network with a switching network, as depicted in Figure 5.4(b). With no loss of generality, assume that $v_{\text{in}} = 1$ (this can be achieved using an appropriate re-scaling). Denote the Y-parameter matrix of the antenna configuration at the given frequency f_0 with Y_s. Moreover, let y_{in} represent the input admittance of the PCS antenna of node 0. From now on, suppose that the adjustable passive controller of the PCS antenna must be linear and strictly passive because the goal is to implement this controller by an interconnection of a number of variable resistors and possibly some variable capacitors and inductors.

Decompose the complex-valued matrix Y_s in a block form as

$$Y_s = \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix},$$



(a)

(b)

Figure 5.4: (a): Equivalent circuit model of the passively controllable smart antenna system proposed in the present work (named Circuit 1). (b): Equivalent circuit model of the switch-based antenna system proposed in [2] and [3] (named Circuit 2)

where $W_{11} \in \mathbf{C}^{z\times z}$, $W_{22} \in \mathbf{C}^{(n-z)\times(n-z)}$, and $W_{33} \in \mathbf{C}$. Let y_{in}^{d} denote the admittance of the source delivering power to the dipole antenna of the PCS antenna system of node 0 (the standard value of y_{in}^{d} is $\frac{1}{50}$ Ω^{-1}). Given a complex vector $\boldsymbol{\alpha} \in \mathbf{C}^{1\times z}$, it is desired to investigate whether the PCS antenna of node 0 can be programmed so that it generates the voltage vector $(v_1, v_2, ..., v_z) = \boldsymbol{\alpha}$ and concurrently the relation $y_{\text{in}} = y_{\text{in}}^{d}$ is satisfied. It is worth mentioning that the constraint $y_{\text{in}} = y_{\text{in}}^{d}$ is said to be the *impedance matching* constraint, whose role is to minimize the power reflected by the antenna in order to save power consumption.

Theorem 1 *Given $\boldsymbol{\alpha} \in \mathbf{C}^{1\times z}$, the PCS antenna of node 0 can be programmed to make the voltages $v_1, v_2, ..., v_z$ and the antenna input admittance y_{in} satisfy the relations*

$$(v_1, v_2, ..., v_z) = \boldsymbol{\alpha}, \quad y_{in} = y_{in}^{d} \tag{5.1}$$

if and only if there exist symmetric matrices $M, N \in \mathbf{R}^{(n-z)\times(n-z)}$ such that

$$\begin{bmatrix} \left(Re\{W_{22} - W_{21}W_{11}^{-1}W_{12}\}\right)^{-1} - M & N \\ N & M \end{bmatrix} \succ 0, \tag{5.2}$$

and

$$-(W_{31}W_{11}^{-1}W_{12} - W_{32})(M + Ni)W_{21}W_{11}^{-1} - W_{31}W_{11}^{-1} = \boldsymbol{\alpha}, \tag{5.3a}$$

$$-(W_{31}W_{11}^{-1}W_{12} - W_{32})(M + Ni)(W_{21}W_{11}^{-1}W_{13} - W_{23}) + W_{33} - W_{31}W_{11}^{-1}W_{13} = y_{in}^{d}. \tag{5.3b}$$

Moreover, if there exist such matrices M, N satisfying the above constraints, then one candidate for the admittance of the passive controller at the frequency f_0, denoted by Y_0, is

$$Y_0 = (M + Ni)^{-1} - W_{22} + W_{21}W_{11}^{-1}W_{12}. \tag{5.4}$$

Proof: The proof is a direct consequence of Theorem 2 in Chapter 4. ■

Theorem 1 states that the PCS antenna of node 0 can generate a specific radiation pattern at the far field subject to an impedance matching constraint if and only if a linear matrix inequality (LMI) problem is feasible (see [13] for the definition of LMI). Since the feasibility region of an LMI problem is convex, the set of all possible voltages $(v_1, v_2, ..., v_z)$ has a convex real-valued representation (see Definition 2). The objective is to further simplify the LMI conditions derived in Theorem 1.

Remark 1 *Assume that the aim is to only check whether the PCS antenna of node* 0 *can generate a voltage vector* $(v_1, v_2, ..., v_z)$ *equal to a pre-specified vector* $\boldsymbol{\alpha}$ *and the impedance matching constraint* $y_{in} = y_{in}^d$ *need not be satisfied. Theorem 1 can be easily adapted to tackle this problem by removing the constraint* $y_{in} = y_{in}^d$ *from the equation (5.1) and eliminating the equation (5.3b).*

Definition 3 *Define* $\mathcal{D}$ *to be the set of all complex-valued* z*-tuples* $(v_1, v_2, ..., v_z)$ *that can be generated by the programmable PCS antenna of node* 0*. Likewise, define* $\tilde{\mathcal{D}}$ *to be the set of all complex-valued* $(z+1)$*-tuples* $(v_1, v_2, ..., v_z, y_{in})$ *that can be produced by the antenna of node* 0*.*

The complex-valued set $\mathcal{D}$ captures the correlation among the individual signals that nodes $1, 2, ..., z$ receive. Besides, the set $\tilde{\mathcal{D}}$ relates the individual signals $v_1, ..., v_z$ not only to each other but also to the input admittance of the transmitting antenna. Hence, the complex-valued sets $\mathcal{D}$ and $\tilde{\mathcal{D}}$ contain important information about the spatial distribution of the signal transmitted by the PCS antenna of node 0. To characterize these sets, two lemmas are required, in addition to Theorem 1, which will be provided in the sequel.

Lemma 1 *Given a scalar* $m \in \mathbf{N}$ *and vectors* $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{R}^{1\times m}$ *with the property* $\|[\ \mathbf{x}_1 \quad \mathbf{x}_2\ ]\| = 1$*, consider the set of all vectors* $\boldsymbol{\alpha} \in \mathbf{R}^{1\times 2m}$ *for which there exist symmetric matrices* $M, N \in \mathbf{R}^{m\times m}$ *such that*

$$\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \begin{bmatrix} M & N \\ N & -M \end{bmatrix} \tag{5.5}$$

and

$$\begin{bmatrix} M & N \\ N & -M \end{bmatrix} \prec I. \tag{5.6}$$

This set is identical to the open unit ball $\{\boldsymbol{\gamma} \in \mathbf{R}^{1\times 2m} \mid \|\boldsymbol{\gamma}\| < 1\}$*.*

Proof: The proof is provided in the appendix. ■

Lemma 2 *Given scalars* $m, k \in \mathbf{N}$*, vectors* $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{R}^{1\times m}$*, and matrices* $G_1, G_2 \in \mathbf{R}^{m\times k}$*, consider the set of all complex vectors* $\boldsymbol{\alpha} \in \mathbf{C}^{1\times k}$ *that can be written as*

$$\boldsymbol{\alpha} = (\mathbf{x}_1 + \mathbf{x}_2 i)(M + Ni)(G_1 + G_2 i) \tag{5.7}$$

for some symmetric matrices $M, N \in \mathbf{R}^{m\times m}$ *with the property*

$$\begin{bmatrix} M & N \\ N & -M \end{bmatrix} \prec I. \tag{5.8}$$

The real-valued representation of this complex set is identical to the ellipsoid

$$\left\{ \mathbf{h} \in \mathbf{R}^{1\times 2k} \middle| \mathbf{h} \left(\begin{bmatrix} G_1^* & -G_2^* \\ G_2^* & G_1^* \end{bmatrix} \begin{bmatrix} G_1 & G_2 \\ -G_2 & G_1 \end{bmatrix} \right)^{-1} \mathbf{h}^* < \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 \right\}, \tag{5.9}$$

provided the matrix $G_1 + G_2 i$ *has full column rank over the field of complex numbers.*

Proof: Observe that the equation (5.7) is tantamount to

$$\begin{bmatrix} \mathrm{Re}\{\boldsymbol{\alpha}\} & \mathrm{Im}\{\boldsymbol{\alpha}\} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & -\mathbf{x}_2 \end{bmatrix} \begin{bmatrix} M & N \\ N & -M \end{bmatrix} \begin{bmatrix} G_1 & G_2 \\ -G_2 & G_1 \end{bmatrix}. \tag{5.10}$$

On the other hand, it follows from Lemma 1 that the set

$$\left\{ \begin{bmatrix} \mathbf{x}_1 & -\mathbf{x}_2 \end{bmatrix} \begin{bmatrix} M & N \\ N & -M \end{bmatrix} \middle| M, N \in \mathbf{S}^{m\times m}, \begin{bmatrix} M & N \\ N & -M \end{bmatrix} \prec I \right\}$$

is an open ball centered at the origin with radius $\sqrt{\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2}$. It can be inferred from this result that the set

$$\left\{ \begin{bmatrix} \mathbf{x}_1 & -\mathbf{x}_2 \end{bmatrix} \begin{bmatrix} M & N \\ N & -M \end{bmatrix} \begin{bmatrix} G_1 & G_2 \\ -G_2 & G_1 \end{bmatrix} \middle| M, N \in \mathbf{S}^{m\times m}, \begin{bmatrix} M & N \\ N & -M \end{bmatrix} \prec I \right\}$$

is equal to the ellipsoid given in (5.9), provided the matrix

$$\begin{bmatrix} G_1^* & -G_2^* \\ G_2^* & G_1^* \end{bmatrix} \begin{bmatrix} G_1 & G_2 \\ -G_2 & G_1 \end{bmatrix}$$

is nonsingular or equivalently $G_1 + G_2\mathrm{i}$ has full column rank. The proof is an immediate consequence of this property and the fact that the equation (5.7) is the same as (5.10). ■

Define some matrices as follows:

$$\begin{aligned}
&K_1 := W_{31}W_{11}^{-1}W_{12} - W_{32}, \quad K_2 := W_{21}W_{11}^{-1}W_{13} - W_{23}, \quad K_3 := W_{31}W_{11}^{-1}W_{13} - W_{33},\\
&K_4 := W_{21}W_{11}^{-1}, \quad K_5 := W_{31}W_{11}^{-1}, \quad K_6 := \begin{bmatrix} K_4 & K_2 \end{bmatrix}, \quad K_7 := \begin{bmatrix} K_5 & K_3 \end{bmatrix},\\
&Q := \left(\mathrm{Re}\left\{W_{22} - W_{21}W_{11}^{-1}W_{12}\right\}\right)^{-1},\\
&\mathbf{o} := -\begin{bmatrix} \mathrm{Re}\left\{\frac{1}{2}K_1QK_4 + K_5\right\} & \mathrm{Im}\left\{\frac{1}{2}K_1QK_4 + K_5\right\} \end{bmatrix},\\
&\tilde{\mathbf{o}} := -\begin{bmatrix} \mathrm{Re}\left\{\frac{1}{2}K_1QK_6 + K_7\right\} & \mathrm{Im}\left\{\frac{1}{2}K_1QK_6 + K_7\right\} \end{bmatrix}.
\end{aligned}$$

Since the matrix Q introduced above is positive definite, one can define $Q^{\frac{1}{2}}$ as the unique symmetric positive definite matrix whose square is equal to Q. The next theorem presents one of the main results of this work, which exploits Lemma 2 and Theorem 1 to characterize the feasibility regions $\mathcal{D}$ and $\tilde{\mathcal{D}}$.

Theorem 2 *The following statements hold:*

i) If the matrix K_4 has full column rank over the field of complex numbers, then the real-valued representation of the complex set $\mathcal{D}$, i.e., $\mathcal{R}(\mathcal{D})$, is equal to the ellipsoid

$$\left\{\mathbf{h} \in \mathbf{R}^{1\times 2z} \middle| (\mathbf{h}-\mathbf{o}) \begin{bmatrix} Re\{K_4^*QK_4\} & Im\{K_4^*QK_4\} \\ -Im\{K_4^*QK_4\} & Re\{K_4^*QK_4\} \end{bmatrix}^{-1} (\mathbf{h}-\mathbf{o})^* < \frac{1}{4}\|K_1Q^{\frac{1}{2}}\|^2 \right\}. \tag{5.11}$$

ii) If the matrix K_6 has full column rank over the field of complex numbers, then the real-valued representation of the complex set $\tilde{\mathcal{D}}$, i.e., $\mathcal{R}(\tilde{\mathcal{D}})$, is equal to the ellipsoid

$$\left\{\mathbf{h} \in \mathbf{R}^{1\times 2(z+1)} \middle| (\mathbf{h}-\tilde{\mathbf{o}}) \begin{bmatrix} Re\{K_6^*QK_6\} & Im\{K_6^*QK_6\} \\ -Im\{K_6^*QK_6\} & Re\{K_6^*QK_6\} \end{bmatrix}^{-1} (\mathbf{h}-\tilde{\mathbf{o}})^* < \frac{1}{4}\|K_1Q^{\frac{1}{2}}\|^2 \right\}. \tag{5.12}$$

Proof of Part (i): It can be concluded from Theorem 1 and Remark 1 that a complex vector $\boldsymbol{\alpha}$ belongs to $\mathcal{D}$ if and only if there exist symmetric matrices $M, N \in \mathbf{R}^{(n-z)\times(n-z)}$

such that

$$\begin{bmatrix} Q-M & N \\ N & M \end{bmatrix} \succ 0 \tag{5.13}$$

and

$$\begin{aligned} \boldsymbol{\alpha} =& -(W_{31}W_{11}^{-1}W_{12}-W_{32})(M+N\mathrm{i})W_{21}W_{11}^{-1}-W_{31}W_{11}^{-1} \\ =& -K_1(M+N\mathrm{i})K_4-K_5. \end{aligned} \tag{5.14}$$

The constraint (5.13) can be re-arranged as

$$\begin{bmatrix} \tilde{M} & \tilde{N} \\ \tilde{N} & -\tilde{M} \end{bmatrix} \prec I, \tag{5.15}$$

where

$$\tilde{M} := 2Q^{-\frac{1}{2}}MQ^{-\frac{1}{2}}-I, \quad \tilde{N} := 2Q^{-\frac{1}{2}}NQ^{-\frac{1}{2}}. \tag{5.16}$$

Moreover, the constraint (5.14) can be expressed in terms of $\tilde{M}$ and $\tilde{N}$ as follows:

$$\begin{aligned} \boldsymbol{\alpha} &= -K_1(M+N\mathrm{i})K_4-K_5 \\ &= -\frac{1}{2}K_1\left(Q^{\frac{1}{2}}\tilde{M}Q^{\frac{1}{2}}+Q+Q^{\frac{1}{2}}\tilde{N}Q^{\frac{1}{2}}\mathrm{i}\right)K_4-K_5 \\ &= -\frac{1}{2}K_1Q^{\frac{1}{2}}(\tilde{M}+\tilde{N}\mathrm{i})Q^{\frac{1}{2}}K_4-\left(\frac{1}{2}K_1QK_4+K_5\right). \end{aligned} \tag{5.17}$$

By applying Lemma 2 to the constraints (5.15) and (5.17) and using the relation

$$\begin{bmatrix} \mathrm{Re}\{K_4^*Q^{\frac{1}{2}}\} & -\mathrm{Im}\{K_4^*Q^{\frac{1}{2}}\} \\ \mathrm{Im}\{K_4^*Q^{\frac{1}{2}}\} & \mathrm{Re}\{K_4^*Q^{\frac{1}{2}}\} \end{bmatrix} \times \begin{bmatrix} \mathrm{Re}\{Q^{\frac{1}{2}}K_4\} & \mathrm{Im}\{Q^{\frac{1}{2}}K_4\} \\ -\mathrm{Im}\{Q^{\frac{1}{2}}K_4\} & \mathrm{Re}\{Q^{\frac{1}{2}}K_4\} \end{bmatrix} = \begin{bmatrix} \mathrm{Re}\{K_4^*QK_4\} & \mathrm{Im}\{K_4^*QK_4\} \\ -\mathrm{Im}\{K_4^*QK_4\} & \mathrm{Re}\{K_4^*QK_4\} \end{bmatrix},$$

it can be deduced that $\mathcal{R}(\mathcal{D})$ is the same as the ellipsoid given in (5.11) (note that the matrices $Q^{\frac{1}{2}}K_4$ and K_4 have the same column rank).

Proof of Part (ii): To prove that $\mathcal{R}(\tilde{\mathcal{D}})$ is identical to the ellipsoid given in (5.12), one can adopt the line of arguments made in the proof of Part (i) and take advantage of the fact that a complex vector $\boldsymbol{\alpha}$ belongs to $\tilde{\mathcal{D}}$ if and only if there exist symmetric matrices

$M, N \in \mathbf{R}^{(n-z)\times(n-z)}$ such that

$$\begin{bmatrix} Q-M & N \\ N & M \end{bmatrix} \succ 0 \tag{5.18}$$

and

$$\begin{aligned} \boldsymbol{\alpha} = & -(W_{31}W_{11}^{-1}W_{12} - W_{32})(M + N\mathrm{i}) \begin{bmatrix} W_{21}W_{11}^{-1} & W_{21}W_{11}^{-1}W_{13} - W_{23} \end{bmatrix} \\ & - \begin{bmatrix} W_{31}W_{11}^{-1} & W_{31}W_{11}^{-1}W_{13} - W_{33} \end{bmatrix} \\ = & -K_1(M + N\mathrm{i})K_6 - K_7. \end{aligned} \tag{5.19}$$

The details are omitted for brevity. ■

So far, it is shown that the PCS antenna of node 0 can be programmed to generate a specific voltage vector $(v_1, v_2, ..., v_z)$ if and only if the real-valued representation of this vector belongs to a particular open ellipsoid. This important result completely characterizes the correlation among the voltages received by different nodes of the network. Note that the eigenvalues and eigenvectors of the describing matrix of the ellipsoid $\mathcal{R}(\mathcal{D})$ determines the strength of this correlation in diverse directions. It is worth mentioning that even though "voltage" and "admittance" are disparate quantities, the set of all possible vectors $(v_1, v_2, ..., v_z, y_{\text{in}})$ is again associated with an *ellipsoidal* feasibility region $\mathcal{R}(\tilde{\mathcal{D}})$. Due to the fundamental similarities between the regions $\mathcal{D}$ and $\tilde{\mathcal{D}}$, the focus of this chapter will be only on the region $\mathcal{D}$.

Remark 2 *Theorem 2 states that $\mathcal{R}(\mathcal{D})$ is an ellipsoid in the case when the matrix K_4 has full column rank. A question arises as to how the region $\mathcal{R}(\mathcal{D})$ looks if this condition is violated. To answer this question, notice that a set of feasible voltages $(v_1, v_2, ..., v_z)$ generated by the PCS antenna of node 0 can be represented as*

$$\begin{bmatrix} v_1 & v_2 & \cdots & v_z \end{bmatrix} = -K_1(M + Ni)K_4 - K_5, \tag{5.20}$$

for some symmetric matrices M and N (see the equation (5.14)). The above relation simply implies that if K_4 loses column rank, some of the far-field voltages $v_1, v_2, ..., v_z$ can always be expressed in term of the remaining ones (for every arbitrary matrices M and N). As a result, in the case when K_4 loses column rank, some of the far-field voltages create an *ellipsoidal feasibility region and the remaining far-field voltages can be linearly written in terms of these voltages. Note that the matrix K_4 losses column rank if $n - z$ is less than z, which signifies that the number of controllable ports of the antenna determines the maximum number of directions towards which independent data can be sent.*

Definition 4 *Given $l \in \mathbf{N}$ and distinct indices $j, k \in \{1, 2, ..., l\}$, define $\mathcal{P}^l_{jk}$ to be the plane consisting of all vectors in $\mathbf{R}^l$ whose elements with the indices in the set $\{1, 2, ..., l\} \backslash \{j, k\}$ are equal to zero.*

Theorem 2 can be used to study whether the PCS antenna of node 0 is capable of transmitting data to an intended node in such a way that unintended nodes all receive a zero signal. This is carried out next.

Corollary 1 *Let node 0 in the wireless network employ a PCS antenna to generate a radiation pattern at the far field. The following statements hold for every $j \in \{1, 2, ..., z\}$:*

i) The real-valued representation of the set of all possible complex voltages v_j that can be induced on the antenna of node j is an ellipse (circle) obtained by projecting the ellipsoid $\mathcal{R}(\mathcal{D})$ (given in (5.11)) on the plane $\mathcal{P}^{2z}_{j(z+j)}$.

ii) The real-valued representation of the set of all possible complex voltages v_j that can be induced on the antenna of node j in such a way that other nodes $1, ..., j-1, j+1, .., z$ receive a zero signal is an ellipse (circle) obtained by intersecting the ellipsoid $\mathcal{R}(\mathcal{D})$ (given in (5.11)) with the plane $\mathcal{P}^{2z}_{j(z+j)}$.

Proof: Due to the analogy between the two parts of this corollary, only Part (i) will be proved here. Recall from Theorem 2 that a voltage vector $(v_1, ..., v_z)$ can be generated at the far field by a PCS antenna if and only if the vector $(\text{Re}\{v_1\}, ..., \text{Re}\{v_z\}, \text{Im}\{v_1\}, ..., \text{Im}\{v_z\})$ belongs to the ellipsoid $\mathcal{R}(\mathcal{D})$. Since the plane $\mathcal{P}^{2z}_{j(z+j)}$ corresponds to the vector $(\text{Re}\{v_j\}, \text{Im}\{v_j\})$, one can argue that the set of all possible vectors $(\text{Re}\{v_j\}, \text{Im}\{v_j\})$ is equal to the projection of the ellipsoid $\mathcal{R}(\mathcal{D})$ on the plane $\mathcal{P}^{2z}_{j(z+j)}$. The proof of Part (i) is completed by noting that this projection leads to a circle. ■

Given $j \in \{1, 2, ..., z\}$, the phrase "real-valued representation of the set of all possible complex voltages v_j" in Corollary 1 indeed refers to a constellation digram for v_j. Hence, Corollary 1 characterizes different constellation diagrams that are associated with node j.

Although the projection of the ellipsoid $\mathcal{R}(\mathcal{D})$ on the plane $\mathcal{P}^{2z}_{j(z+j)}$ is a non-empty set, the intersection of these ellipsoid and plane might be a null set. As a result, the PCS antenna of node 0 may not be able to transmit data only to node j so that other nodes all receive a zero signal. However, the feasibility region $\mathcal{R}(\mathcal{D})$ can be used to find a maximum number of nodes that can simultaneously receive a zero signal.

5.3.1 Online Passive Controller Design

So far, different constellation diagrams associated with every node $j \in \{1, 2, ..., z\}$ are obtained and shown to be elliptic (see Corollary 1). Assume that based on these of constellation diagrams, node 0 has decided to generate the far-field voltage vector $(v_1, v_2, ..., v_z) = \boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathcal{D}$. Note that $\boldsymbol{\alpha}$ can, for instance, be a vector with only one non-zero entry corresponding to a directional data transmission. A question arises as to what passive controller should be applied to the PCS antenna of node 0 to make it generate the voltage vector $(v_1, v_2, ..., v_z) = \boldsymbol{\alpha}$. To address this question, let a procedure be introduced.

Procedure 1:

Step 1: Define a vector $\mathbf{u}$ as

$$\mathbf{u} := \left(\begin{bmatrix} \mathrm{Re}\{\boldsymbol{\alpha}\} & \mathrm{Im}\{\boldsymbol{\alpha}\} \end{bmatrix} - \mathbf{o} \right) \begin{bmatrix} \mathrm{Re}\{K_4^* Q K_4\} & \mathrm{Im}\{K_4^* Q K_4\} \\ -\mathrm{Im}\{K_4^* Q K_4\} & \mathrm{Re}\{K_4^* Q K_4\} \end{bmatrix}^{-1} \\ \times \begin{bmatrix} \mathrm{Re}\{K_4^* Q^{\frac{1}{2}}\} & -\mathrm{Im}\{K_4^* Q^{\frac{1}{2}}\} \\ \mathrm{Im}\{K_4^* Q^{\frac{1}{2}}\} & \mathrm{Re}\{K_4^* Q^{\frac{1}{2}}\} \end{bmatrix}.$$

Step 2: By using the procedure presented in the proof of Lemma 1, compute two symmetric matrices $\tilde{M}, \tilde{N} \in \mathbf{R}^{(n-z)\times(n-z)}$ such that

$$\mathbf{u} = \begin{bmatrix} -\frac{1}{2} K_1 Q^{\frac{1}{2}} & \frac{1}{2} K_1 Q^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \tilde{M} & \tilde{N} \\ \tilde{N} & -\tilde{M} \end{bmatrix}$$

and

$$\begin{bmatrix} \tilde{M} & \tilde{N} \\ \tilde{N} & -\tilde{M} \end{bmatrix} \prec I.$$

Step 3: One candidate for the admittance of the passive controller at the frequency

f_0, denoted by Y_0, is

$$Y_0 = 2\left(Q^{\frac{1}{2}}\tilde{M}Q^{\frac{1}{2}} + Q + Q^{\frac{1}{2}}\tilde{N}Q^{\frac{1}{2}}\mathrm{i}\right)^{-1} - W_{22} + W_{21}W_{11}^{-1}W_{12}.$$

The proofs of Lemma 1, Lemma 2, Theorem 1, and Theorem 2 can all be combined in a clear way to deduce why Procedure 1 described above produces a correct admittance Y_0. After obtaining the matrix Y_0, the next question would be how to design a passive controller (circuit) with the admittance Y_0 at the frequency f_0. This can be carried out using the next procedure.

Procedure 2:

Step 1: Decompose Y as

$$Y = T_1 + T_2\mathrm{i},$$

where $T_1, T_2 \in \mathbf{R}^{(n-z)\times(n-z)}$ are both symmetric and T_1 is positive definite.

Step 2: Define an admittance transfer function $Y(s)$ as

$$Y(s) = T_1 + \frac{1}{\omega_0}T_2 s, \quad \forall s \in \mathbf{C}.$$

Step 3: Implement $Y(s)$ by the parallel connection of two $(n-z)$-port networks: (i) a resistive network with the conductance matrix T_1 and (ii) a reactive network with the susceptance matrix $\frac{1}{\omega_0}T_2$.

Step 4: Since $Y(2\pi f_0\mathrm{i}) = Y_0$, the obtained parallel connection of these two separate networks is a candidate for the passive controller being sought.

Hence, Procedures 1 and 2 should be taken to find a passive controller under which the PCS antenna of node 0 generates a desired far-field voltage vector.

5.3.2 Real-Time Data Transmission via a PCS Antenna

How can a real-time data transmission be performed using a PCS Antenna? As before, assume that node 0 is equipped with a PCS Antenna for transmitting possibly distinct pieces of data to the individual nodes $1, 2, ..., z$. Given an index $j \in \{1, 2, .., z\}$, some preprocessing needs to be carried out to find how many different symbols can be sent from node 0 to node j in such a way that many of the remaining nodes receive a zero signal concurrently. To this end, one can exploit Corollary 1 or the ellipsoidal feasibility region $\mathcal{R}(\mathcal{D})$ directly to find some nodes $j_1, j_2, ..., j_k$ that can all receive a zero signal in the course of data transmission to node j. Now, obtain the associated elliptic constellation diagram for node j and pick a maximum number of points inside this constellation diagram such that every point has a certain minimum distance from the origin (or a minimum power) and every two points have a minimum point-to-point distance d, where d is a positive number whose value depends on the noise level of the communication channel. Each of these points in the constellation diagram corresponds to a symbol that can be transmitted to node j using the smart antenna of node 0 in such a way that nodes $j_1, j_2, ..., j_k$ all receive a zero signal. Utilize Procedures 1 and 2 to design controllers corresponding to all these different symbols and then store the values of the controllers' parameters in a look-up table. Note that since Procedures 1 and 2 have low computational complexities, this table can be formed at a very high speed by means of a low-power processor. Now, at every time slot that any of these symbols needs to be transmitted from node 0 to node j, node 0 selects a suitable passive controller (resistive-capacitive-inductive network) from the table and applies it to its PCS antenna to make node j receive the underlying symbol while nodes $j_1, j_2, ..., j_k$ all get a zero signal.

Remark 3 *The real-time data transmission scheme spelled out above is contingent upon the availability of the matrix Y_s to node 0. Unlike Procedures 1 and 2, computing the matrix Y_s is time-consuming and cannot be accomplished in a fraction of second. However, the configuration-dependent matrix Y_s is easily computable (approximately) from a normalized Y-parameter matrix that can be computed offline and stored on the memory of the sensor device. This normalized Y-parameter matrix, denoted by Y_n, is defined as follows:*

- *Discretize the continuous angle interval $[0°, 360°)$ by some quantization level, say 1 degree, and assume that any receiving antenna around the PCS antenna system lies on a discretized direction.*

- *Visualize a configuration where a receiving antenna is placed in every discretized direction at the distance of 1 unit from the PCS antenna. This unit is arbitrary as long as it is larger than a few multiples of the signal wavelength.*

- *Let Y_n be defined as the Y-parameter matrix of this visualized antenna configuration.*

To find an online approximation of Y_s associated with nodes $0, 1, 2, ..., z$ in terms of the normalized matrix Y_n, the following steps should be taken:

- *For every $j \in \{1, 2, ..., z\}$, map the direction from node j to node 0 to the closest quantized direction.*

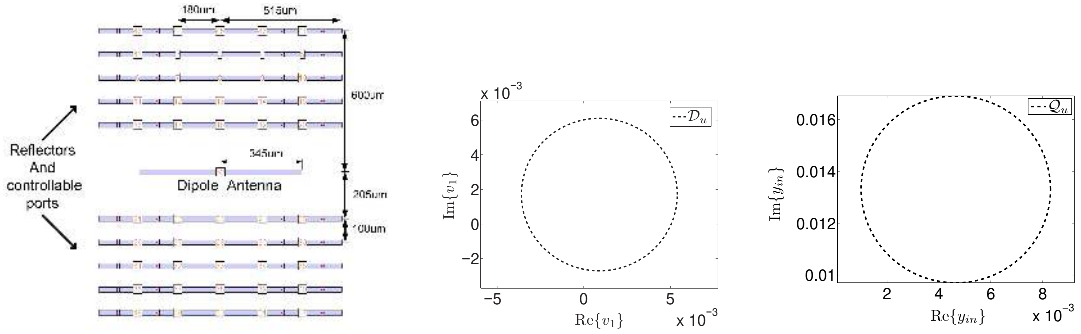
- *After identifying the mapped quantized directions, eliminate the unnecessary rows and columns of the matrix Y_n (associated with the unmapped directions).*

- *The resulting matrix corresponds to the case when nodes $1, 2, ..., z$ were placed at the distance of 1 unit from node 0. To consider the real distances of node $1, 2, .., z$ to node 0, it is enough to note that changing the distance between nodes 0 and j (without changing the direction) only rotates and re-scales the constellation digram (due to power attenuation/amplification and delay). Hence, Y_s can be easily estimated from the obtained matrix by taking the real node-to-node distances into account.*

Remark 4 *Assume that nodes $1, 2, ..., z$ are not static relative to the reference node 0 and all nodes are allowed to be mobile. In this case, it is no longer true that the passive controllers can be designed only one time so that their parameters are stored in a look-up table. Instead, real-time online computation is required to find a proper controller for every signal transmission. The allowable mobility rate of each node mainly depends on how fast the proposed procedures can be completed to design a passive controller, which in turn relies on the speed/power of the node's processor.*

5.3.3 Simultaneous Data Transmission via a PCS Antenna

To describe the idea of simultaneous data transmission using a PCS antenna, consider a simple scenario where node 0 intends to send a symbol A_1 corresponding to the complex voltage α_1 to node 1 and another symbol A_2 corresponding to the complex voltage α_2 to node 2. Assume that $(\alpha_j, 0, 0, ..., 0)$ is inside the ellipsoid $\mathcal{D}$, for $j = 1, 2$. One strategy for this data transmission is that the PCS Antenna of node 0 transmits A_1 and A_2 at two different time slots. However, for the sake of saving time and energy, it is really preferable to somehow transmit these symbols in a single time slot and by only a one-time programming of the smart antenna. To do so, note that if $(\alpha_1, \alpha_2, 0, ..., 0)$ belongs to the ellipsoid $\mathcal{D}$,



(a) (b) (c)

Figure 5.5: (a): The PCS antenna system studied in Example 1; (b): The constellation diagram $\mathcal{R}(\mathcal{D}_u)$; (c): The admittance feasibility region $\mathcal{R}(\mathcal{Q}_u)$

then it is possible to program the PCS antenna of node 0 so that it generates the far-field voltages $v_1 = \alpha_1$, $v_2 = \alpha_2$ and $v_3 = \cdots = v_z = 0$. In other words, it could be possible to simultaneously transmit data to nodes 1 and 2 in such a way that other nodes receive a zero signal. This idea of simultaneous data transmission can be easily generalized to more than two nodes.

5.4 Simulation Results

Two examples will be presented in this section, where Example 1 illustrates the modulation capability of a PCS antenna in a single direction and Example 2 demonstrates the efficacy of a PCS antenna in a wireless sensor network. For the sake of brevity, the detailed specifications of a PCS antenna required for its implementation on a silicon chip is outlined only in Example 2.

Example 1: Consider the PCS antenna system depicted in Figure 5.5(a), which consists of a transmitting dipole antenna and 10 metal reflectors each with 5 ports (antenna parasitic elements). The objective is to transmit data from this PCS antenna to a receiving dipole antenna located at the far field in the upward direction. Assume that the PCS antenna system is driven by a sinusoidal voltage source with the frequency 60 GHz and the amplitude of 1 volt. The circuit model of the antenna system can be extracted at the desired frequency 60 GHz (using localized differential lumped ports) by means of the electromagnetic software

IE3D [41]. This circuit model is in the form of the circuit given in Figure 5.4(a) with 51 output ports, where

- Port 1 is aimed to sample the radiation pattern of the transmitting antenna on the receiving antenna.
- Ports 2 to 51 are the controllable ports of the PCS antenna and are intended to change the boundary condition of the transmitting antenna.

The two parameters of interest in this problem are the voltage induced on the receiving antenna, i.e., v_1, and the input admittance of the PCS antenna, i.e., y_{in}. The goal is to understand what values can be generated for v_1 and y_{in} via a passive control of the PCS antenna and, moreover, how these two parameters are related to one another. To this end, Theorem 2 can be employed to deduce that a complex pair (v_1, y_{in}) can be produced by programming the PCS antenna if and only if

$$\left\| \begin{bmatrix} \text{Re}\{v_1\} & \text{Re}\{y_{\text{in}}\} & \text{Im}\{v_1\} & \text{Im}\{y_{\text{in}}\} \end{bmatrix} \Phi \right\| < 0.0427, \tag{5.21}$$

where

$$\Phi = \begin{bmatrix} 10.2871 & -1.0002 & 0.0000 & 2.6625 \\ -1.0002 & 12.9429 & -2.6625 & -0.0000 \\ 0.0000 & -2.6625 & 10.2871 & -1.0002 \\ 2.6625 & -0.0000 & -1.0002 & 12.9429 \end{bmatrix}.$$

In other words, the real-valued representation of all possible complex vectors (v_1, y_{in}) forms an ellipsoid given by (5.21). Note that the eigenvalues and eigenvectors of the positive-definite matrix Φ completely specify the correlation between v_1 and y_{in}. In what follows, different problems will be studied.

The first goal is to identify two feasibility regions $\mathcal{D}_u$ and $\mathcal{Q}_u$, defined as:

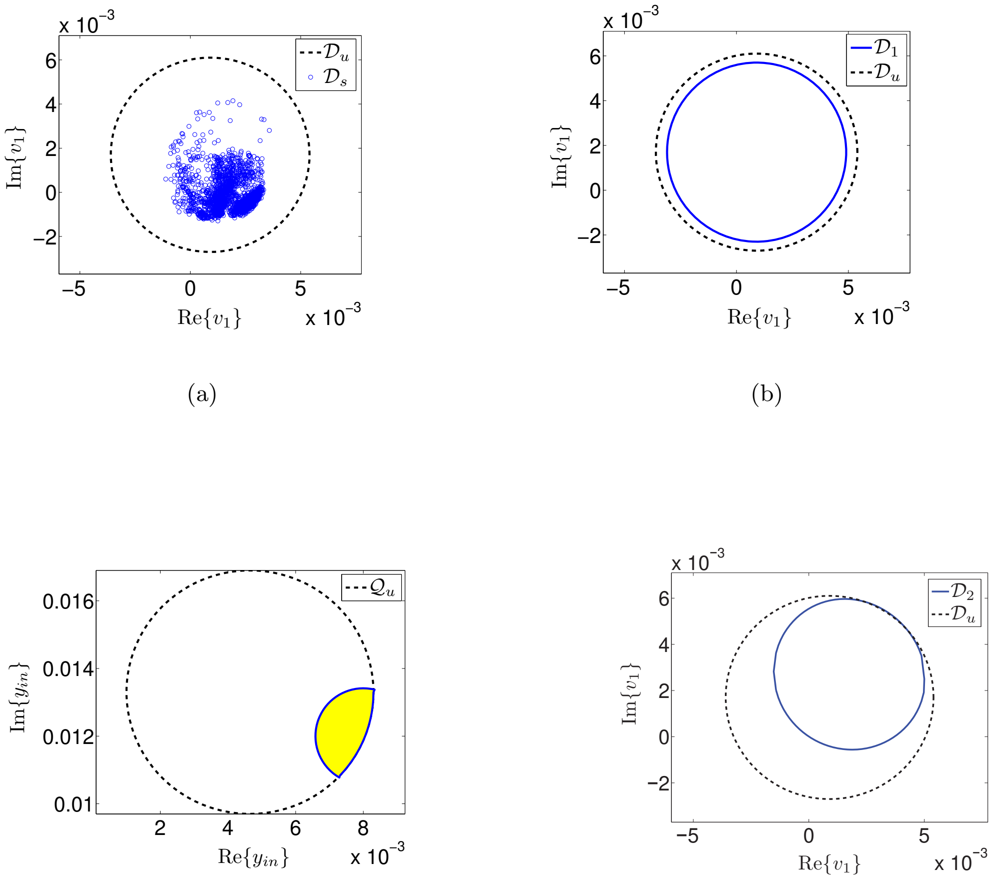
- $\mathcal{D}_u$ (unconstrained constellation diagram): the set of all complex voltages v_1 that can be generated via the underlying PCS antenna;
- $\mathcal{Q}_u$ (admittance feasibility region): the set of all complex input admittances y_{in} that can be generated via the underlying PCS antenna.

One can argue that $\mathcal{R}(\mathcal{D}_u)$ and $\mathcal{R}(\mathcal{Q}_u)$ are indeed the projection of the ellipsoid given in (5.21) on the planes $\mathcal{P}^4_{13}$ and $\mathcal{P}^4_{24}$, respectively. These two regions turn out to be both circular, as depicted in Figures 5.5(b) and 5.5(c). It is noteworthy that in light of the circular shape of $\mathcal{R}(\mathcal{D}_u)$, it is easy to find an optimal number of modulation points in the unconstrained constellation diagram $\mathcal{R}(\mathcal{D}_u)$ so that every point-to-point distance is greater than a prescribed number.

Recall that if the passive controller of the PCS antenna is confined to be only a decoupled switching network, then the resulting PCS antenna reduces to the antenna system proposed in [2] and [3]. It is desired to compare the achievable performances of the PCS antenna given here and the switch-based antenna suggested in the aforementioned papers. To this end, let $\mathcal{D}_s$ denote the set of all values of v_1 that can be generated by the PCS antenna subject to the constraint that each of its controllable ports is connected to an ideal on/off switch. Finding the exact shape of $\mathcal{R}(\mathcal{D}_s)$ requires computing v_1 for 2^{50} switching combinations, which is almost impossible. However, a number of switching combinations are generated at random and the corresponding values of v_1 are plotted in Figure 5.6(a). It can be seen that even though a passive network has far more free parameters than a switching network, the discrete set $\mathcal{R}(\mathcal{D}_s)$ is fairly dense in a big part of the continuous set $\mathcal{R}(\mathcal{D}_u)$. This observation is no longer valid if the number of receiving antennas is not as small as 1, in which case the discrete set $\mathcal{R}(\mathcal{D}_s)$ will become sparse in the high-dimensional set $\mathcal{R}(\mathcal{D}_u)$. As mentioned earlier, the programming of the PCS antenna under a switching controller is an NP-complete problem, whereas its programming under an arbitrary passive controller can be cast as a simple convex optimization problem whose solution is known analytically.

As shown in Figure 5.5(c), the admittance feasibility region is a circle. Assume that the input admittance of the PCS antenna is required to be matched with the admittance corresponding to the center of this circle. Since this matching constraint enforces y_{in} to be fixed, one may speculate that the corresponding constrained constellation diagram for v_1, denoted by $\mathcal{D}_1$, is noticeably smaller than $\mathcal{D}_u$. However, it is interesting to note that the set $\mathcal{D}_1$ is only a little smaller than $\mathcal{D}_u$, as illustrated in Figure 5.6(b). This demonstrates one of the advantages of deploying a *passively controllable* smart antenna.

As the last scenario, assume that the impedance of the input voltage source of the PCS antenna is equal to the standard value 50Ω. Since $\frac{1}{50}$ is outside the admittance feasibility region given in Figure 5.5(c), the objective is to find an optimal antenna input admittance minimizing the reflection factor $\|T\|$. Based on the circular region $\mathcal{R}(\mathcal{Q}_u)$, it can be shown that the optimal input admittance of the PCS antenna is equal to $0.008 + 0.012\mathrm{i}$, which is



(a)

(b)

(c)

(d)

Figure 5.6: (a): The constellation diagrams $\mathcal{R}(\mathcal{D}_s)$ and $\mathcal{R}(\mathcal{D}_u)$; (b): The constellation diagrams $\mathcal{R}(\mathcal{D}_1)$ and $\mathcal{R}(\mathcal{D}_u)$; (c): The admittance feasibility region associated with the constraint that the antenna input admittance belongs to a circle centered at $0.008 + 0.012\text{i}$ with radius $\sqrt{2} \times 10^{-3}$ (colored area); (d): The constellation diagrams $\mathcal{R}(\mathcal{D}_2)$ and $\mathcal{R}(\mathcal{D}_u)$

not attainable due to lying on the boundary of the open region $\mathcal{R}(\mathcal{Q}_u)$. In order to have a near-optimal antenna input admittance, suppose that y_{in} is permitted to deviate from $0.008 + 0.012\text{i}$ by at most $\sqrt{2} \times 10^{-3}$. The corresponding constrained admittance feasibility region is shown in Figure 5.6(c) (see the colored area). The set of all possible voltages v_1 generated by the PCS antenna under this admittance matching constraint, denoted by $\mathcal{D}_2$, can be obtained using the ellipsoid given in (5.21), which is plotted in Figure 5.6(d). It can be seen that $\mathcal{R}(\mathcal{D}_2)$ covers a big part of $\mathcal{R}(\mathcal{D}_u)$ and that $\mathcal{R}(\mathcal{D}_2)$ is a non-circular region.

Example 2: The goal of this example is to demonstrate the efficacy of PCS antennas on a cluster of 12 wireless sensor nodes, labeled as $0, 1, ..., 11$. Let the configuration of the sensor nodes in the x-y plane be as follows:

- Node 0 lies at the origin of the x-y plane.

- Nodes $1, 2, ..., 11$ are all located at the distance of 20 units from node 0 with the angular directions $15°, 30°, 45°, 60°$, $75°, 90°, 105°, 120°, 135°, 150°, 165°$ (with respect to the x-axis).

The configuration associated with the above-mentioned scenario is depicted in Figure 5.7(a). It is desired to transmit data from node 0 to node 6 (i.e., in the vertical direction) in such a way that some of the remaining nodes $\{1, ..., 5, 7, ..., 11\}$ receive a zero signal, if possible. To this end, equip node 0 with the PCS antenna system given in Figure 5.7(b), which has the following specifications:

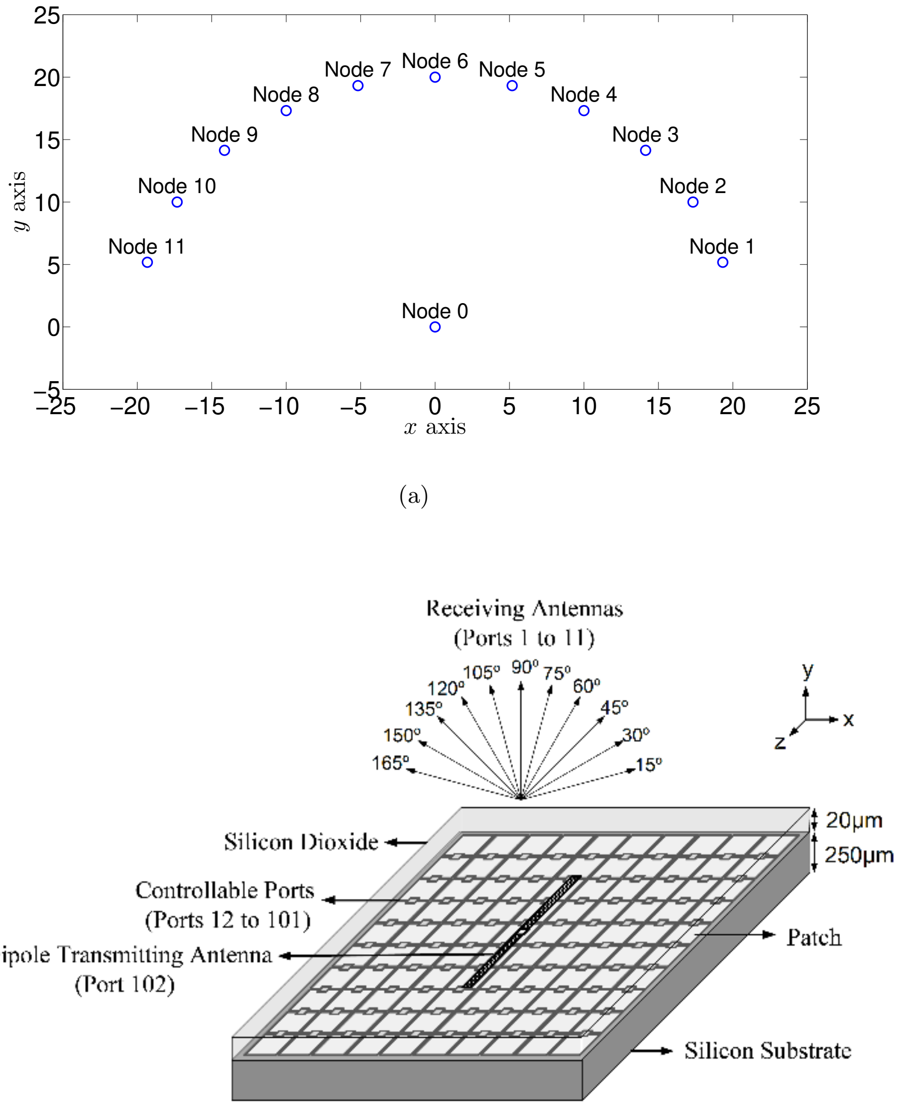
- It consists of a 500 μm on-chip transmitting dipole antenna located at 20 μm above a 1mm × 1mm on-chip patch array.

- The patch array is located on the x-z plane and comprises 100 metal squares, each with a dimension of 95 μm by 95 μm.

- Every two adjacent patches in the z-direction are connected with a controllable port resulting in a total number of 90 ports.

- The ground plane is located at $z = 0$, and a 250 μm 10 Ω-cm silicon substrate is located right above the ground layer.

- A 20 μm silicon-dioxide (SiO_2) layer is placed on top of the silicon substrate as shown in Figure 5.7(b).



(a)

(b)

Figure 5.7: (a): The configuration of the sensor nodes studied in Example 2; (b): The smart antenna deployed for node 0 in Example 2

- The transmitting dipole antenna is placed just above the SiO_2 layer.

Note that although the theory developed in the present work relies on reflectors to shape the electromagnetic field around the PCS antenna, a patch array can alternatively be used instead of reflectors, as done in this example. Let the transmitting dipole antenna of the PCS antenna system be driven by a 300 GHz sinusoidal signal with a fixed amplitude of 1 volt. Notice that since the PCS antenna is designed to be very small, the operating frequency of the antenna is chosen very high to make its size comparable to the wavelength of the transmitted signal. Assume that each of the nodes $1, 2, ..., 11$ has a receiving dipole antenna perpendicular to the x-y place with the length of 0.44 mm and the fixed terminal impedance 50 Ω. If node 0 knows the locations of the other nodes (practically, their directions), then the equivalent circuit model of the described wireless network can be extracted at node 0. This leads to the circuit given in Figure 5.4(a) with 101 output ports, where

- For every $j \in \{1, 2, ..., 11\}$, output port j measures the voltage induced by the smart antenna of node 0 on the center of the receiving dipole antenna of node j.
- Output ports 12 to 101 are the 90 controllable ports on the smart antenna of node 0 that are to be controlled by a passive controller for every signal transmission.

The real-valued representation of the set of all possible voltages $(v_1, v_2, ..., v_{11})$ that can be generated by the PCS antenna of node 0 is an ellipsoid denoted by $\mathcal{R}(\mathcal{D})$, which can be obtained using Theorem 2. Based on this ellipsoid, a number of problems are addressed in the sequel.

First, consider the problem of maximizing the power of the signal received by node 6, which amounts to the maximization of the scalar $\|v_6\|$. Although the maximization of the 2-norm of a signal is normally a non-convex problem, it will be shown here that the underlying power optimization problem will be convex. Indeed, the real-valued representation of the set of all possible complex voltages v_6 that can be generated by the PCS antenna of node 0 is a circle obtained by projecting the ellipsoid $\mathcal{R}(\mathcal{D})$ on the plane $\mathcal{P}^{22}_{6,17}$. This circle is centered at $\left(-0.53 \times 10^{-3}, 0.87 \times 10^{-3}\right)$ with radius 4.96×10^{-3}. Hence, the problem of maximizing the power received by node 6 reduces to finding the farthest point of this circle from the origin. This yields the point $\left(-3.11 \times 10^{-3}, 5.10 \times 10^{-3}\right)$ or, equivalently, the complex voltage $v_6 = (-3.11 + 5.10\mathrm{i}) \times 10^{-3}$.

Table 5.1: Maximizing the power of the signal received by node 6 while forcing one of the nodes $\{1, 2, ..., 11\}\backslash\{6\}$ to receive a zero signal (Example 2)

	Constraints	Circular Constellation Diagram for v_6	Optimal value of v_6
Case 1	No constraint	Center: $\left(-0.53 \times 10^{-3}, 0.87 \times 10^{-3}\right)$, Radius: 4.96×10^{-3}	$v_6 = (-3.11 + 5.10\mathrm{i}) \times 10^{-3}$
Case 2	$v_1 = 0$	Empty	None
Case 3	$v_2 = 0$	Empty	None
Case 4	$v_3 = 0$	Center: $\left(-1.49 \times 10^{-3}, 1.04 \times 10^{-3}\right)$, Radius: 2.96×10^{-3}	$v_6 = (-3.92 + 2.74\mathrm{i}) \times 10^{-3}$
Case 5	$v_4 = 0$	Center: $\left(-1.53 \times 10^{-3}, 0.60 \times 10^{-3}\right)$, Radius: 3.91×10^{-3}	$v_6 = (-5.17 + 2.03\mathrm{i}) \times 10^{-3}$
Case 6	$v_5 = 0$	Center: $\left(-0.14 \times 10^{-3}, 0.60 \times 10^{-3}\right)$, Radius: 2.91×10^{-3}	$v_6 = (-0.81 + 3.43\mathrm{i}) \times 10^{-3}$
Case 7	$v_7 = 0$	Center: $\left(-0.14 \times 10^{-3}, 0.60 \times 10^{-3}\right)$, Radius: 2.91×10^{-3}	$v_6 = (-0.81 + 3.43\mathrm{i}) \times 10^{-3}$
Case 8	$v_8 = 0$	Center: $\left(-1.53 \times 10^{-3}, 0.60 \times 10^{-3}\right)$, Radius: 3.91×10^{-3}	$v_6 = (-5.17 + 2.03\mathrm{i}) \times 10^{-3}$
Case 9	$v_9 = 0$	Center: $\left(-1.49 \times 10^{-3}, 1.04 \times 10^{-3}\right)$, Radius: 2.96×10^{-3}	$v_6 = (-3.92 + 2.74\mathrm{i}) \times 10^{-3}$
Case 10	$v_{10} = 0$	Empty	None
Case 11	$v_{11} = 0$	Empty	None

Given a node $j \in \{1, 2, ..., 11\}\backslash\{6\}$, the goal of the second problem is to maximize the power of the signal received by node 6 subject to the constraint that node j receives a zero signal. Similar to the previous problem, one needs to find the intersection of the ellipsoid $\mathcal{R}(\mathcal{D})$ with the plane $\mathcal{P}^{22}_{6,17}$ and then search for the farthest point of the resultant circular region from the origin. The results obtained for this problem (for different values of j) together with that of the previous problem are summarized in Table 5.1. It can be observed that the circular constellation diagrams corresponding to Cases $2, 3, 10, 11$ are all empty, meaning that it is not possible to program the antenna of node 0 to send data to node 6 so that any of the nodes $1, 2, 10, 11$ receives a zero signal. In contrast, it is possible to send a zero signal to each of the nodes $3, 4, 5, 7, 8, 9$. Comparing Case 1 with Cases 4–9 in Table 5.1, one can draw the interesting conclusion that imposing an extra constraint $v_j = 0$, $j \in \{3, 4, 5, 7, 8, 9\}$, does not shrink the constellation diagram noticeably (i.e., the radius of the circle corresponding to each of Cases 4–10 is not far smaller than that for Case 1).

Due to the symmetry of the locations of nodes $1, 2, ..., 11$ with respect to the reference node 0, it was not possible to simultaneously transmit data to node 6 and send a zero signal to some of the unintended nodes $\{1, 2, ..., 11\}\backslash\{6\}$. Now, let the configuration of the sensor nodes be distorted by moving node 0 to the coordinate $(0, 5)$. The problem of transmitting data to node 6 while sending a zero signal to one of the nodes $\{1, 2, ..., 11\}\backslash\{6\}$ is investigated for the new configuration and the results are summarized in Table 5.2. Two observations can be made here as follows:

Table 5.2: Studying the possibility of transmitting data to node 6 while forcing one of the nodes $\{1, 2, ..., 11\} \backslash \{6\}$ to receive a zero signal (corresponding to the second configuration in Example 2)

	Constraints	Circular Constellation Diagram for v_6
Case 1	No constraint	Center: $\big(1.02 \times 10^{-3}, -1.73 \times 10^{-3}\big)$, Radius: 4.97×10^{-3}
Case 2	$v_1 = 0$	Center: $\big(1.06 \times 10^{-3}, -1.93 \times 10^{-3}\big)$, Radius: 4.69×10^{-3}
Case 3	$v_2 = 0$	Center: $\big(1.16 \times 10^{-3}, -1.82 \times 10^{-3}\big)$, Radius: 4.71×10^{-3}
Case 4	$v_3 = 0$	Center: $\big(1.27 \times 10^{-3}, -1.52 \times 10^{-3}\big)$, Radius: 4.64×10^{-3}
Case 5	$v_4 = 0$	Center: $\big(0.56 \times 10^{-3}, -1.58 \times 10^{-3}\big)$, Radius: 3.81×10^{-3}
Case 6	$v_5 = 0$	Center: $\big(2.80 \times 10^{-3}, +0.35 \times 10^{-3}\big)$, Radius: 2.21×10^{-3}
Case 7	$v_7 = 0$	Center: $\big(1.60 \times 10^{-3}, -0.95 \times 10^{-3}\big)$, Radius: 2.78×10^{-3}
Case 8	$v_8 = 0$	Center: $\big(1.08 \times 10^{-3}, -1.10 \times 10^{-3}\big)$, Radius: 3.62×10^{-3}
Case 9	$v_9 = 0$	Center: $\big(-0.32 \times 10^{-3}, 0.16 \times 10^{-3}\big)$, Radius: 1.18×10^{-3}
Case 10	$v_{10} = 0$	Center: $\big(0.83 \times 10^{-3}, -1.26 \times 10^{-3}\big)$, Radius: 4.40×10^{-3}
Case 11	$v_{11} = 0$	Center: $\big(0.95 \times 10^{-3}, -1.12 \times 10^{-3}\big)$, Radius: 4.36×10^{-3}

- It is always possible to transmit data to node 6 in such a way that any of the remaining nodes $1, ..., 5, 7, ..., 11$ receives a zero signal.

- The radius of the constellation diagram corresponding to the set of all possible values of v_6 (Case 1) is not remarkably larger than the radius of the constellation diagram corresponding to each of Cases 2-11 in which a node $j \in \{1, ..., 5, 7, ..., 11\}$ is required to receive a zero signal.

It can be shown that it is not possible to send data to node 6 such that all of the remaining nodes receive a zero signal concurrently. However, as an example, the smart antenna of node 0 can be programmed so that nodes $1, 2, 3, 7, 8$ all receive a zero signal simultaneously. More precisely, the real-valued representation of the set of all possible values of v_6 subject to the constraint $v_1 = v_2 = v_3 = v_7 = v_8 = 0$ is a circle centered at $(1.13 \times 10^{-3}, -0.83 \times 10^{-3})$ with radius 0.70×10^{-3}. Now, one can select some points in this circular constellation diagram corresponding to the number of symbols to be sent from node 0 to node 6. For transmitting each of these symbols, it is enough to apply a proper passive controller to the smart antenna of node 0 to make node 6 receive the correct symbol while nodes $1, 2, 3, 7, 8$ all receive a zero signal. Detailed discussions are provided in Sections 5.3.1 and 5.3.2 to shed light on the design of this passive controller and the real-time data transmission via a PCS antenna.

5.5 Summary

This work proposes a new type of smart antenna system, referred to as *passively controllable smart (PCS) antenna*, which can be used as an efficient transmission device in wireless sensor networks. A PCS antenna system is accompanied by a tunable passive controller whose adjustment at every signal transmission generates a possibly unique radiation pattern. To reduce co-channel interference and optimize the transmitted power, this antenna can be programmed to transmit data in a desired direction in such a way that no signal is transmitted (to the far field) at many pre-specified undesired directions. In particular, it is shown that a set of voltage signals can be sent to different directions if and only if a linear matrix inequality problem is feasible. Later on, this result is exploited to prove that a set of voltages can be generated at the far field if and only if the associated vector of voltages belongs to an ellipsoidal region. This region can be computed at a very high speed online by the transmitting sensor node in order to program its PCS antenna for sending data towards an intended node in such a way that a zero signal is sent in several undesired directions. The PCS antenna proposed here is made of only one active element and its programming has a low complexity. These two properties differentiate a PCS antenna from the existing smart antennas, and make it possible to implement a PCS antenna on a cheap, small-sized, low-power silicon chip.

5.6 Appendix

Proof of Lemma 1: Denote the open unit ball $\{\boldsymbol{\gamma} \in \mathbf{R}^{1\times 2m} \mid \|\boldsymbol{\gamma}\| < 1\}$ with $\mathcal{B}^{2m}$. In order to show that the set of every vector $\boldsymbol{\alpha}$ representable in the form (5.5) subject to the constraint (5.6) is equal to $\mathcal{B}^{2m}$, it suffices to prove that every point in this set belongs to $\mathcal{B}^{2m}$ and vice versa. This will be performed in two phases. First, consider an arbitrary vector $\boldsymbol{\alpha} \in \mathbf{R}^{1\times 2m}$ for which there exist symmetric matrices $M, N \in \mathbf{R}^{m\times m}$ such that the relations (5.5) and (5.6) both hold. The goal of this step is to prove that $\boldsymbol{\alpha}$ is in the open ball $\mathcal{B}^{2m}$. Notice that since the matrix

$$\begin{bmatrix} M & N \\ N & -M \end{bmatrix} \tag{5.22}$$

is Hamiltonian, its eigenvalues are all symmetric with respect to the imaginary axis in the complex plane. This property, together with the inequality (5.6), yields that the eigenvalues of this Hamiltonian (and Hermitian) matrix all lie in the interval $(-1,1)$. As a result,

$$\begin{bmatrix} M & N \\ N & -M \end{bmatrix}^2 \prec I.$$

Therefore,

$$\boldsymbol{\alpha}\boldsymbol{\alpha}^* = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \begin{bmatrix} M & N \\ N & -M \end{bmatrix}^2 \begin{bmatrix} \mathbf{x}_1^* \\ \mathbf{x}_2^* \end{bmatrix} < \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^* \\ \mathbf{x}_2^* \end{bmatrix} = 1.$$

This proves that $\boldsymbol{\alpha}$ belongs to $\mathcal{B}^{2m}$. As the second step of the proof, assume that $\boldsymbol{\beta} \in \mathbf{R}^{1\times 2m}$ is an arbitrary vector in the ball $\mathcal{B}^{2m}$. The objective is to show that there exist two symmetric matrices $M, N \in \mathbf{R}^{m\times m}$ satisfying the relation (5.6) such that

$$\boldsymbol{\beta} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \begin{bmatrix} M & N \\ N & -M \end{bmatrix}. \tag{5.23}$$

A constructive proof will be provided here. Decompose the vector $\boldsymbol{\beta}$ as $[\ \boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2\ ]$, where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbf{R}^{1\times m}$. Since the symmetric matrix $\|\boldsymbol{\beta}\|^2\mathbf{x}_1^*\mathbf{x}_1 + \|\boldsymbol{\beta}\|^2\mathbf{x}_2^*\mathbf{x}_2 - \boldsymbol{\beta}_1^*\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^*\boldsymbol{\beta}_2$ is the sum of four rank-one matrices, it has at most four nonzero eigenvalues. Denote the eigenvalues of this matrix with $\gamma_1, \gamma_2, ..., \gamma_m$, where $\gamma_5 = \cdots = \gamma_m = 0$. Let $\mathbf{q}_j$ represent the unit right eigenvector of the above matrix corresponding to the eigenvalue γ_j, for every $j \in \{1,2,...,m\}$. To simplify the proof by avoiding special cases, assume that $m \geq 4$. Define

$$\begin{aligned} &\mathbf{p}_1 := \frac{1}{2}(-\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 + \mathbf{q}_4), \quad \mathbf{p}_2 := \frac{1}{2}(+\mathbf{q}_1 - \mathbf{q}_2 + \mathbf{q}_3 + \mathbf{q}_4), \\ &\mathbf{p}_3 := \frac{1}{2}(+\mathbf{q}_1 + \mathbf{q}_2 - \mathbf{q}_3 + \mathbf{q}_4), \quad \mathbf{p}_4 := \frac{1}{2}(+\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 - \mathbf{q}_4), \\ &\mathbf{p}_j := \mathbf{q}_j, \quad \forall j \in \{5,...,m\}. \end{aligned} \tag{5.24}$$

It is straightforward to verify that

$$\mathbf{p}_j^*\mathbf{p}_j = 1, \quad \mathbf{p}_j^*\mathbf{p}_k = 0, \quad \forall j,k \in \{1,2,...,m\},\ j \neq k. \tag{5.25}$$

Define the matrix P as $\begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_m \end{bmatrix}$. It can be concluded from (5.25) that $PP^* = I$. Let $\lambda_1, ..., \lambda_m, \bar{\lambda}_1, ..., \lambda_m$ be some scalars given by the equation

$$\begin{bmatrix} \lambda_j \\ \bar{\lambda}_j \end{bmatrix} = \frac{1}{\|\boldsymbol{\beta}\|} \begin{bmatrix} \mathbf{x}_1\mathbf{p}_j & \mathbf{x}_2\mathbf{p}_j \\ -\mathbf{x}_2\mathbf{p}_j & \mathbf{x}_1\mathbf{p}_j \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\beta}_1\mathbf{p}_j \\ \boldsymbol{\beta}_2\mathbf{p}_j \end{bmatrix}, \quad \forall j \in \{1, 2, ..., m\}. \tag{5.26}$$

It is desired to show that the relations (5.6) and (5.23) are satisfied if M and N are taken as follows:

$$M = \|\boldsymbol{\beta}\| P \times \operatorname{diag}(\lambda_1, \lambda_2, ..., \lambda_m) \times P^*, \quad N = \|\boldsymbol{\beta}\| P \times \operatorname{diag}(\bar{\lambda}_1, \bar{\lambda}_2, ..., \bar{\lambda}_m) \times P^*. \tag{5.27}$$

For this purpose, it results from the equation (5.26) that

$$\lambda_j^2 + \bar{\lambda}_j^2 = \frac{\|\boldsymbol{\beta}_1\mathbf{p}_j\|^2 + \|\boldsymbol{\beta}_2\mathbf{p}_j\|^2}{\|\boldsymbol{\beta}\|^2\left(\|\mathbf{x}_1\mathbf{p}_j\|^2 + \|\mathbf{x}_2\mathbf{p}_j\|^2\right)}, \quad \forall j \in \{1, 2, ..., m\}. \tag{5.28}$$

On the other hand, one can write

$$\mathbf{p}_j^*\left(\|\boldsymbol{\beta}\|^2\mathbf{x}_1^*\mathbf{x}_1 + \|\boldsymbol{\beta}\|^2\mathbf{x}_2^*\mathbf{x}_2 - \boldsymbol{\beta}_1^*\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^*\boldsymbol{\beta}_2\right)\mathbf{p}_j = 0, \quad \forall j \in \{5, ..., m\}, \tag{5.29}$$

due to the equalities $\mathbf{p}_j = \mathbf{q}_j$ and $\gamma_j = 0$. Given an index $j \in \{1, 2, 3, 4\}$, it can be verified that

$$\begin{aligned} \mathbf{p}_j^*\left(\|\boldsymbol{\beta}\|^2\mathbf{x}_1^*\mathbf{x}_1 + \|\boldsymbol{\beta}\|^2\mathbf{x}_2^*\mathbf{x}_2 - \boldsymbol{\beta}_1^*\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^*\boldsymbol{\beta}_2\right)\mathbf{p}_j &= \sum_{k=1}^{4}\gamma_k = \sum_{k=1}^{m}\gamma_k \\ &= \operatorname{trace}\left(\|\boldsymbol{\beta}\|^2\mathbf{x}_1^*\mathbf{x}_1 + \|\boldsymbol{\beta}\|^2\mathbf{x}_2^*\mathbf{x}_2 - \boldsymbol{\beta}_1^*\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^*\boldsymbol{\beta}_2\right) \\ &= \|\boldsymbol{\beta}\|^2\|\mathbf{x}_1\|^2 + \|\boldsymbol{\beta}\|^2\|\mathbf{x}_2\|^2 - \|\boldsymbol{\beta}_1\|^2 - \|\boldsymbol{\beta}_2\|^2 \\ &= \|\boldsymbol{\beta}\|^2 - \|\boldsymbol{\beta}_1\|^2 - \|\boldsymbol{\beta}_2\|^2 = 0. \end{aligned} \tag{5.30}$$

Hence, it can be concluded from (5.29) and (5.30) that

$$\|\boldsymbol{\beta}\|^2\left(\|\mathbf{x}_1\mathbf{p}_j\|^2 + \|\mathbf{x}_2\mathbf{p}_j\|^2\right) = \|\boldsymbol{\beta}_1\mathbf{p}_j\|^2 + \|\boldsymbol{\beta}_2\mathbf{p}_j\|^2, \quad \forall j \in \{1, 2, ..., m\}. \tag{5.31}$$

Combining (5.28) and (5.31) yields

$$\lambda_j^2 + \bar{\lambda}_j^2 = 1, \quad \forall j \in \{1, 2, ..., m\}.$$

The above equation, along with the relation PP^*, leads to the fact that the matrices M and N introduced earlier satisfy the equality

$$\begin{bmatrix} M & N \\ N & -M \end{bmatrix}^2 = \|\boldsymbol{\beta}\|^2 I,$$

which implies that the Hamiltonian matrix (5.22) has m eigenvalues at $\|\boldsymbol{\beta}\|$ and m eigenvalues at $-\|\boldsymbol{\beta}\|$. Since $\|\boldsymbol{\beta}\|$ is strictly less than 1, it can be inferred that the inequality (5.6) holds for this choice of M and N. Now, it remains to show that the equation (5.23) is also satisfied. To this end, simplify the equation (5.26) to obtain

$$\|\boldsymbol{\beta}\|\lambda_j\mathbf{x}_1\mathbf{p}_j + \|\boldsymbol{\beta}\|\bar{\lambda}_j\mathbf{x}_2\mathbf{p}_j = \boldsymbol{\beta}_1\mathbf{p}_j,\ \forall j \in \{1, ..., m\}, \tag{5.32a}$$

$$\|\boldsymbol{\beta}\|\bar{\lambda}_j\mathbf{x}_1\mathbf{p}_j - \|\boldsymbol{\beta}\|\lambda_j\mathbf{x}_2\mathbf{p}_j = \boldsymbol{\beta}_2\mathbf{p}_j,\ \forall j \in \{1, ..., m\}, \tag{5.32b}$$

or equivalently

$$\mathbf{x}_1 M + \mathbf{x}_2 N = \boldsymbol{\beta}_1, \quad \mathbf{x}_1 N - \mathbf{x}_2 M = \boldsymbol{\beta}_2. \tag{5.33a}$$

The above equations show the validity of (5.23), which completes the proof. ■

Part III

Distributed Computation

Chapter 6

Quantized Consensus by Means of Gossip Algorithm

This chapter deals with the distributed averaging problem over a connected network of agents, subject to a quantization constraint. It is assumed that at each time update, only a pair of agents can update their own states in terms of the quantized data being exchanged. The agents are also required to communicate with one another in a stochastic fashion. It is shown that a quantized consensus is reached for an arbitrary quantizer by means of the stochastic gossip algorithm proposed in a recent paper. The expected value of the time at which a quantized consensus is reached is lower and upper bounded in terms of the topology of the graph for a uniform quantizer. In particular, it is shown that these bounds are related to the principal submatrices of the weighted Laplacian matrix. A convex optimization is also proposed to determine a set of probabilities used to pick a pair of agents that leads to a fast convergence of the gossip algorithm.

6.1 Introduction

During the past few decades, there has been a particular interest in the area of distributed computation, which aims to compute some quantity over a network of processors in a decentralized fashion [107, 108, 69, 105]. The distributed averaging problem, as a particular case, is concerned with computing the average of numbers owned by the agents of a group [82, 81]. This problem has been investigated through the notion of *consensus* in several papers, motivated by different applications [57, 103, 8, 89, 96, 80]. For instance, the synchronization of coupled oscillators, arising in biophysics, neurobiology, and systems biology, is studied in [57] and [103] to explore how to reach a consensus on the frequencies of some agents. Moreover, the problem of aligning the heading angles of a group of mobile agents (e.g., a flock of birds) is treated in [47]. Given a sensor network comprising a set of sensors measuring the same quantity in a noisy environment, the problem of consensus on state estimates is discussed in [100]. The consensus problem for networks of dynamic agents with fixed and switching topologies is tackled in [82], where it is shown that the convergence rate is related to the algebraic connectivity of the network. The work [20] elaborates the relationship between the amount of information exchanged by the agents and the rate of convergence to a consensus. A more complete survey on this topic is given in the recent paper [81].

Consider the distributed average consensus problem in which the values associated with a set of agents are to be averaged in a distributed fashion. Since it may turn out in some applications that all agents cannot update their numbers synchronously, gossip algorithms have been widely exploited by researchers to handle the averaging problem asynchronously [107, 11]. This type of algorithm selects a pair of agents at each time instant and updates their values based on some averaging policy. The consensus problem in the context of gossip algorithms has been thoroughly investigated in the literature [12, 7, 52, 28]. For instance, the work [12] studies the convergence of a general randomized gossip algorithm, and derives conditions under which the algorithm converges. That paper also shows that the averaging time of a gossip algorithm depends on the second largest eigenvalue of a doubly stochastic matrix characterizing the algorithm.

In light of communication constraints, the data being exchanged between a pair of agents is normally quantized. This has given rise to the emergence of quantized gossip algorithms. The notion of *quantized consensus* is introduced in [52] for the case when quantized values (integers) are to be averaged over a connected network with digital communication channels. That paper shows that a quantized gossip algorithm leads to reaching a quantized consensus. This result is extended in [28] to the case when the quantization is uniform, and the initial values of the agents are reals (as opposed to being integers). The paper [28] shows that the quantized gossip algorithm works for a particular choice of the updating parameter, although it conjectures that this result is true for a wide range of updating parameters. A related paper on quantized consensus gives a synchronous algorithm in order to reach a consensus with arbitrary precision, at the cost of not preserving the average of the initial numbers [22].

In this chapter, a weighted connected graph is considered together with a set of scalars sitting on its vertices. The weight of each edge represents the probability of establishing a communication between its corresponding vertices through the updating procedure. First, it is shown that a quantized consensus is reached under the stochastic gossip algorithm proposed in [28], for a wide range of updating parameters and an arbitrary quantizer. The convergence time of the gossip algorithm is then studied. More precisely, consider the expected value of the time at which a quantized consensus is reached, and take its maximum over all possible initial states belonging to a given hypercube. Lower and upper bounds on this quantity are provided for a uniform quantizer, which turn out to be related to the Laplacian of the weighted graph. The upper bound is then minimized in order to obtain the best weights resulting in a small convergence time. To do so, a convex optimization problem is proposed, which can be solved by a semidefinite program.

6.2 Problem Formulation

Consider an undirected connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the set of vertices $\mathcal{V} := \{1, 2, ..., n\}$ and the set of edges $\mathcal{E} \subseteq \{(i,j)|\ i,j \in \mathcal{V}\}$. Suppose that every edge $(i,j) \in \mathcal{E}$ of the graph is associated with a strictly positive number $p_{(i,j)}$ such that

$$\sum_{(i,j)\in\mathcal{E},\ i<j} p_{(i,j)} = 1$$

(note that since $\mathcal{G}$ is undirected, if $(i,j) \in \mathcal{E}$, then $(j,i) \in \mathcal{E}$). These numbers induce a discrete probability distribution $\mathcal{P} := \{p_{(i,j)}|\ (i,j) \in \mathcal{E},\ i < j\}$ on the edges of the graph $\mathcal{G}$, which can be used to specify by what probability an edge can be chosen at random from the set $\mathcal{E}$. Assume that a real number x_i has been assigned to vertex i of $\mathcal{G}$, for every $i \in \mathcal{V}$. Define $\mathbf{X}_0$ as $\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$. In this chapter, the sets of natural, integer and real numbers are denoted by $\mathbf{N}$, $\mathbf{Z}$ and $\mathbf{R}$, respectively. Let $q(x) : \mathbf{R} \to \mathbf{R}$ be a general quantization operator characterized as

$$q(x) = \begin{cases} L_i & \text{if} \quad x \in [L_i, \bar{L}_i] \\ L_{i+1} & \text{if} \quad x \in (\bar{L}_i, L_{i+1}] \end{cases} \qquad \forall i \in \mathbf{Z}, \tag{6.1}$$

where $\{L_i\}_{-\infty}^{\infty}$ is a monotonically increasing sequence of integers representing the quantization levels (that is unbounded from both below and above) and

$$\bar{L}_i := \frac{L_i + L_{i+1}}{2}, \quad \forall i \in \mathbf{Z}.$$

The scalar quantities L_i and $\bar{L}_i$ will be referred to as *level* and *splitting level*, respectively. Assume that the numbers on the vertices of $\mathcal{G}$ are updated at each discrete time instant according to some rule. For every time $k \in \{0\} \cup \mathbf{N}$, let $x_i[k]$ denote the number associated with vertex i at time k and $\mathbf{X}[k] := \begin{bmatrix} x_1[k] & x_2[k] & \cdots & x_n[k] \end{bmatrix}$ denote the state of the graph system at time k.

Given a fixed parameter ε, define an action function $A : \mathbf{R}^n \times \mathcal{E} \to \mathbf{R}^n$ on the graph $\mathcal{G}$ as follows: for every arbitrary vector $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix} \in \mathbf{R}^n$ and edge $(i,j) \in \mathcal{E}$, the quantity $A(\boldsymbol{\alpha}, (i,j))$ is an n-tuple whose p^{th} entry is equal to α_p for every $p \in \mathcal{V} \backslash \{i,j\}$, and whose i^{th} and j^{th} entries are equal to $\alpha_i + \varepsilon(q(\alpha_j) - q(\alpha_i))$ and $\alpha_j + \varepsilon(q(\alpha_i) - q(\alpha_j))$, respectively. The action function A is intended to operate on the graph $\mathcal{G}$ to update the state of the graph system at each time instant such that only two numbers are updated at each discrete time in terms of the quantized data of each other. Note that this action function is average preserving, i.e., the average of the entries of $\boldsymbol{\alpha}$ is the same as that of $A(\boldsymbol{\alpha}, (i,j))$ for every $\boldsymbol{\alpha} \in \mathbf{R}^n$ and $(i,j) \in \mathcal{E}$. The action function A is employed in the gossip algorithm introduced in the sequel.

Stochastic Gossip (SG) Algorithm:

Step 1: Set $k = 0$ and $\mathbf{X}[0] = \mathbf{X}_0$.

Step 2: Pick an edge $(i,j) \in \mathcal{E}$ of $\mathcal{G}$ at random from the probability distribution $\mathcal{P}$ (i.e., with probability $p_{(i,j)}$). Define $\mathbf{X}[k+1]$ to be $A(\mathbf{X}[k], (i,j))$.

Step 3: Increase k by 1 and jump to step 2.

It is said that *a quantized consensus is reached almost surely (with probability 1) for the graph $\mathcal{G}$ with the initial state* $\mathbf{X}[0] = X_0$ *under the SG algorithm* if almost surely there exist a natural number $\tilde{k}$ and an integer p such that either

$$x_i[k] \in [L_p, L_{p+1}], \quad \forall\, k \geq \tilde{k}, i \in \mathcal{V} \tag{6.2}$$

or

$$x_i[k] \in (\bar{L}_p, \bar{L}_{p+1}], \quad \forall\, k \geq \tilde{k}, i \in \mathcal{V} \tag{6.3}$$

holds. Roughly speaking, a quantized consensus is reached if all numbers on the vertices of the graph $\mathcal{G}$ ultimately lie between two consecutive levels or splitting levels.

As a special case, let $q(x)$ be a uniform quantizer that rounds every number $x \in \mathbf{R}$ to its nearest integer (by convention, assume that $q(p+0.5) = p$ for every integer p). Relax the above-mentioned definition of quantized consensus by replacing the inequalities (6.2) and (6.3) with

$$x_i[k] \in (x_{\text{ave}} - 1, x_{\text{ave}} + 1), \quad \forall\, k \geq \tilde{k}, i \in \mathcal{V}, \tag{6.4}$$

where $x_{\text{ave}} := \frac{x_1+x_2+\cdots+x_n}{n}$. It is shown in [28] that if the quantizer $q(x)$ is uniform, a quantized consensus in the relaxed sense given above is reached almost surely for the graph $\mathcal{G}$ under the SG algorithm, provided $\varepsilon = 0.5$. That paper also conjectures that the same result holds true for every positive number $\varepsilon < 0.5$, while it may not be true for $\varepsilon > 0.5$ (as simulation confirms). The primary objective of the present work is to prove reaching a quantized consensus with probability 1 for every general quantizer $q(x)$ in the form of (6.1) and every fixed parameter $\varepsilon \in (0, 0.5]$. Another goal is to bound the expected value of the convergence time, i.e., the time at which a quantized consensus is reached.

6.3 Convergence Proof

Assume for now that $q(x)$ is a uniform quantizer, as defined above. The results will be later extended to the general case. Consider a tuning factor $\varepsilon \in (0, 0.5]$, and define x_{max} and x_{min} as

$$x_{\text{max}} := \max_{i \in \mathcal{V}} \lceil x_i \rceil, \quad x_{\text{min}} := \min_{i \in \mathcal{V}} \lfloor x_i \rfloor,$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor operators, respectively. If $\mathbf{X}[0]$ is taken as $\mathbf{X}_0$, then two simple observations can be made about $x_i[k]$ ($i \in \mathcal{V}, k \in \mathbf{N}$) as follows:

- The scalar $x_i[k]$ belongs to the interval $[x_{\text{min}}, x_{\text{max}}]$.
- The difference $x_i[k] - x_i$ is an integer multiple of ε (in light of the action function A used in step 2 of the SG algorithm).

These facts imply that if $\mathbf{X}[0] = \mathbf{X}_0$, then the state $\mathbf{X}[k]$, $\forall k \in \mathbf{N}$, belongs to a finite-dimensional set $\mathcal{S}$ that can be defined as the collection of all n-tuple $(\alpha_1, \alpha_2, ..., \alpha_n)$ such that $\alpha_i \in [x_{\min}, x_{\max}]$ and that $\alpha_i - x_i$ is an integer multiple of ε for every $i \in \mathcal{V}$. To present the main results, it is necessary to define two more sets:

$$\begin{aligned} \mathcal{C} &:= \Big\{(\alpha_1, \alpha_2, ..., \alpha_n) \in \mathcal{S} \Big| \; \alpha_i \in (q(x_{\text{ave}}) - 0.5, q(x_{\text{ave}}) + 0.5], \;\; \forall i \in \mathcal{V}\Big\}, \\ \mathcal{C}(\mu) &:= \Big\{(\alpha_1, \alpha_2, ..., \alpha_n) \in \mathcal{S} \Big| \; \alpha_i \in (\mu - \varepsilon, \mu + \varepsilon], \; \forall i \in \mathcal{V}\Big\}, \quad \forall \mu \in \mathbf{R}. \end{aligned} \tag{6.5}$$

Using the definition of quantized consensus provided earlier, one can easily verify that if $\mathbf{X}[k]$ belongs to each of the sets $\mathcal{C}$, $\mathcal{C}(q(x_{\text{ave}}) - 0.5)$ or $\mathcal{C}(q(x_{\text{ave}}) + 0.5)$ for some time instant $k = \tilde{k} \in \mathbf{N}$, then a quantized consensus is reached for the graph $\mathcal{G}$ with the initial state $\mathbf{X}_0$ under the SG algorithm. Hence, to prove reaching a quantized consensus under the SG algorithm with probability 1, it suffices to show that almost surely there exists a time instant $\tilde{k}$ such that $\mathbf{X}[\tilde{k}]$ is contained in one of the sets $\mathcal{C}$, $\mathcal{C}(q(x_{\text{ave}}) - 0.5)$ or $\mathcal{C}(q(x_{\text{ave}}) + 0.5)$. The existence of such a time $\tilde{k}$ with probability 1 is studied in the sequel. Finding appropriate lower and upper bounds on the expected value of $\tilde{k}$ will be addressed in the next section.

A Lyapunov-type argument will be used to prove the convergence to a quantized consensus. For every $\mu \in \mathbf{R}$, let $d(\cdot, \mathcal{C}(\mu)) : \mathcal{S} \to \mathbf{Z}$ be a distance function defined as

$$d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) := \min_{\boldsymbol{\beta} \in \mathcal{C}(\mu)} \frac{\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1}{\varepsilon}, \quad \forall \boldsymbol{\alpha} \in \mathcal{S},$$

where $\|\cdot\|_1$ denotes the L_1 norm. Note that the number $d(\boldsymbol{\alpha}, \mathcal{C}(\mu))$ quantifies the distance of the vector $\boldsymbol{\alpha} \in \mathcal{S}$ from the discrete set $\mathcal{C}(\mu)$. A Lyapunov function will be later introduced in terms of $d(\cdot, \mathcal{C}(q(x_{\text{ave}}) - 0.5))$ and $d(\cdot, \mathcal{C}(q(x_{\text{ave}}) + 0.5))$ to prove the convergence. Define $\mathbf{Q}$ to be the set $\{k + 0.5 | \; k \in \mathbf{Z}\}$.

Lemma 1 *Given $\mu \in \mathbf{Q}$, the inequality*

$$d(A(\boldsymbol{\alpha}, (i,j)), \mathcal{C}(\mu)) \leq d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) \tag{6.6}$$

holds for every vector $\boldsymbol{\alpha} \in \mathcal{S}$ and edge $(i,j) \in \mathcal{E}$.

Proof: Let the short-hand notation $\boldsymbol{\beta}$ be used for $d(A(\boldsymbol{\alpha}, (i,j))$. Denote the p^{th} entries of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with α_p and β_p, respectively, for every $p \in \mathcal{V}$. To simplify the proof, suppose that neither $\alpha_i - \mu$ nor $\alpha_j - \mu$ is an integer multiple of ε. If $\boldsymbol{\alpha}$ belongs to $\mathcal{C}(\mu)$, then

$$d(\boldsymbol{\beta}, \mathcal{C}(\mu)) = d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) = 0.$$

Hence, with no loss of generality we assume that $\boldsymbol{\alpha} \notin \mathcal{C}(\mu)$ and $\alpha_j \leq \alpha_i$. If $\alpha_i, \alpha_j > \mu$ or $\alpha_i, \alpha_j < \mu$, then

$$d(\boldsymbol{\beta}, \mathcal{C}(\mu)) = d(\boldsymbol{\alpha}, \mathcal{C}(\mu)).$$

For the remaining case $\alpha_i > \mu$ and $\alpha_j < \mu$, it suffices to prove the inequality

$$d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) - d(\boldsymbol{\beta}, \mathcal{C}(\mu)) \geq \min\left\{2\big(q(\alpha_i) - q(\alpha_j)\big), 2\left\lfloor \frac{\alpha_i - \mu}{\varepsilon} \right\rfloor + 1, 2\left\lfloor \frac{\mu - \alpha_j}{\varepsilon} \right\rfloor + 1\right\}. \tag{6.7}$$

To this end, a number of possibilities can be considered as follows:

i) *$\beta_i > \mu$ and $\beta_j < \mu$:* Given a vector $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_n \end{bmatrix} \in \mathcal{S}$, if none of the numbers $\gamma_1 - \mu, \gamma_2 - \mu, ..., \gamma_n - \mu$ is an integer multiple of ε, then it can be shown that

$$d(\boldsymbol{\gamma}, \mathcal{C}(\mu)) = \sum_{i=1}^{n} \left\lfloor \frac{|\gamma_i - \mu|}{\varepsilon} \right\rfloor. \tag{6.8}$$

Use the above equality twice for $\boldsymbol{\gamma} = \boldsymbol{\alpha}$ and $\boldsymbol{\gamma} = \boldsymbol{\beta}$ to obtain

$$\begin{aligned} d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) - d(\boldsymbol{\beta}, \mathcal{C}(\mu)) &= \left\lfloor \frac{\alpha_i - \mu}{\varepsilon} \right\rfloor + \left\lfloor \frac{\mu - \alpha_j}{\varepsilon} \right\rfloor \\ &- \left\lfloor \frac{\alpha_i + \varepsilon(q(\alpha_j) - q(\alpha_i)) - \mu}{\varepsilon} \right\rfloor - \left\lfloor \frac{\mu - \alpha_j - \varepsilon(q(\alpha_i) - q(\alpha_j))}{\varepsilon} \right\rfloor \\ &= 2\big(q(\alpha_i) - q(\alpha_j)\big). \end{aligned} \tag{6.9}$$

ii) *$\beta_i > \mu$ and $\beta_j > \mu$:* The equality (6.8) yields

$$\begin{aligned} d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) - d(\boldsymbol{\beta}, \mathcal{C}(\mu)) &= \left\lfloor \frac{\alpha_i - \mu}{\varepsilon} \right\rfloor + \left\lfloor \frac{\mu - \alpha_j}{\varepsilon} \right\rfloor \\ &- \left\lfloor \frac{\alpha_i + \varepsilon(q(\alpha_j) - q(\alpha_i)) - \mu}{\varepsilon} \right\rfloor - \left\lfloor \frac{\alpha_j + \varepsilon(q(\alpha_i) - q(\alpha_j)) - \mu}{\varepsilon} \right\rfloor \\ &= \left\lfloor \frac{\mu - \alpha_j}{\varepsilon} \right\rfloor - \left\lfloor \frac{\alpha_j - \mu}{\varepsilon} \right\rfloor = 2\left\lfloor \frac{\mu - \alpha_j}{\varepsilon} \right\rfloor + 1 \end{aligned} \tag{6.10}$$

(the assumption that $\alpha_j - \mu$ is not an integer multiple of ε is used to derive the last line of the above inequality).

iii) $\beta_i < \mu$ *and* $\beta_j < \mu$*:* Similar to the previous case, one can write

$$d(\boldsymbol{\alpha}, \mathcal{C}(\mu)) - d(\boldsymbol{\beta}, \mathcal{C}(\mu)) = 2\left\lfloor \frac{\alpha_i - \mu}{\varepsilon} \right\rfloor + 1. \tag{6.11}$$

iv) $\beta_i < \mu$ *and* $\beta_j > \mu$*:* This case is possible only if $\boldsymbol{\alpha} \in \mathcal{C}(\mu)$, which is in contradiction to the assumption made earlier.

The proof is completed by noting that the inequality (6.7) follows immediately from (6.9), (6.10), and (6.11). ■

Remark 1 *Suppose we run the SG algorithm on the graph $\mathcal{G}$ with the initial state $\mathbf{X}[0] = \mathbf{X}_0$ to obtain an infinite sequence of states $\mathbf{X}[0], \mathbf{X}[1], \mathbf{X}[2], \ldots.$ Given $\mu \in \mathbf{Q}$, it follows from Lemma 1 that*

$$d(\mathbf{X}[0], \mathcal{C}(\mu)) \geq d(\mathbf{X}[1], \mathcal{C}(\mu)) \geq d(\mathbf{X}[2], \mathcal{C}(\mu)) \geq \cdots$$

In other words, as time elapses, the state of the graph system can never become farther from the discrete set $\mathcal{C}(\mu)$.

The single-action function A was already defined. For a vector $\boldsymbol{\alpha} \in \mathcal{S}$ and a sequence of edges $(i_1, j_1), (i_2, j_2), (i_3, j_3), \ldots$, the multi-action functions $A^2, A^2, A^4, \ldots$ can also be defined analogously as

$$\begin{aligned} A^2(\boldsymbol{\alpha}, (i_1, j_1), (i_2, j_2)) &:= A(A(\boldsymbol{\alpha}, (i_1, j_1)), (i_2, j_2)), \\ A^3(\boldsymbol{\alpha}, (i_1, j_1), (i_2, j_2), (i_3, j_3)) &:= A(A^2(\boldsymbol{\alpha}, (i_1, j_1), (i_2, j_2)), (i_3, j_3)), \\ &\vdots \end{aligned} \tag{6.12}$$

Let r denote the cardinality of the set $\mathcal{E}$. For every $\mu \in \mathbf{Q}$, define a deterministic gossip algorithm as follows for an initial state $\mathbf{X}[0] \in \mathcal{S}$.

μ-Deterministic Gossip (μ-DG) Algorithm:

Step 1: Set $k = 0$.

Step 2: Find a sequence of r edges $(i_1, j_1), (i_2, j_2), \ldots., (i_r, j_r) \in \mathcal{E}$ (not necessarily distinct) such that $d(A^r(\mathbf{X}[k], (i_1, j_1), ..., (i_r, j_r)), \mathcal{C}(\mu))$ has the least possible value. Define $\mathbf{X}[k+1]$ to be $A(\mathbf{X}[k], (i_1, j_1))$.

Step 3: Increase k by 1 and jump to step 2.

Note that although the SG algorithm picks an edge at random from the probability distribution $\mathcal{P}$ at each time update, an optimal edge is selected by the μ-DG algorithm at each iteration in such a way that if the next $r-1$ time updates were taken in an optimal way, then the resulting state of the graph system would be as closely as possible to the set $\mathcal{C}(\mu)$. As a result, this deterministic algorithm takes an optimal strategy relative to the set $\mathcal{C}(\mu)$.

Lemma 2 *Given $\mu \in \mathbf{Q}$, apply the μ-DG algorithm to the graph $\mathcal{G}$ with an initial state $\mathbf{X}[0] \in \mathcal{S}$. There exists a natural number k_0 for which either of the following cases occurs:*

i) $\mathbf{X}[k]$ *belongs to set $C(\mu)$, for every $k \geq k_0$.*

ii) $x_1[k]-\mu, x_2[k]-\mu, ..., x_n[k]-\mu$ *are either all negative or all strictly positive, for every $k \geq k_0$.*

Proof: The proof is provided in Appendix 1. ■

For notational simplicity, let η_1 and η_2 denote $q(x_{\text{ave}})-0.5$ and $q(x_{\text{ave}})+0.5$, respectively. The next theorem presents a key result that will be later used to prove the almost sure convergence to a quantized consensus under the SG algorithm.

Theorem 1 *Apply the η_1-DG algorithm to the graph $\mathcal{G}$ with an initial state $\mathbf{x}[0] \in \mathcal{S}$. Stop the algorithm at some iteration $k = k_0$ where the integer-valued non-increasing (nonnegative) function $d(\mathbf{X}[k], \mathcal{C}(\eta_1))$ attains its minimum. Then, run the η_2-DG algorithm on the graph $\mathcal{G}$ (by starting from the current state $\mathbf{X}[k_0]$) until the function $d(\mathbf{X}[k], \mathcal{C}(\eta_2))$ reaches its minimum at some iteration $k = k_1$. One of the following cases takes place:*

i) $\mathbf{X}[k]$ *belongs to the set $\mathcal{C}$, for every $k \geq k_1$.*

ii) $\mathbf{X}[k]$ *belongs to the set $\mathcal{C}(\eta_1)$, for every $k \geq k_1$.*

iii) $\mathbf{X}[k]$ *belongs to the set $\mathcal{C}(\eta_2)$, for every $k \geq k_1$.*

Proof: Since the η_1-DG algorithm makes $d(\mathbf{X}[k], \mathcal{C}(\eta_1))$ reach its minimum at $k = k_0$, it can be concluded from (the proof of) Lemma 2 that one of the following cases happens:

i) $\mathbf{X}[k]$ *belongs to the set* $\mathcal{C}(\eta_1)$ *for every* $k \geq k_0$*:* This corresponds to case (ii) of the theorem.

ii) $x_1[k] - \eta_1, x_2[k] - \eta_1, ..., x_n[k] - \eta_1$ *are all negative for every* $k \geq k_0$*:* One can write

$$\eta_1 \geq \frac{1}{n}\sum_{i=1}^{n} x_i[k] = x_{ave} > q(x_{ave}) - 0.5 = \eta_1.$$

The above contradiction does not allow this case to take place.

iii) $x_1[k] - \eta_1, x_2[k] - \eta_1, ..., x_n[k] - \eta_1$ *are all strictly positive for every* $k \geq k_0$*:* Using Lemma 2 for the η_2-DG algorithm, it can be concluded that one of the following cases occurs:

- $\mathbf{x}[k]$ *belongs to the set* $\mathcal{C}(\eta_2)$ *for every* $k \geq k_1$*:* If this is the case, the proof is complete.

- $x_1[k] - \eta_2, x_2[k] - \eta_2, ..., x_n[k] - \eta_2$ *are all negative for every* $k \geq k_1$*:* This case simply implies that $\mathbf{X}[k]$ belongs to the set $\mathcal{C}$, which corresponds to case (i) of the theorem.

- $x_1[k] - \eta_2, x_2[k] - \eta_2, ..., x_n[k] - \eta_2$ *are all strictly positive for every* $k \geq k_1$*:* The inequality

$$\eta_2 < \frac{1}{n}\sum_{i=1}^{n} x_i[k] = x_{ave} \leq q(x_{ave}) + 0.5 = \eta_2$$

is a contradiction for this case, as before. ■

The next theorem presents the main result of this section.

Theorem 2 *Apply the SG algorithm to the graph* $\mathcal{G}$ *with the initial state* $\mathbf{x}[0] = \mathbf{X}_0$*. With probability 1, there exists a natural number* $\tilde{k}$ *such that one of the following cases occurs:*

i) $\mathbf{x}[k]$ *belongs to the set* $\mathcal{C}$*, for all* $k \geq \tilde{k}$*.*

ii) $\mathbf{x}[k]$ *belongs to the set* $\mathcal{C}(q(x_{ave}) - 0.5)$*, for all* $k \geq \tilde{k}$*.*

iii) $\mathbf{x}[k]$ *belongs to the set* $\mathcal{C}(q(x_{ave}) + 0.5)$*, for all* $k \geq \tilde{k}$*.*

Proof: Construct a transition graph $\tilde{\mathcal{G}}$ as follows:

- Put $|\mathcal{S}|$ vertices corresponding to the elements of the set $\mathcal{S}$ (where $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$).

- For every $\boldsymbol{\alpha} \in \mathcal{S}$ and $(i,j) \in \mathcal{E}$ such that $i < j$, draw a directed edge from vertex $\boldsymbol{\alpha}$ to vertex $A(\boldsymbol{\alpha}, (i,j))$ in $\tilde{\mathcal{G}}$, and assign the weight $p_{(i,j)}$ to this edge.

It is easy to verify that every run of the SG algorithm on the graph $\mathcal{G}$ with the initial state $\mathbf{X}[0] = \mathbf{X}_0$ is equivalent to a random walk on the graph $\tilde{\mathcal{G}}$ starting from vertex $\mathbf{X}_0$, where the weight of every edge shows its probability of being chosen through the walk. Let $\tilde{\mathcal{G}}_0$ denote an induced subgraph of $\tilde{\mathcal{G}}$ with the set of vertices $\mathcal{C} \cup \mathcal{C}(q(x_{\text{ave}}) - 0.5) \cup \mathcal{C}(q(x_{\text{ave}}) + 0.5)$. The problem now reduces to proving that every random walk in $\tilde{\mathcal{G}}$ starting from vertex $\mathbf{X}_0$ almost surely ends in the subgraph $\tilde{\mathcal{G}}_0$. To prove this statement, two observations are needed. First, note that if a walk enters the subgraph $\tilde{\mathcal{G}}_0$, it can never leave this subgraph (due to its vertices forming a quantized set, as discussed earlier). Second, it can be deduced from Theorem 1 that there is a directed path from every vertex of $\tilde{\mathcal{G}}$ to (some vertex of) the subgraph $\tilde{\mathcal{G}}_0$ (this path can, for instance, be obtained using the η_1-DG and η_2-DG algorithms). These properties imply that the subgraph $\tilde{\mathcal{G}}_0$ is an absorbing set and, therefore, it follows from a well-known theorem in the Markov chain theory that every infinite random walk almost surely ends up in this absorbing set [65]. This completes the proof. ■

Define the diameter of a discrete set $\mathcal{M}$ as the supremum of the infinity norm of the difference between every two points in $\mathcal{M}$, i.e.,

$$\sup_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}} \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_\infty$$

where $\|\cdot\|_\infty$ denotes the L_∞ norm. Moreover, for every natural number $p \in \mathbf{N}$, define $\mathcal{M}^p$ to be the product set $\underbrace{\mathcal{M} \times \mathcal{M} \times \cdots \times \mathcal{M}}_{p \text{ times}}$.

Remark 2 *The relaxed definition of quantized consensus provided in [28] for a uniform quantizer is less precise than the one introduced in this work. Indeed, the work [28] states that if a quantized consensus is reached at time $\tilde{k}$, then the state $\mathbf{X}[k]$ belongs to the box $(x_{ave} - 1, x_{ave} + 1)^n$ for every $k \geq \tilde{k}$. In contrast, Theorem 2 proves that there exists a* *positive integer $\tilde{k}$ such that $\mathbf{X}[k]$, $\forall k \geq \tilde{k}$, belongs to one of the sets $\mathcal{C}$, $\mathcal{C}(\eta_1)$ or $\mathcal{C}(\eta_2)$. In this regard, two points can be made as follows:*

- *The diameter of the set $(x_{ave} - 1, x_{ave} + 1)^n$ is equal to 2, whereas the diameter of each of the sets $\mathcal{C}$, $\mathcal{C}(\eta_1)$, or $\mathcal{C}(\eta_2)$ is at most 1.*
- *If $\mathbf{X}[k]$ in the steady state (for large enough k's) is not constant and oscillates (with probability 1), it should then belong to either $\mathcal{C}(\eta_1)$ or $\mathcal{C}(\eta_2)$, which are both of diameter ε. Since the diameters of these sets can become arbitrarily small by rendering an appropriate ε, running the SG algorithm for a small ε either makes the steady state constant or permits it to oscillate in a set with a small diameter (ε). In the latter case, each number $x_i[k]$ can oscillate between only two numbers of difference ε (due to the definition of $\mathcal{C}(\mu)$, $\mu \in \mathbf{R}$).*

To clarify Remark 2, consider the nominal values $x_{\text{ave}} = 10.6$ and $\varepsilon = 0.2$. The definition of consensus borrowed from [28] states that there exists a positive integer $\tilde{k}$ such that

$$9.6 < x_1[k], ..., x_n[k] < 11.6, \quad \forall k \geq \tilde{k}.$$

In contrast, Theorem 2 asserts that there exists a number $\tilde{k}$ so that

$$10.3 < x_1[k], ..., x_n[k] \leq 10.7, \quad \forall k \geq \tilde{k} \tag{6.13}$$

or

$$10.5 < x_1[k], ..., x_n[k] \leq 11.5, \quad \forall k \geq \tilde{k} \tag{6.14}$$

(note that case (iii) in Theorem 2 is ruled out in this example, as the average of the entries of $\mathbf{X}[k]$ cannot be smaller than all entries of $\mathbf{X}[k]$). Comparing (6.13) with (6.13) and (6.14), one can simply observe that a more informative description of the steady-state values on the vertices of $\mathcal{G}$ is delineated by (6.13) and (6.14).

6.3.1 Generalization to Arbitrary Quantizers

To prove reaching a quantized consensus with probability 1 under the SG algorithm for a general quantizer $q(x)$ given by (6.1), the definitions of η_1, η_2, $\mathcal{C}$, and $\mathcal{C}(\mu)$ (where $\mu \in \mathbf{R}$) should be revised. This is carried out in the sequel.

Let η_1 and η_2 be equal to

$$\eta_1 = \max\left\{\bar{L}_i \middle|\ i \in \mathbf{Z},\ \bar{L}_i \leq x_{\text{ave}}\right\},$$
$$\eta_2 = \min\left\{\bar{L}_j \middle|\ j \in \mathbf{Z}, \bar{L}_j \geq x_{\text{ave}}\right\}.$$

As before, define $\mathcal{S}$ to be the set of all n-tuple $(\alpha_1, \alpha_2, ..., \alpha_n)$ such that $\alpha_i \in [x_{\min}, x_{\max}]$ and that $\alpha_i - x_i$ is an integer multiple of ε for every $i \in \mathcal{V}$. Moreover

$$\mathcal{C} := \left\{(\alpha_1, \alpha_2, ..., \alpha_n) \in \mathcal{S} \middle|\ \alpha_i \in (\eta_1, \eta_2],\quad \forall i \in \mathcal{V}\right\},$$
$$\mathcal{C}(\bar{L}_j) := \left\{(\alpha_1, \alpha_2, ..., \alpha_n) \in \mathcal{S} \middle|\ \alpha_i \in \left(\bar{L}_j - \varepsilon(L_{j+1} - L_j), \bar{L}_j + \varepsilon(L_{j+1} - L_j)\right],\ \forall i \in \mathcal{V}\right\},\ \forall j \in \mathbf{Z}.$$

One can adopt an approach similar to the one proposed earlier to prove all lemmas and theorems given so far for a general quantizer $q(x)$. This leads to the conclusion that a quantized consensus is reached almost surely for the graph $\mathcal{G}$ under the SG algorithm with the initial state $\mathbf{X}[0] = \mathbf{X}_0$ and, more specifically, $\mathbf{X}[k]$ belongs to one of the quantized sets $\mathcal{C}$, $\mathcal{C}(\eta_1)$ or $\mathcal{C}(\eta_2)$ for large enough k's.

6.4 Convergence Time

Let $\boldsymbol{\mathcal{E}}\{\cdot\}$ and $\boldsymbol{\mathcal{E}}\{\cdot|\cdot\}$ denote the *expectation* and *conditional expectation* operators, respectively. For simplicity, assume that $q(x)$ is a uniform quantizer (the results can be extended to the general case similarly to what was done earlier). Suppose that $\mathbf{X}_0$ is an unknown initial state that belongs to the given hyperrectangle $[x_{\min}, x_{\max}]^n$. Since the time $\tilde{k}$ at which the state of the graph system belongs to one of the quantized sets $\mathcal{C}$, $\mathcal{C}(\eta_1)$, or $\mathcal{C}(\eta_2)$ is a random variable by virtue of the stochastic nature of the SG algorithm, the goal is to find the expected value of $\tilde{k}$ corresponding to the worst initial state $\mathbf{X}_0$ in the hyperrectangle $[x_{\min}, x_{\max}]^n$. In other words, the objective is to study the quantity t_c, where

$$t_c := \max\left\{\boldsymbol{\mathcal{E}}\{\tilde{k}|\mathbf{X}[0] = \mathbf{X}_0\} \ \middle|\ \mathbf{X}_0 \in [x_{\min}, x_{\max}]^n\right\}. \tag{6.15}$$

Recall that the sets $\mathcal{C}$ and $\mathcal{C}(\mu)$ (where $\mu \in \mathbf{Q}$) defined earlier both depend on the initial state X_0. To show this dependency explicitly (as required later in this work), let $\mathcal{C}_{\boldsymbol{\alpha}}$ and $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$ be defined similarly to $\mathcal{C}$ and $\mathcal{C}(\mu)$, respectively, but for an arbitrary initial state $\mathbf{X}[0] = \boldsymbol{\alpha} \in [x_{\min}, x_{\max}]^n$.

Given a number $\mu \in \mathbf{Q}$, consider a run of the SG algorithm on the graph $\mathcal{G}$ with an initial state $\mathbf{X}[0] = \boldsymbol{\alpha} \in [x_{\min}, x_{\max}]^n$, where an edge $(i,j) \in \mathcal{E}$ is chosen at the k^{th} update ($k \in \mathbf{N}$). If $\mathbf{X}[k]$ defined as $A(\mathbf{X}[k-1], (i,j))$ has the property that

$$d(\mathbf{X}[k], \mathcal{C}_{\boldsymbol{\alpha}}(\mu)) \leq d(\mathbf{X}[k-1], \mathcal{C}_{\boldsymbol{\alpha}}(\mu)) - 1,$$

then the action taken at the k^{th} update is said to be a *positive action* with respect to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$, otherwise it is called a *trivial action* meaning that

$$d(\mathbf{X}[k], \mathcal{C}_{\boldsymbol{\alpha}}(\mu)) = d(\mathbf{X}[k-1], \mathcal{C}_{\boldsymbol{\alpha}}(\mu))$$

(see Remark 1). It can be deduced from the proof of Lemma 1 (particularly the inequality (6.7)) that a positive action takes place at the k^{th} update with respect to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$ if and only if none of the following sets of relations holds:

$$x_i[k-1], x_j[k-1] \leq \mu, \tag{6.16a}$$

$$x_i[k-1], x_j[k-1] > \mu, \tag{6.16b}$$

$$x_i[k-1], x_j[k-1] \in (\mu - \varepsilon, \mu + \varepsilon]. \tag{6.16c}$$

This motivates the introduction of a set $\tilde{\mathcal{C}}(\mu)$ defined as the collection of all vectors $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix} \in [x_{\min}, x_{\max}]^n$ for which there are two indices $i, j \in \mathcal{V}$ such that none of the sets of relations

$$\alpha_i, \alpha_j \leq \mu, \tag{6.17a}$$

$$\alpha_i, \alpha_j > \mu, \tag{6.17b}$$

$$\alpha_i, \alpha_j \in (\mu - \varepsilon, \mu + \varepsilon] \tag{6.17c}$$

holds. Given $\mu \in \boldsymbol{Q}$, let $T(\mu)$ denote the first time instant at which a positive action occurs with respect to the set $\mathcal{C}_{\mathbf{X}[0]}(\mu)$ under running the SG algorithm on the graph $\mathcal{G}$. Note that $T(\mu)$ is a random variable that depends on the initial state $\mathbf{X}[0]$ and the probability distribution $\mathcal{P}$. Define now

$$\Phi(\mu) := \max\left\{ \boldsymbol{\mathcal{E}}\{T(\mu) \big| \mathbf{X}[0] = \boldsymbol{\alpha}\} \;\Big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu) \right\}, \quad \forall \mu \in \mathbf{Q}. \tag{6.18}$$

Roughly speaking, $\Phi(\mu)$ characterizes the maximum of the expected number of time updates that are required for the SG algorithm to take a possible action on the graph $\mathcal{G}$ with respect to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$ by starting from every state $\boldsymbol{\alpha}$ for which a positive action can be taken in the future (this is guaranteed by the condition $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$ in the above definition). The quantity $\Phi(\mu)$ and its relevance to the desired term t_c are investigated in the sequel.

Theorem 3 *Given $\mu \in \mathbf{Q}$, the number $\Phi(\mu)$ is equal to*

$$\max \Big\{ \boldsymbol{\mathcal{E}}\big\{ T(\mu) \big| \mathbf{X}[0] = \boldsymbol{\alpha} \big\} \Big\},$$

where the maximum is taken over all n-tuple $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix}$ satisfying the relation $\{\alpha_1, \alpha_2, ..., \alpha_n\} = \{\mu - \varepsilon, \mu, ..., \mu, \mu + \varepsilon\}$ in which the value μ appears $n-2$ times.

Proof: The proof is provided in Appendix 2. ■

The following definition and notation will be later used to express $\Phi(\mu)$ in terms of the topology of the graph and the probability set $\mathcal{P}$:

- Let P denote the Laplacian of the weighted graph $\mathcal{G}$, i.e.,

$$p_{ij} = \begin{cases} -p_{(i,j)} & \text{if } (i,j) \in \mathcal{E} \\ \sum_{(i,u)\in\mathcal{E}} p_{(i,u)} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

  where p_{ij} represents the (i,j) entry of the matrix P for every $i, j \in \mathcal{V}$.

- For every $i \in \mathcal{V}$ and $M \in \mathbf{R}^{n\times n}$, define $M_{\sim i}$ to be a matrix obtained from M by removing its i^{th} row and i^{th} column.

Theorem 4 *Given $\mu \in \mathbf{Q}$, the quantity $\Phi(\mu)$ can be obtained as*

$$\Phi(\mu) = \max_{j\in\mathcal{V}} \left\| (P_{\sim j})^{-1} E \right\|_{\infty},$$

where $E \in \mathbf{R}^{n-1}$ is a vector of 1's.

Proof: For every $i, j \in \mathcal{V},\ i \neq j$, let $\boldsymbol{\beta}_{ij}$ denote an n-dimensional vector whose elements are all equal to μ, except for its i^{th} and j^{th} entries that are $\mu + \varepsilon$ and $\mu - \varepsilon$, respectively. It follows from Theorem 3 that

$$\Phi(\mu) = \max_{i,j\in\mathcal{V},\ i\neq j} \boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\big\}. \tag{6.19}$$

It is useful to contrive a recursive equation for $\boldsymbol{\mathcal{E}}\{T(\mu)|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\}$. To this end, run the SG algorithm on the graph $\mathcal{G}$ with the initial state $\boldsymbol{\beta}_{ij}$. The expected value of the time at which the first positive action is taken with respect to the set $\mathcal{C}_{\boldsymbol{\beta}_{ij}}(\mu)$ is $\boldsymbol{\mathcal{E}}\{T(\mu)|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\}$. To count this number in another way, run the algorithm only one iteration. Assume that the edge $e \in \mathcal{E}$ is chosen in the first update. There are a number of possibilities as given below:

- *e is equal to the edge (i,u), for some $u \in \mathcal{V}\backslash\{j\}$:* In this case, due to the equality $\mathbf{X}[0] = \boldsymbol{\beta}_{ij}$, the vector $\mathbf{X}[1]$ can be obtained as $\boldsymbol{\beta}_{uj}$. Hence, it is expected to take the first positive action after $\boldsymbol{\mathcal{E}}\{T(\mu)|\mathbf{X}[0] = \boldsymbol{\beta}_{uj}\}$ time updates (in addition to the first time update taken at the beginning).

- *e is equal to the edge (i,j):* This means that a positive action is already taken at the first time update.

- *e is equal to the edge (u,l), for some $u,l \in \mathcal{V}\backslash\{i\}$:* In this case, it is easy to show that $\mathbf{X}[1] = \mathbf{X}[0] = \boldsymbol{\beta}_{ij}$. This implies that it is expected to take the first positive action after $\boldsymbol{\mathcal{E}}\{T(\mu)|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\}$ time updates (other than the first one already taken).

The above reasoning yields the recursive equation

$$\begin{aligned}\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\big\} =&1 + \sum_{(i,u)\in\mathcal{E}} p_{(i,u)}\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{uj}\big\}\\ &+ \Big(1 - \sum_{(i,u)\in\mathcal{E}} p_{(i,u)}\Big)\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\big\}, \quad \forall i \in \mathcal{V}\backslash\{j\}.\end{aligned} \tag{6.20}$$

This equation can be arranged in a matrix form to obtain

$$\max_{i\in\mathcal{V}\backslash\{j\}} \boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{ij}\big\} = \big\|(P_{\sim j})^{-1}E\big\|_{\infty}, \quad \forall j \in \mathcal{V}. \tag{6.21}$$

The proof follows immediately from (6.19) and (6.21). ■

An important implication of Theorem 4 is that the quantity $\Phi(\mu)$ does not depend on μ or ε. Hence, the notation Φ will be used henceforth for $\Phi(\mu)$. Given $\boldsymbol{\alpha} = \left[\begin{array}{cccc} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{array}\right]$, define $V(\boldsymbol{\alpha})$ as

$$V(\boldsymbol{\alpha}) := \sum_{i=1}^{n} (\alpha_i - x_{\text{ave}})^2.$$

Apply the SG algorithm to the graph $\mathcal{G}$ with the initial state $\mathbf{X}[0] = \mathbf{X}_0 \in [x_{\min}, x_{\max}]^n$. In line with the method developed in [28], it is easy to show that

$$\boldsymbol{\mathcal{E}}\big\{V(\mathbf{X}[k])\big\} \leq \big(1 - \varepsilon\lambda_2(P)\big)^k V(\mathbf{X}_0) + \frac{\varepsilon}{\lambda_2(P)}, \quad \forall k \in \mathbf{N}, \tag{6.22}$$

where $\lambda_2(P)$ denotes the second smallest eigenvalue of the positive semidefinite matrix P. As discussed in [28] for the special case $\varepsilon = 0.5$, the right side of the above inequality is composed of two terms that can be interpreted as follows:

- The first term is $\big(1-\varepsilon\lambda_2(P)\big)^k V(\mathbf{X}_0)$, which depends on the initial state. Since $\lambda_2(p)$ is always less than 2 (if $n \geq 3$) and $\varepsilon \in (0, 0.5]$, the number $1 - \varepsilon\lambda_2(P)$ belongs to the interval $(0, 1)$. This means that the term $\big(1 - \varepsilon\lambda_2(P)\big)^k V(\mathbf{X}_0)$ goes to zero exponentially fast.

- The second term is $\frac{\varepsilon}{\lambda_2(P)}$, which does not depend on the initial state.

The inequality (6.22) implies that the effect of the initial state $\mathbf{X}_0$ disappears in the quantity $\boldsymbol{\mathcal{E}}\big\{V(\mathbf{X}[k])\big\}$ exponentially fast so that the state of the graph system becomes close to a quantized set at an exponential rate and then the convergence rate begins to slow down until a quantized consensus is reached. The bias term $\frac{\varepsilon}{\lambda_2(P)}$ in (6.22) makes it impossible to find t_c directly. However, the parameter $\Phi(\mu)$ introduced earlier will be used in the sequel to bypass this issue.

Theorem 5 *The number t_c can be lower and upper bounded as*

$$t_c \geq \max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty, \tag{6.23a}$$

$$t_c \leq \left\lceil -\frac{\log\big(n(x_{\max} - x_{\min})^2\big)}{\log(1 - \varepsilon\lambda_2(P))} \right\rceil + \frac{2}{\varepsilon}\left(n + \sqrt{n\left(1 + \frac{\varepsilon}{\lambda_2(P)}\right)}\right) \max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty. \tag{6.23b}$$

Proof: It can be concluded from (6.15) and (6.18) that Φ is a lower bound for t_c. The inequality (6.23a) follows from this fact and Theorem 4 that relates Φ to the Laplacian matrix

P. To prove the inequality (6.23b), consider an arbitrary initial state $\mathbf{X}[0] \in [x_{\min}, x_{\max}]^n$. Define $\bar{k}$ as

$$\bar{k} := \left\lceil -\frac{\log\left(n(x_{\max} - x_{\min})^2\right)}{\log(1 - \varepsilon\lambda_2(P))} \right\rceil. \tag{6.24}$$

Notice that

$$\left(1 - \varepsilon\lambda_2(P)\right)^{\bar{k}} V(\mathbf{X}_0) \leq \left(1 - \varepsilon\lambda_2(P)\right)^{\bar{k}} n(x_{\max} - x_{\min})^2 \leq 1.$$

This inequality implies that the effect of the initial state on $\boldsymbol{\mathcal{E}}\{V(\mathbf{X}[k])\}$ mainly diminishes by the time $k = \bar{k}$; more specifically, the relation (6.22) yields

$$\boldsymbol{\mathcal{E}}\{V(\mathbf{X}[\bar{k}])\} \leq 1 + \frac{\varepsilon}{\lambda_2(P)}. \tag{6.25}$$

On the other hand, one can write

$$\begin{aligned} d(\mathbf{X}[\bar{k}], \mathcal{C}(\eta_1)) + d(\mathbf{X}[\bar{k}], \mathcal{C}(\eta_2)) &\leq \sum_{i=1}^{n} \frac{|x_i[\bar{k}] - \eta_1| + |x_i[\bar{k}] - \eta_2|}{\varepsilon} \\ &\leq 2\sum_{i=1}^{n} \frac{|x_i[\bar{k}] - x_{\text{ave}}| + 1}{\varepsilon} \\ &\leq \frac{2}{\varepsilon}\left(n + \sqrt{nV(\mathbf{X}[\bar{k}])}\right). \end{aligned} \tag{6.26}$$

It follows from (6.25), (6.26) and the concavity of the function $\sqrt{x}$ that

$$\begin{aligned} \boldsymbol{\mathcal{E}}\left\{d(\mathbf{X}[\bar{k}], \mathcal{C}(\eta_1)) + d(\mathbf{X}[\bar{k}], \mathcal{C}(\eta_2))\right\} &\leq \frac{2}{\varepsilon}\boldsymbol{\mathcal{E}}\left\{n + \sqrt{nV(\mathbf{X}[\bar{k}])}\right\} \\ &\leq \frac{2}{\varepsilon}\left(n + \sqrt{n\boldsymbol{\mathcal{E}}\{V(\mathbf{X}[\bar{k}])\}}\right) \\ &\leq \frac{2}{\varepsilon}\left(n + \sqrt{n\left(1 + \frac{\varepsilon}{\lambda_2(P)}\right)}\right). \end{aligned} \tag{6.27}$$

Consider the state $\mathbf{X}[k]$ at the time instant $k = \bar{k}$. It can be inferred from Theorem 1 that if sufficient positive actions are taken by the SG algorithm by starting from the state $\mathbf{X}[\bar{k}]$ so that $d(\mathbf{X}[k], \mathcal{C}(\eta_1))$ reaches its minimum at some time instant $k = \tilde{k}_0$ and then more positive actions are taken to make $d(\mathbf{X}[k], \mathcal{C}(\eta_2))$ attain its minimum at some time $k = \tilde{k}$, then a quantized consensus is reached at the time instant $\tilde{k}$. This property together with the fact that at most Φ iterations (in expectation) are required to take a positive action leads to

$$t_c \leq \bar{k} + \boldsymbol{\mathcal{E}}\left\{d(\mathbf{X}[\bar{k}], \mathcal{C}(\eta_1)) + d(\mathbf{X}[\bar{k}], \mathcal{C}(\eta_2))\right\}\Phi. \tag{6.28}$$

The inequality (6.23b) follows immediately from (6.24), (6.27), (6.28), and Theorem 4. ■

Note that the upper bound on t_c provided in Theorem 5 is composed of two terms. The first one depends on $x_{\max}$ and $x_{\min}$, and is indeed identical to the available upper bound on t_c in the case when the standard unquantized gossip algorithm is applied to the graph. The second term in the upper bound does not depend on the geometry of the initial state and is only contingent upon the topology of the graph. This term is due to the quantized nature of the SG algorithm and corresponds to the number of iterations required to reach a quantized consensus once the state of the graph system is already close to a quantized set. The next theorem relates the upper bound on t_c to the spectral of the principal submatrices of the Laplacian P.

Theorem 6 *The scalar t_c satisfies the inequality*

$$t_c \leq \left\lceil -\frac{\log\left(n(x_{\max} - x_{\min})^2\right)}{\log(1 - \varepsilon\lambda_2(P))}\right\rceil + \frac{2\sqrt{n-1}}{\varepsilon}\left(n + \sqrt{n\left(1 + \frac{\varepsilon}{\lambda_2(P)}\right)}\right)\left(\max_{j\in\mathcal{V}} \frac{1}{\lambda_{min}\{P_{\sim j}\}}\right), \tag{6.29}$$

where $\lambda_{\min}(\cdot)$ represents the smallest eigenvalue of a matrix.

Proof: Given $j \in \mathcal{V}$, One can write

$$\|(P_{\sim j})^{-1}E\|_\infty \leq \|(P_{\sim j})^{-1}\|_\infty \|E\|_\infty = \|(P_{\sim j})^{-1}\|_\infty \leq \sqrt{n-1}\|(P_{\sim j})^{-1}\|_2,$$

where $\|\cdot\|_2$ stands for the L_2 norm. Since the graph $\mathcal{G}$ is connected, the principal submatrix $P_{\sim j}$ is positive definite. As a result, it can be deduced from the above inequality that

$$\|(P_{\sim j})^{-1}E\|_\infty \leq \sqrt{n-1}\frac{1}{\lambda_{\min}\{P_{\sim j}\}}. \tag{6.30}$$

The proof is completed by combining the inequalities (6.23b) and (6.30). ■

Remark 3 *Theorem 6 states that the expected value of the convergence time in the worst case (i.e., t_c) is related to the $(n-1)^{th}$ order submatrices of the Laplacian of the graph, in addition to the Laplacian itself. Since the graph $\mathcal{G}$ is connected, $\lambda_2(P)$ is strictly positive.*

The interlacing theorem yields

$$0 < \lambda_{\min}\{P_{\sim j}\} \leq \lambda_2(P). \tag{6.31}$$

This means that unlike the unquantized consensus whose convergence mainly depends on $\lambda_2(P)$, a more subtle dependency on $\lambda_2(P)$ is governed for the quantized case. To be more precise, the convergence time depends on the minimum of $\lambda_{\min}\{P_{\sim j}\}$ (in addition to $\lambda_2(P)$), which is not directly related to $\lambda_2(P)$.

6.4.1 Special Graphs

This subsection aims to obtain lower and upper bounds on the quantity t_c for both complete and path graphs in the case when all edges have the same weight. In this regard, assume that every edge is associated with the same weight p.

Corollary 1 *For a complete graph $\mathcal{G}$ with equally weighted edges, the quantity t_c satisfies the inequalities (6.23a) and (6.23b), where*

$$\begin{aligned} \lambda_2(P) &= \frac{2}{n-1}, \\ \max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty &= \frac{n(n-1)}{2}. \end{aligned} \tag{6.32}$$

Proof: It is easy to verify that $\lambda_2(P)$ and $\max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty$ satisfy the equalities given in (6.32) in the case when $\mathcal{G}$ is a complete graph. The proof is completed by using Theorem 5. ■

Corollary 2 *Let $\mathcal{G}$ be a path graph with equally weighted edges such that vertex i is connected to vertex $i+1$ for every $i \in \{1, 2, ..., n-1\}$ (these are the only edges of the graph). The quantity t_c satisfies the inequalities (6.23a) and (6.23b), where*

$$\begin{aligned} \lambda_2(P) &= \frac{2}{n}\left(1 - \cos\frac{2\pi}{n}\right), \\ \max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty &= \frac{n(n-1)^2}{2}. \end{aligned} \tag{6.33}$$

Proof: The proof is a consequence of Theorem 5 after showing that the quantities $\lambda_2(P)$ and $\max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty$ satisfy the relations given in (6.33). The computation of $\lambda_2(P)$

for a path graph is straightforward. To find $\max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty$, first notice that the weight p is equal to $\frac{1}{n-1}$. On the other hand, it is evident that

$$\max_{j\in\mathcal{V}} \|(P_{\sim j})^{-1}E\|_\infty = \Phi = \boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{1n}\big\},$$

where $\boldsymbol{\beta}_{ij}$ is introduced in Theorem 4 for every $i, j \in \mathcal{V}$ such that $i \neq j$. The set of equations given in (6.20) gives rise to

$$p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{1n}\big\} - p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{2n}\big\} = 1, \tag{6.34a}$$

$$-p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{(i-1)n}\big\} + 2p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{in}\big\} - p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{(i+1)n}\big\} = 1, \tag{6.34b}$$

$$-p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{(n-2)n}\big\} + 2p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{(n-1)n}\big\} = 1, \tag{6.34c}$$

where the argument i in the equation (6.34b) belongs to the set $\{2, 3, ..., n-2\}$. Adding up these equalities results in the relation

$$p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{(n-1)n}\big\} = n - 1. \tag{6.35}$$

The (backward) recursive equation (6.34b) can be solved using conventional techniques to conclude that there exist two constants a and b such that

$$p\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{in}\big\} = a + bi - \frac{i^2}{2}, \quad i = n-1, n-2, ..., 1.$$

One can employ the final conditions given by (6.34c) and (6.35) to arrive at

$$a = \frac{n^2 - n}{2}, \quad b = \frac{1}{2}.$$

This implies that

$$\Phi = \boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\beta}_{1n}\big\} = \frac{n^2 - n}{2p} = \frac{n(n-1)^2}{2},$$

which completes the proof. ■

6.4.2 Optimal Edge Weights

In this subsection, it is desired to find out what probabilities the edges of $\mathcal{G}$ should possess so that the consensus is reached quickly. For this purpose, observe that the quantity t_c has been related to the spectral of the submatrices of the Laplacian in (6.29). Letting $x_{\min}$, $x_{\max}$ and ε be fixed, it follows from the upper bound on t_c provided in Theorem 6 and the inequality (6.31) that in order to minimize the convergence time in the worst case, a heuristic method is to minimize the term

$$\max_{i \in \mathcal{V}} \frac{1}{\lambda_{\min}\left\{P_{\sim i}\right\}}. \tag{6.36}$$

Hence, the goal is to minimize the function (6.36) over all possible (discrete) probability distributions captured by P for the sake of finding a sub-optimal edge-selection probability distribution. This is accomplished in the sequel.

Problem 1: Minimize the scalar variable $-\mu$ subject to the constraints

$$\lambda_{\min}\{P_{\sim i}\} \geq \mu, \quad i = 1, 2, ..., n,$$

where P is a matrix variable representing the Laplacian of the weighted graph $\mathcal{G}$. Denote the global minimizer of this optimization with (μ^*, P^*) (note that there are some implicit constraints stating that the weights on the edges are positive and sum up to 1).

Since the operator $\lambda_{\min}(\cdot)$ is concave with respect to its symmetric argument, it is easy to show that Problem 1 is convex. More precisely, the constraint $\lambda_{\min}\{P_{\sim i}\} \geq \mu$ can be expressed as $P_{\sim i} \succeq \mu I$, which is a semidefinite constraint. Hence, the solution P^* can be found efficiently. On the other hand, one can verify that

$$\mu^* = \max_{P} \min_{i} \lambda_{\min}\{P_{\sim i}\}$$

or equivalently

$$\frac{1}{\mu^*} = \min_{P} \max_{i} \frac{1}{\lambda_{\min}\{P_{\sim i}\}}.$$

This implies that the solution P^* corresponds to a sub-optimal edge-selection probability distribution (resulting in a fast convergence of the SG algorithm), because of minimizing the term given in (6.36). Note that solving Problem 1 needs the entire information of the graph

$\mathcal{G}$, which means that a standard algorithm for solving this problem will be a centralized one whose implementation in practice might be impossible. However, in line with the method discussed in [12], one can devise a distributed algorithm for finding P^*.

Remark 4 *The stochastic gossip algorithm studied in this work requires that an edge be selected at each time instant according to a pre-specified probability. Thus, one may speculate that a global coordinator is needed to be responsible for the edge-selection task, which makes the SG algorithm not really distributed. To address this issue, consider the distributed randomized gossip algorithm investigated in [12]: provide every vertex of the graph with a clock that ticks at the times of a rate 1 Poisson process such that whenever its clock ticks, it contacts one of its neighboring vertices using some pre-specified (local) probabilities to exchange quantized data. In other words, the distributed stochastic gossip algorithm given in [12] chooses first a vertex and then an edge connected to that vertex, instead of selecting an edge directly. Nonetheless, it is easy to show that every such distributed gossip algorithm corresponds to the stochastic gossip algorithm given here with some specific probability distribution $\mathcal{P}$ on the edges. Thus, the analysis provided in the present work is applicable to the distributed stochastic gossip algorithm discussed in [12].*

6.5 Simulation Results

Example 1: Consider a complete graph $\mathcal{G}$ with $n = 40$. Assume that all edges possess the same weight equal to $\frac{2}{n(n-1)}$ and that the initial values sitting on the vertices of $\mathcal{G}$ are uniformly distributed in the box $[0, 100]^n$. The intent is to understand how these values evolve under the SG algorithm. For this purpose, let $q(x)$ be a uniform quantizer and $\varepsilon = 0.2$. Two sets of initial states have been randomly generated, which are analyzed in the sequel:

- As the first trial, the initial values randomly generated are depicted in Figure 6.1a. Note that the x-axis of this plot shows the index i changing from 1 to 40, and the y-axis shows the corresponding value of $x_i[0]$. The average number x_{ave} and time instant $\tilde{k}$ introduced in Theorem 2 turned out to be equal to 45.98 and 658, respectively, for one particular run of the SG algorithm. The final values at the time $\tilde{k}$ are plotted in Figure 6.1b. Since these numbers are spread in the interval $[45.5, 46.5]$, the point $\mathbf{X}[\tilde{k}]$

belongs to the set $\mathcal{C}$. This implies that the steady state of the vector $\mathbf{X}[k]$ is fixed, i.e., $\mathbf{X}[k] = \mathbf{X}[\tilde{k}]$, for every $k \geq \tilde{k}$. The distance function $d(\mathbf{X}[k], \mathcal{C})$ is sketched in Figure 6.3a to illustrate how it attenuates to zero in a (non-strictly) decreasing way.

- As the second trial, the initial values randomly generated are shown in Figure 6.2a. The corresponding final values at time $\tilde{k} = 959$ are depicted in Figure 6.2b. This plot demonstrates that $\mathbf{X}[\tilde{k}]$ belongs to the set $\mathcal{C}(\eta_1)$, rather than $\mathcal{C}$ (note that $x_{\text{ave}} = 42.57$). Therefore, the steady-state behavior of the vector $\mathbf{X}[k]$ is oscillatory with probability 1. However, $x_i[k]$ ($i \in \mathcal{V}$ and $k \geq \tilde{k}$) can take only two possible values with the difference $\varepsilon = 0.2$, in light of the definition of $\mathcal{C}(\eta_1)$. The distance function $d(\mathbf{X}[k], \mathcal{C}(\eta_1))$ is plotted in Figure 6.3b to illustrate the convergence rate of the SG algorithm.

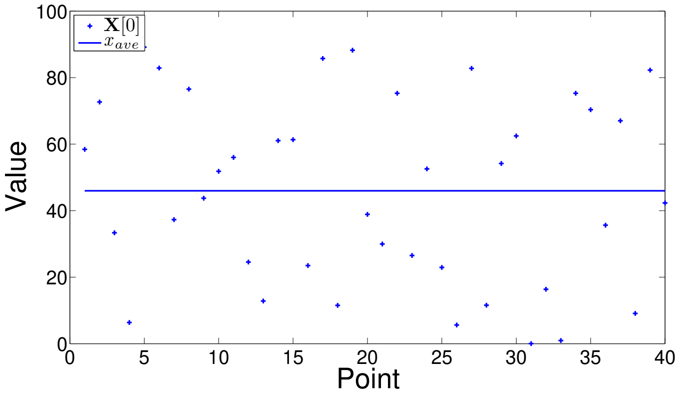
Example 2: Consider the graph $\mathcal{G}$ drawn in Figure 6.4. The objective is to find out what probabilities should be assigned to the edges of $\mathcal{G}$ so that the consensus is reached quickly under the SG algorithm with $q(x)$ being a uniform quantizer. To this end, let the convex optimization provided in Problem 1 be solved. This yields the following probability distribution:

$$p_{(1,2)} = p_{(1,5)} = 0.2087, \quad p_{(2,3)} = p_{(2,4)} = p_{(4,5)} = p_{(3,5)} = 0.1146, \quad p_{(3,4)} = 0.1241.$$

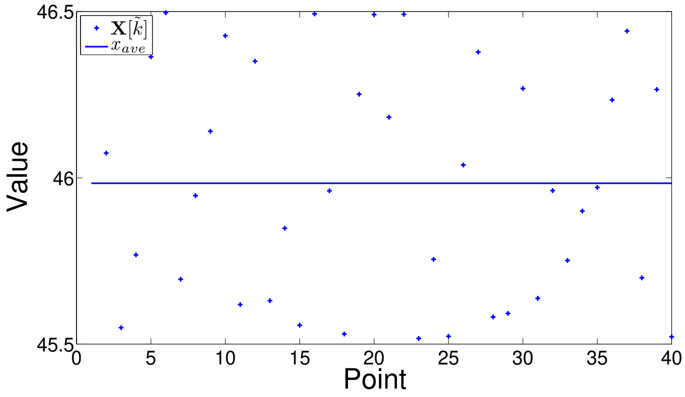
The quantity Φ corresponding to this set of edge-selection probabilities is 14.1770. One can make a comparison with two heuristic methods for designing the probability set $\mathcal{P}$, which are spelled out below:

- The most naive approach is to assume that the edges of the graph are equally weighted. This leads to the probability $p = \frac{1}{7}$ on each edge. The associated quantity Φ is obtained as 17.5.

- Another technique is to devise the probability distribution $\mathcal{P}$ in such a way that all

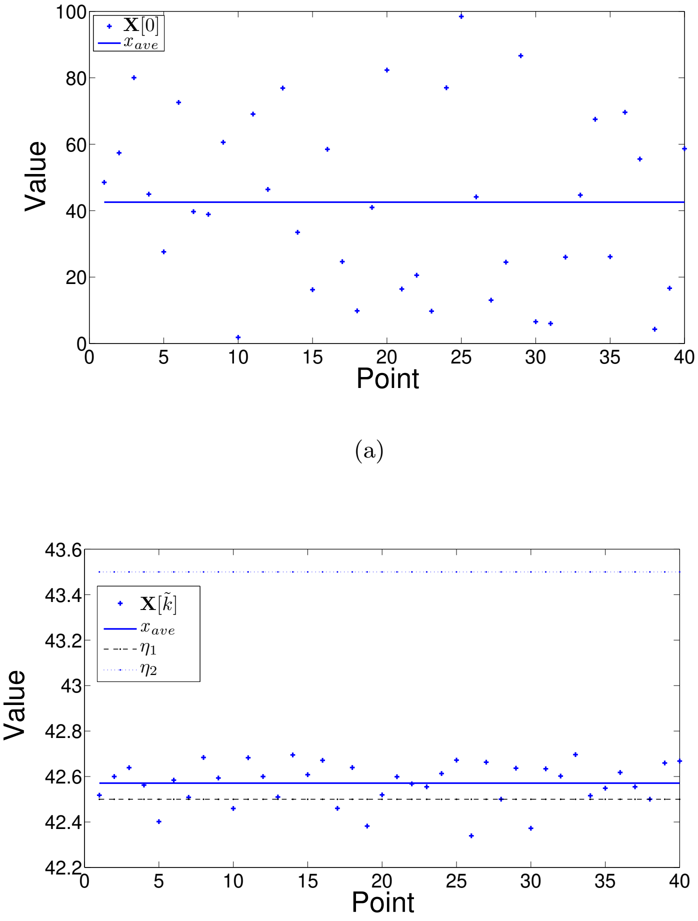


(a)



(b)

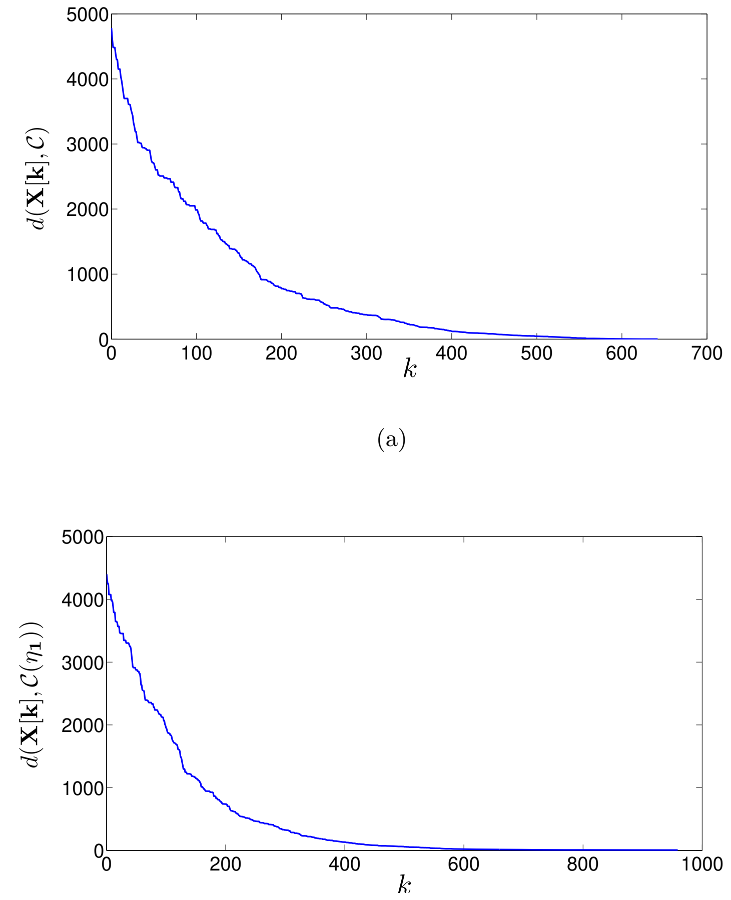
Figure 6.1: (a): The initial values on the vertices of the graph $\mathcal{G}$ for the first trial in Example 1; (b): the final values on the vertices of the graph $\mathcal{G}$ (at time $\tilde{k}$) for the first trial in Example 1



(a)

(b)

Figure 6.2: (a): The initial values on the vertices of the graph $\mathcal{G}$ for the second trial in Example 1; (b): the final values on the vertices of the graph $\mathcal{G}$ (at time $\tilde{k}$) for the second trial in Example 1



(a)

(b)

Figure 6.3: (a): The distance function $d(\mathbf{X}[k], \mathcal{C})$ for the first trial in Example 1; (b): the distance function $d(\mathbf{X}[k], \mathcal{C}(\eta_1))$ for the second trial in Example 1

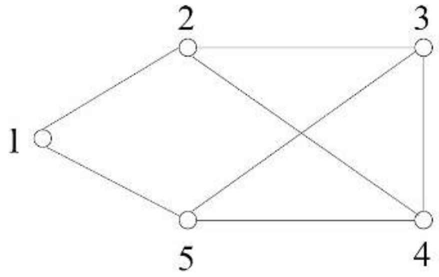


Figure 6.4: The graph $\mathcal{G}$ studied in Example 2

vertices have the same probability of being chosen at each time update, i.e.,

$$\begin{aligned} p_{(1,2)} + p_{(1,5)} &= p_{(2,1)} + p_{(2,3)} + p_{(2,4)} \\ &= p_{(3,2)} + p_{(3,4)} + p_{(3,5)} \\ &= p_{(4,2)} + p_{(4,3)} + p_{(4,5)} \\ &= p_{(5,1)} + p_{(5,3)} + p_{(5,4)}. \end{aligned}$$

Note that $p_{(i,j)} = p_{(j,i)}$, $\forall (i,j) \in \mathcal{E}$. The above set of equations has a unique symmetric solution (complying with the symmetry of the graph $\mathcal{G}$) as follows:

$$p_{(1,2)} = p_{(1,5)} = 0.2, \quad p_{(2,3)} = p_{(2,4)} = p_{(4,5)} = p_{(3,5)} = 0.1, \quad p_{(3,4)} = 0.2.$$

The corresponding Φ is equal to 15.

Hence, the value of Φ for the sub-optimal solution is better from the ones obtained using these two rudimentary techniques.

An interesting fact about the edge selection can be seen in this example. Remove the edge $(1,5)$ from the graph $\mathcal{G}$. In this case, Problem 1 leads to the solution

$$p_{(1,2)} = 0.3781, \quad p_{(2,3)} = p_{(2,4)} = 0.1757, \quad p_{(3,4)} = 0, \quad p_{(5,3)} = p_{(5,4)} = 0.1352 \tag{6.37}$$

for the new graph, associated with $\Phi = 23.1292$. Notice that $p_{(3,4)} = 0$, which indicates that although a complete graph has the best convergence, if some edges do not exist (e.g., the edge $(1,5)$), it might be better to ignore some other edges too (e.g., the edge $(3,4)$). This is interesting as it reveals the fact that some communications are redundant. If all edges of this new graph are assumed to have the same weight, Φ will be obtained as 36. Therefore, there is a noticeable improvement in the value of Φ via the solution of Problem 1.

For the purpose of simulation, the points

$$x_1[0] = 20.1185,\ x_2[0] = 13.6221,\ x_3[0] = 97.8356,\ x_4[0] = 45.5033,\ x_5[0] = 45.9224$$

have been randomly generated in the interval $[0, 100]$. The stochastic gossip algorithm was run 1000 times on the graph $\mathcal{G}$ with its edge (1,5) removed and the average of the random variable $\tilde{k}$ was calculated accordingly. This value for the probability distribution (6.37) was obtained as 48.5710, while this turned out to be 65.3580 for the identical probability distribution (equal edge weights). This demonstrates that one can save significantly in the convergence time if the solution of Problem 1 is deployed, which also obviates the usage of the edge $(3, 4)$.

6.6 Summary

This chapter deals with the distributed averaging problem over a connected weighted graph. The governing policy is that an edge of the graph is chosen at each time update with the probability equal to its weight, and then the values on its ending vertices are updated in terms of the quantized data of each other. A quantized stochastic gossip algorithm was proposed in a recent paper, which was shown to work in a particular case. In this chapter, it is proved that a quantized consensus is reached in the general case using this algorithm. Some steady-state properties of the numbers sitting on the vertices of the graph are obtained. Lower and upper bounds on the expected value of the convergence time in the worst case are also derived, which depend on the principal submatrices of the Laplacian matrix of the weighted graph. These bounds are explicitly computed for equally weighted complete and path graphs. Finally, a convex optimization is provided to obtain a set of weights on the edges of the graph that results in a fast convergence of the gossip algorithm.

6.7 Appendix 1

Proof of Lemma 2: Since the distance function $d(\cdot, \mathcal{C}(\mu))$ is always integer-valued and nonnegative, it follows from Lemma 1 that there exists a number $k_0 \in \mathbf{N}$ with the property that $d(\mathbf{X}[k], \mathcal{C}(\mu)) = d(\mathbf{X}[k_0], \mathcal{C}(\mu))$ for every natural number $k \geq k_0$ (see Remark 1). If $d(\mathbf{X}[k_0], \mathcal{C}(\mu)) = 0$, then case (i) given in the statement of the lemma definitely occurs. It remains to prove that if $d(\mathbf{X}[k_0], \mathcal{C}(\mu)) \neq 0$, then case (ii) takes place. To this end, notice that if $x_1[k] - \mu, ..., x_n[k] - \mu$ are negative (strictly positive) for some time k, then $x_1[k+1] - \mu, ..., x_n[k+1] - \mu$ are negative (strictly positive) as well. This implies that it suffices to prove case (ii) only for $k = k_0$.

By contradiction, assume that $x_1[k_0] - \mu, ..., x_n[k_0] - \mu$ are neither all negative nor all strictly positive. Thus, there are two indices $i, j \in \mathcal{V}$ such that $x_i[k_0] > \mu$ and $x_j[k_0] \leq \mu$. Consider a path between vertices i and j in the connected graph $\mathcal{G}$. There exists an edge $(i', j') \in \mathcal{E}$ in this path such that $x_{i'}[k_0] > \mu$ and $x_{j'}[k_0] \leq \mu$. If $x_{i'}[k_0] > \mu + \varepsilon$ or $x_{j'}[k_0] \leq \mu - \varepsilon$, then the optimality of the μ-DG algorithm together with Lemma 1 (on using the inequality (6.7)) yields

$$\begin{aligned} d(\mathbf{X}[k_0 + r], \mathcal{C}(\mu)) &\leq d(A^r(\mathbf{X}[k_0], (i', j'), ..., (i', j')), \mathcal{C}(\mu)) \\ &\leq d(A(\mathbf{X}[k_0], (i', j')), \mathcal{C}(\mu)) < d(\mathbf{X}[k_0], \mathcal{C}(\mu)). \end{aligned}$$

This contradicts the assumption that $d(\mathbf{X}[k], \mathcal{C}(\mu)) = d(\mathbf{X}[k_0], \mathcal{C}(\mu))$ for every $k \geq k_0$. As a result, the only remaining possibility is the following:

$$\mu - \varepsilon < x_{j'}[k_0] \leq \mu < x_{i'}[k_0] \leq \mu + \varepsilon. \tag{6.38}$$

Let j'_1 be a vertex connected to vertex j' in the graph $\mathcal{G}$. It is desired to show that

$$\mu - \varepsilon < x_{j'_1}[k_0] \leq \mu + \varepsilon. \tag{6.39}$$

To prove this by contradiction, consider the following scenarios:

- $x_{j'_1}[k_0]$ *is greater than* $\mu + \varepsilon$: As before, this case leads to a contradiction, because

$$\begin{aligned} d(\mathbf{X}[k_0], \mathcal{C}(\mu)) = d(\mathbf{X}[k_0 + r], \mathcal{C}(\mu)) &\leq d(A^r(\mathbf{X}[k_0], (j', j'_1), ..., (j', j'_1)), \mathcal{C}(\mu)) \\ &\leq d(A(\mathbf{X}[k_0], (j', j'_1)), \mathcal{C}(\mu)) < d(\mathbf{X}[k_0], \mathcal{C}(\mu)). \end{aligned}$$

- $x_{j_1'}[k_0]$ *is less than or equal to* $\mu - \varepsilon$: One can write

$$d(\mathbf{X}[k_0], \mathcal{C}(\mu)) = d(\mathbf{X}[k_0 + r], \mathcal{C}(\mu)) \leq d(A^r(\mathbf{X}[k_0], (i', j'), (j', j_1'), ..., (j', j_1')), \mathcal{C}(\mu))$$
$$\leq d(A^2(\mathbf{X}[k_0], (i', j'), (j', j_1')), \mathcal{C}(\mu)) < d(\mathbf{X}[k_0], \mathcal{C}(\mu)).$$

This is a contradiction as well.

So far, the validity of the inequality (6.39) is shown. Since the graph $\mathcal{G}$ is connected, there is a path from vertex j' to every other vertex of $\mathcal{G}$. Continuing the argument made above about vertex j_1' for all vertices of such paths successively and using the fact that every simple path has at most r edges give rise to

$$\mu - \varepsilon < x_p[k_0] \leq \mu + \varepsilon, \quad \forall p \in \mathcal{V}. \tag{6.40}$$

This inequality implies that $d(\mathbf{X}[k_0], \mathcal{C}(\mu))$ is equal to zero, while $d(\mathbf{X}[k_0], \mathcal{C}(\mu))$ was earlier assumed to be nonzero. This contradiction completes the proof. ■

6.8 Appendix 2

This appendix derives a number of results to prove Theorem 3. Consider an arbitrary infinite sequence of edges $\mathcal{H} \in \mathcal{E}^\infty$. Similar to the μ-DG algorithm, one can define an $\mathcal{H}$-G algorithm for every initial state $\mathbf{X}[0] \in [x_{\min}, x_{\max}]^n$ as follows.
$\mathcal{H}$-Gossip ($\mathcal{H}$-G) Algorithm:

Step 1: Set $k = 0$.

Step 2: Define $\mathbf{X}[k+1]$ to be $A(\mathbf{X}[k], (i, j))$, where (i, j) denotes the $(k+1)^{\text{th}}$ element of $\mathcal{H}$.

Step 3: Increase k by 1 and jump to step 2.

Note that unlike the μ-DG algorithm that picks an optimal edge relative to the set $\mathcal{C}(\mu)$ at each time update, the $\mathcal{H}$-G algorithm selects an edge from the sequence $\mathcal{H}$ in turn. For every $\mu \in \mathbf{Q}$, $\boldsymbol{\alpha} \in [x_{\min}, x_{\max}]^n$ and $\mathcal{H} \in \mathcal{E}^\infty$, let $T_\mu(\boldsymbol{\alpha}, \mathcal{H})$ denote the first time update at which a positive action occurs with respect to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$ if the $\mathcal{H}$-G algorithm is applied to the graph $\mathcal{G}$ with the initial state $\mathbf{X}[0] = \boldsymbol{\alpha}$.

For every $\mu \in \mathbf{Q}$, $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix} \in [x_{\min}, x_{\max}]^n$ and $i \in \mathcal{V}$, define $R_\mu(\boldsymbol{\alpha}, i)$ as

$$R_\mu(\boldsymbol{\alpha}, i) = \begin{cases} \begin{bmatrix} \alpha_1 & \cdots & \alpha_{i-1} & \alpha_i - \varepsilon & \alpha_{i+1} & \cdots & \alpha_n \end{bmatrix} & \text{if} \quad \alpha_i > \mu + \varepsilon \\ \begin{bmatrix} \alpha_1 & \cdots & \alpha_{i-1} & \alpha_i + \varepsilon & \alpha_{i+1} & \cdots & \alpha_n \end{bmatrix} & \text{if} \quad \alpha_i \le \mu - \varepsilon \\ \begin{bmatrix} \alpha_1 & \cdots & \alpha_{i-1} & \alpha_i & \alpha_{i+1} & \cdots & \alpha_n \end{bmatrix} & \text{otherwise} \end{cases}$$

The main idea behind the above definition is to convert a vector $\boldsymbol{\alpha}$ into another vector $R_\mu(\boldsymbol{\alpha}, i)$ that is closer to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$.

Lemma 3 *Given $\mu \in \mathbf{Q}$, $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$ and $\mathcal{H} \in \mathcal{E}^\infty$, the inequality*

$$T_\mu\big(\boldsymbol{\alpha}, \mathcal{H}\big) \le T_\mu\big(R_\mu(\boldsymbol{\alpha}, i), \mathcal{H}\big) \tag{6.41}$$

holds for every $i \in \mathcal{V}$.

Proof: Denote $\boldsymbol{\alpha}$ as $\begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix}$, and assume that $\alpha_i > \mu + \varepsilon$. Moreover, let $\mathbf{W}[k] := \begin{bmatrix} w_1[k] & w_2[k] & \cdots & w_n[k] \end{bmatrix}$ and $\bar{\mathbf{W}}[k] := \begin{bmatrix} \bar{w}_1[k] & \bar{w}_2[k] & \cdots & \bar{w}_n[k] \end{bmatrix}$ denote the states of the graph system at time $k \in \{0\} \cup \mathbf{N}$ under the $\mathcal{H}$-G algorithm with the initial states $\boldsymbol{\alpha}$ and $R_\mu(\boldsymbol{\alpha}, i)$, respectively. For notational simplicity, define

$$m := T_\mu\big(R_\mu(\boldsymbol{\alpha}, i), \mathcal{H}\big). \tag{6.42}$$

To prove the inequality (6.41) by contradiction, assume that $T_\mu\big(\boldsymbol{\alpha}, \mathcal{H}\big) > m$, which implies

$$d(\mathbf{W}[0], \mathcal{C}_{\boldsymbol{\alpha}}(\mu)) = d(\mathbf{W}[k], \mathcal{C}_{\boldsymbol{\alpha}}(\mu)), \quad \forall k \in \{0, 1, 2, ..., m\}. \tag{6.43}$$

In light of (6.42), (6.43), and the fact that $\boldsymbol{\alpha}$ and $R_\mu(\boldsymbol{\alpha}, i)$ are identical in $n-1$ entries, two properties can be proved by using an induction on the time instant k as follows:

i) $w_j[k]$ is always greater than or equal to $\bar{w}_j[k]$, for every $j \in \mathcal{V}$ and $k \in \{1, ..., m\}$.

ii) The relation $w_j[k] = \bar{w}_j[k]$ holds if $w_j[k] \le \mu$ or $\bar{w}_j[k] \le \mu$, for every $j \in \mathcal{V}$ and $k \in \{0, 1, ..., m-1\}$.

Let $(u, p) \in \mathcal{E}$ denote the m^{th} element of $\mathcal{H}$ such that $\bar{w}_u[m-1] \ge \bar{w}_p[m-1]$. Since a positive action occurs at time m with respect to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$ for the initial state $R_\mu(\boldsymbol{\alpha}, i)$

(due to the equation (6.42)), the proof of Lemma 1 (especially the inequality (6.7)) can be used to conclude that either

$$\bar{w}_u[m-1] > \mu + \varepsilon, \quad \bar{w}_p[m-1] \leq \mu \tag{6.44}$$

or

$$\bar{w}_u[m-1] > \mu, \quad \bar{w}_p[m-1] \leq \mu - \varepsilon$$

must hold. Assume that the relations given in (6.44) hold (the other case is similar). Properties (i) and (ii) mentioned above yield

$$\begin{aligned} &w_u[m-1] \geq \bar{w}_u[m-1] > \mu + \varepsilon, \\ &w_p[m-1] = \bar{w}_p[m-1] \leq \mu. \end{aligned}$$

Since the same edge (u, p) is chosen at the m^{th} time update by the $\mathcal{H}$-G algorithm for the initial state $\boldsymbol{\alpha}$, it follows immediately from the above relations and the inequality (6.7) that a positive action occurs at this time. This contradicts the assumption that $T_\mu\big(\boldsymbol{\alpha}, \mathcal{H}\big) > m$, which completes the proof for the case when $\alpha_i > \mu + \varepsilon$. The proof is similar for the other cases. ■

Lemma 3 states that if an initial state $\boldsymbol{\alpha}$ is replaced by $R_\mu(\boldsymbol{\alpha}, i)$ that is closer to the set $\mathcal{C}_{\boldsymbol{\alpha}}(\mu)$, then more iterations are required to take a positive action. Similar to the multi-action function A^k that was defined in terms of the single-action function A in (6.12) (where $k \in \mathbf{N}$), one can define the multi-action function R_μ^k from R_μ. For every $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix} \in \tilde{\mathcal{C}}(\mu)$ and $\mu \in \mathbf{Q}$, let $R_\mu(\boldsymbol{\alpha})$ be a vector obtained using the following procedure:

Step 1: Identify the smallest index $j \in \mathcal{V}$ such that $\alpha_j > \mu + \varepsilon$ or $\alpha_j \leq \mu - \varepsilon$ (such an index exists due to the definition of $\tilde{\mathcal{C}}(\mu)$).

Step 2: Set $\boldsymbol{\gamma}$ to be $\boldsymbol{\alpha}$.

Step 3: For every $i \in \mathcal{V}$, find the smallest number $r \in \mathbf{N}$ such that

$$R_\mu^r(\boldsymbol{\gamma}, \underbrace{i, i, ..., i}_{r \text{ times}}) = R_\mu^{r-1}(\boldsymbol{\gamma}, \underbrace{i, i, ..., i}_{r-1 \text{ times}}).$$

Update the new value of $\boldsymbol{\gamma}$ as $R_\mu^r(\boldsymbol{\gamma}, i, i, ..., i)$ if $i \in \mathcal{V}\backslash\{j\}$, and otherwise as $R_\mu^{r-1}(\boldsymbol{\gamma}, i, i, ..., i)$ (by convection, assume that $R_\mu^0(\boldsymbol{\gamma}, i) = \boldsymbol{\gamma}$).

Step 4: Define $R_\mu(\boldsymbol{\alpha})$ as $\boldsymbol{\gamma}$.

Remark 5 *The vector $R_\mu(\boldsymbol{\alpha})$ is derived from $\boldsymbol{\alpha}$ in such a way that all entries of $R_\mu(\boldsymbol{\alpha})$ lie in the interval $(\mu - \varepsilon, \mu + \varepsilon]$, except for only one entry that belongs to either $(\mu + \varepsilon, \mu + 2\varepsilon]$ or $(\mu - 2\varepsilon, \mu - \varepsilon]$. Note that $R_\mu(\boldsymbol{\alpha})$ satisfies the inequality*

$$T_\mu\big(\boldsymbol{\alpha}, \mathcal{H}\big) \leq T_\mu\big(R_\mu(\boldsymbol{\alpha}), \mathcal{H}\big) \tag{6.45}$$

for every $\mathcal{H} \in \mathcal{E}^\infty$, because of Lemma 3.

For every $\mu \in \mathbf{Q}$, $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix} \in [x_{\min}, x_{\max}]^n$ and $i, j \in \mathcal{V}$, define $R'_\mu(\boldsymbol{\alpha}, i; j)$ as

$$R'_\mu(\boldsymbol{\alpha}, i; j) = \begin{cases} \begin{bmatrix} \alpha_1 & \cdots & \alpha_{i-1} & \alpha_i + \varepsilon & \alpha_{i+1} & \cdots & \alpha_n \end{bmatrix} & \text{if} \quad \alpha_i \leq \mu, \ \alpha_j > \mu + \varepsilon \\ \begin{bmatrix} \alpha_1 & \cdots & \alpha_{i-1} & \alpha_i - \varepsilon & \alpha_{i+1} & \cdots & \alpha_n \end{bmatrix} & \text{if} \quad \alpha_i > \mu, \ \alpha_j \leq \mu - \varepsilon \\ \begin{bmatrix} \alpha_1 & \cdots & \alpha_{i-1} & \alpha_i & \alpha_{i+1} & \cdots & \alpha_n \end{bmatrix} & \text{otherwise.} \end{cases}$$

The main idea behind the definition of $R'_\mu(\boldsymbol{\alpha}, i; j)$ is to make the i^{th} entry of $\boldsymbol{\alpha}$ become closer to its j^{th} entry. The next lemma presents a useful relationship between the R_μ and R'_μ functions.

Lemma 4 *Given $\mu \in \mathbf{Q}$, $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$ and $\mathcal{H} \in \mathcal{E}^\infty$, let $j \in \mathcal{V}$ denote the index of the unique entry of $R_\mu(\boldsymbol{\alpha})$ that does not lie in the interval $(\mu - \varepsilon, \mu + \varepsilon]$. The inequality*

$$T_\mu\big(R_\mu(\boldsymbol{\alpha}), \mathcal{H}\big) \leq T_\mu\big(R'_\mu(R_\mu(\boldsymbol{\alpha}), i; j), \mathcal{H}\big) \tag{6.46}$$

holds for every $i \in \mathcal{V}$.

Proof: let $\mathbf{U}[k] := \begin{bmatrix} w_1[k] & w_2[k] & \cdots & w_n[k] \end{bmatrix}$ and $\bar{\mathbf{U}}[k] := \begin{bmatrix} \bar{w}_1[k] & \bar{w}_2[k] & \cdots & \bar{w}_n[k] \end{bmatrix}$ denote the states of the graph system at time $k \in \{0\} \cup \mathbf{N}$ under the $\mathcal{H}$-G algorithm with the initial states $R_\mu(\boldsymbol{\alpha})$ and $R'_\mu(R_\mu(\boldsymbol{\alpha}), i; j)$, respectively. Define also

$$g := T_\mu\big(R'_\mu(R_\mu(\boldsymbol{\alpha}), i; j), \mathcal{H}\big).$$

To prove (6.46) by contradiction, assume that $T_\mu\big(\boldsymbol{\alpha}, \mathcal{H}\big) > g$. Let $R_\mu(\boldsymbol{\alpha})$ be denoted as $\left[\begin{array}{ccc} \beta_1 & .... & \beta_n \end{array}\right]$ and with no loss of generality assume that $\beta_j \in (\mu+\varepsilon, \mu+2\varepsilon]$. Two observations can be made as follows:

i) The relations

$$u_j[k] = \bar{u}_j[k] = \beta_j,$$
$$u_p[k], \bar{u}_p[k] \in (\mu-\varepsilon, \mu+\varepsilon], \quad \forall p \in \mathcal{V}\backslash\{j\},$$

hold for every $k \in \{0, 1, ..., g-1\}$.

ii) Using property (i) and by means of an induction on k, one can show that if $\bar{u}_p[k] \leq \mu$ for some $p \in \mathcal{V}$ and $k \in \{0, 1, ..., g-1\}$, then $u_p[k] = \bar{u}_p[k]$.

Let the g^{th} element of $\mathcal{H}$ be the edge (r_1, r_2), where $r_1 < r_2$. It results from the definition of g and property (i) that $r_1 = j$ and $\bar{u}_{r_2}[g-1] \leq \mu$ (this is the only way to generate a positive action at time g for the initial state $R'_\mu(R_\mu(\boldsymbol{\alpha}), i; j))$. Therefore, by properties (i) and (ii), one can write

$$u_{r_2}[g-1] = \bar{u}_{r_2}[g-1] \leq \mu,$$
$$u_{r_1}[g-1] > \mu + \varepsilon.$$

The inequality (6.7) can be used to conclude that selecting the edge (r_1, r_2) at time g results in a positive action for the graph $\mathcal{G}$ with the initial state $R_\mu(\boldsymbol{\alpha})$, which implies that $T_\mu\big(R_\mu(\boldsymbol{\alpha}), \mathcal{H}\big) = g$. This contradicts the aforementioned assumption. ■

For every $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$, let $R'_\mu(R_\mu(\boldsymbol{\alpha}))$ be an n-dimensional vector that is obtained from $R_\mu(\boldsymbol{\alpha})$ using the following procedure:

Step 1: Identify the unique index $j \in \mathcal{V}$ such that the j^{th} entry of $R_\mu(\boldsymbol{\alpha})$ is not in the range $(\mu-\varepsilon, \mu+\varepsilon]$.

Step 2: Identify the smallest index $i \in \mathcal{V}$ such that

- The i^{th} entry of $R_\mu(\boldsymbol{\alpha})$ is greater than μ if the j^{th} entry of $R_\mu(\boldsymbol{\alpha})$ is less than or equal to μ.
- The i^{th} entry of $R_\mu(\boldsymbol{\alpha})$ is less than or equal to μ if the j^{th} entry of $R_\mu(\boldsymbol{\alpha})$ is greater than μ.

Step 3: Set $\boldsymbol{\gamma}$ to be $\boldsymbol{\alpha}$.

Step 4: For every $p \in \mathcal{V}\backslash\{i,j\}$, update the new value of $\boldsymbol{\gamma}$ as $R'_\mu(R_\mu(\boldsymbol{\gamma}), p; j)$.

Step 4: Define $R'_\mu(R_\mu(\boldsymbol{\alpha}))$ as $\boldsymbol{\gamma}$.

Proposition 1 *Given $\mu \in \mathbf{Q}$, $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$ and $\mathcal{H} \in \mathcal{E}^\infty$, the inequality*

$$T_\mu(\boldsymbol{\alpha}, \mathcal{H}) \leq T_\mu(R'_\mu(R_\mu(\boldsymbol{\alpha})), \mathcal{H}) \tag{6.47}$$

holds.

Proof: The proof is a direct consequence of the inequality (6.45) and Lemma 4. ■

Proposition 1 will be used in the sequel to present the main result of this appendix, which is a proof for Theorem 3.

Proof of Theorem 3: It follows from the inequality (6.47) that

$$\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\alpha}\big\} \leq \boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = R'_\mu(R_\mu(\boldsymbol{\alpha}))\big\},$$

where $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$. Hence

$$\begin{aligned}\Phi(\mu) &= \max\Big\{\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\alpha}\big\} \;\Big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)\Big\} \\ &\leq \max\Big\{\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = R'_\mu(R_\mu(\boldsymbol{\alpha}))\big\} \;\Big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)\Big\}.\end{aligned} \tag{6.48}$$

On the other hand, the simple set inclusion property

$$\big\{R'_\mu(R_\mu(\boldsymbol{\alpha}))\big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)\big\} \subseteq \tilde{\mathcal{C}}(\mu)$$

yields

$$\begin{aligned}&\max\Big\{\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = R'_\mu(R_\mu(\boldsymbol{\alpha}))\big\} \;\Big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)\Big\} \\ &\qquad \leq \max\Big\{\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = \boldsymbol{\alpha}\big\} \;\Big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)\Big\} = \Phi(\mu).\end{aligned} \tag{6.49}$$

It can be concluded from (6.48) and (6.49) that

$$\Phi(\mu) = \max\Big\{\boldsymbol{\mathcal{E}}\big\{T(\mu)\big|\mathbf{X}[0] = R'_\mu(R_\mu(\boldsymbol{\alpha}))\big\} \;\Big|\; \boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)\Big\}.$$

The rest of the proof replies on the above equality and the fact that for every $\boldsymbol{\alpha} \in \tilde{\mathcal{C}}(\mu)$, the vector $R'_\mu(R_\mu(\boldsymbol{\alpha}))$ satisfies either of the following properties:

- $n-2$ entries of $R'_\mu(R_\mu(\boldsymbol{\alpha}))$ are in the interval $(\mu, \mu+\varepsilon]$, and the two other entries are in the intervals $(\mu+\varepsilon, \mu+2\varepsilon]$ and $(\mu-\varepsilon, \mu]$.

- $n-2$ entries of $R'_\mu(R_\mu(\boldsymbol{\alpha}))$ are in the interval $(\mu-\varepsilon, \mu]$, and the two other entries are in the intervals $(\mu-2\varepsilon, \mu-\varepsilon]$ and $(\mu, \mu+\varepsilon]$.

The details are omitted for brevity. ■

Chapter 7

Congestion Control Algorithms from Optimal Control Perspective

This chapter is concerned with understanding the connection between the existing Internet congestion control algorithms and the optimal control theory. The available resource allocation controllers are mainly devised to derive the state of the system to a desired equilibrium point and, therefore, they are oblivious to the transient behavior of the closed-loop system. To take into account the real-time performance of the system, rather than merely its steady-state performance, the congestion control problem should be solved by maximizing a proper utility functional as opposed to a utility function. For this reason, this work aims to investigate what utility functionals the existing congestion control algorithms maximize. In particular, it is shown that there exist meaningful utility functionals whose maximization leads to the celebrated primal, dual, and primal/dual algorithms. An implication of this result is that a real network problem may be solved by regarding it as an optimal control problem on which some practical constraints, such as a real-time link capacity constraint, are imposed.

7.1 Introduction

There has been a growing interest in studying the Internet congestion control ever since the first congestion collapse occurred [46]. Many algorithms have been proposed in the literature to allocate the available network resources in a fair manner among the competing users, without overloading the network. The main idea behind all these algorithms is more or less the same: each user measures some feedback signal, such as packet loss or queueing delay, and accordingly adapts its transmission rate. Among the existing transmission control protocols (TCP) for congestion control, one can name TCP-Tahoe, Reno, New Reno, and Vegas [87, 14]. More complete surveys of this topic can be found in [98], [101] and [23].

The seminal papers [53] and [54] sparked remarkable process in mathematical modeling and analysis of the Internet congestion control. This advancement is due to the convex programming theory, which allows for solving a utility maximization problem by means of the Lagrangian technique. The available resource allocation algorithms, such as the primal, dual, and primal/dual algorithms, are particularly designed to solve the underlying problem in a distributed way asymptotically. In other words, these algorithms guarantee that the asymptotic transmission rate of each user is the fairest rate that can be utilized without congesting the network. Having regarded the network as a system, this result implies that the control system possesses a unique globally asymptotically stable equilibrium point that corresponds to the solution of the static utility maximization problem. Nonetheless, it is not clear how well the system operates during its transient time. As a result, the capacity link constraints can, for instance, be violated in this period. Furthermore, these algorithms have not been derived in such a way that they can be generalized systematically to include real-time constraints such as a link capacity requirement. This work aims to revisit the congestion control problem from the standpoint of the optimal control theory.

This chapter proves that the controllers proposed by the primal, dual, and primal/dual algorithms all maximize some meaningful dynamical behaviors. More precisely, there exist natural utility functionals whose maximization leads to these celebrated controllers. This result opens the possibility of tackling network problems directly as optimal control problems, which not only take the dynamics into account, but which also allow to impose physical constraints. Other applications of dealing with utility functionals directly are in deducing the stability of the control system for free, gaining insight into how to perform joint routing and congestion control, etc. It is noteworthy that the development of this work relies on the inverse optimal control theory, which has a very ancient history [94, 75].

7.2 Preliminaries

Consider a network with the set of sources $\mathcal{S}$ and the set of links $\mathcal{L}$, where each source is identified by an origin and a destination between which data can be transferred. For every

$r \in \mathcal{S}$, let x_r denote the transmission rate corresponding to source r and $\mathcal{L}(r)$ denote the collection of links belonging to its fixed route. Assume that each link $l \in \mathcal{L}$ has a finite capacity c_l. Form a vector of transmission rates, denoted by $\mathbf{x}$, where its r^{th} element is equal to x_r for all $r \in \mathcal{S}$. The resource allocation problem is concerned with solving the optimization

$$\max_{\mathbf{x}} \sum_{r \in \mathcal{S}} U_r(x_r)$$

subject to

$$\sum_{r:\ l \in \mathcal{L}(r)} x_r \leq c_l, \quad \forall\, l \in \mathcal{L}$$

$$x_r \geq 0, \quad \forall\, r \in \mathcal{S},$$

where $U_r : \Re \rightarrow \Re$ is a strictly concave, increasing and twice differentiable utility function associated with source r. Define R to be a routing matrix whose (l, r) entry $(r \in \mathcal{S},\ l \in \mathcal{L})$ is equal to 1 if $l \in \mathcal{L}(r)$, and is 0 otherwise. Define also the aggregate flow rate y_l, the route price q_r and the Lagrangian $L(\mathbf{x}, \mathbf{p})$ as

$$\begin{aligned} y_l &:= \sum_{r:\ l \in \mathcal{L}(r)} x_r, \quad l \in \mathcal{L}, \\ q_r &:= \sum_{l \in \mathcal{L}(r)} p_l, \quad r \in \mathcal{S}, \\ L(\mathbf{x}, \mathbf{p}) &:= \sum_{r \in \mathcal{S}} U_r(x_r) - \sum_{l \in \mathcal{L}} p_l \left(y_l - c_l\right), \end{aligned} \tag{7.1}$$

where $\mathbf{p}$ is the vector of Lagrange multipliers $p_l,\ l \in \mathcal{L}$. The Karush-Kuhn-Tucker (KKT) conditions for the utility maximization problem are

$$\begin{aligned} U'(x_r) &= q_r, \\ p_l(y_l - c_l) &= 0, \\ y_l - c_l &\leq 0, \\ x_r, p_l &\geq 0, \end{aligned}$$

for all $l \in \mathcal{L}$ and $r \in \mathcal{S}$. Having assumed that R has full row rank, the above KKT equations have a unique solution $(\mathbf{x}^*, \mathbf{p}^*)$ [101]. Since each user $r \in \mathcal{S}$ must obtain its optimal transmission rate x_r^* in terms of the available *local* information, a number of distributed algorithms have been proposed in the literature to enable every user to adaptively find its optimal transmission rate. One of these algorithms is briefly outlined in the sequel.

7.2.1 Dual Algorithm

Assume that each link $l \in \mathcal{L}$ updates its associated price p_l based on the rule

$$\dot{p}_l(t) = h_l(p_l(t))(y_l(t) - c_l)^+_{p_l(t)}, \tag{7.2}$$

where $h_l : \Re \rightarrow \Re^+$ is a given non-decreasing continuous function and

$$(y_l(t) - c_l)^+_{p_l(t)} = \begin{cases} y_l(t) - c_l & p_l(t) > 0 \\ \max(y_l(t) - c_l, 0) & p_l(t) = 0. \end{cases}$$

Moreover, suppose that the user of each source $r \in \mathcal{S}$ is provided with the aggregate price along its route to update its transmission rate as

$$x_r(t) = U_r'^{-1}(q_r(t)). \tag{7.3}$$

It is well-known that the interconnected system specified by (7.2) and (7.3) is globally asymptotically stable with the unique equilibrium point $(\mathbf{x}^*, \mathbf{p}^*)$ [101].

7.3 Motivation and Problem Formulation

The main idea behind the existing congestion control algorithms is to contrive a distributed control system which has a unique equilibrium point $(\mathbf{x}^*, \mathbf{p}^*)$ that is globally asymptotically stable. However, this interesting technique is oblivious to the transient behavior of the system and merely targets its steady-state behavior. As a result, the link capacity constraints may be violated during the transient time. Moreover, these indirect congestion control algorithms cannot be generalized systematically. For instance, it is pragmatic to impose a buffer size constraint or to assume that each source has a certain amount of data to transfer. These practical constraints, along with many other ones, cannot be incorporated into the aforementioned algorithms in light of the fact that these algorithms essentially rely on the static utility maximization problem to which these constraints cannot be applied. By regarding the network as a system with a specific topology, a question arises as to whether one can define an optimal control problem for this system whose solution leads to a distributed controller solving the utility maximization problem. This chapter aims to show that the answer to this fundamental question is affirmative, and that working directly with the network problems in the context of optimal control theory allows the designer to incorporate other physical constraints and deduce some properties for free such as stability.

The objective is to prove that the updating policies proposed by the primal, dual, and primal/dual algorithms can all be obtained by maximizing appropriate utility functionals which take the transient response of the system into account. Nevertheless, it is well-understood that even though an optimal control problem normally has a unique solution, there might be an infinite number of optimal control problems which all lead to the same solution. For instance, consider the simple first-order system $\dot{p}(t) = x(t)$, where $p(t)$ and $x(t)$ are its state and input, respectively. Note that although $x(t)$ is a standard notation for representing the state of a system, this chapter needs to use this notation to denote the input of a system (as it corresponds to the transmission rate that acts as an input). Given positive numbers k and T, there exists a unique controller that maximizes the utility functional

$$-\int_0^T \left(\frac{x(t)^2}{k} + kp(t)^2\right) dt - p(T)^2.$$

This controller turns out to be $x(t) = -kp(t)$. However, there are other utility functionals whose maximization leads to this controller. For example, the trivial term $(x(t) + kp(t))^2$ can be added to the integrand of the above utility functional without altering the optimal solution. It can be shown in this example that all such functionals can be characterized systematically, provided the terminal utility is fixed as $-p(T)^2$. To be more precise, assume that the maximization of the utility functional

$$\int_0^T g(p(t), x(t))dt - p(T)^2 \tag{7.4}$$

yields the controller $x(t) = -kp(t)$, where $g(p(t), x(t))$ is some appropriate function. One can verify that there exist a function $\hat{g}(p(t), x(t))$ and a constant number μ such that

$$g(p(t), x(t)) = \mu + \hat{g}(p(t), x(t)) - \frac{x(t)^2}{k} - kp(t)^2, \tag{7.5}$$

where $\hat{g}(p(t), x(t))$ is equal to zero along all trajectories of the optimal closed-loop system.

This simple toy example implies that there are an infinite number of utility functionals which solve the inverse optimal problem; nevertheless, they all share some key part that determines the trade-off between the state and the input which has caused the optimal controller to be identical to the given one.

The above discussion signifies that there might be numerous utility functionals associated with the static utility maximization problem. The primary objective is to identify their common part that has meaningful physical interpretations. It will be later shown that there is a close parallel (term by term) between the functionals solving the utility maximization problem and the ones characterized in (7.4) and (7.5).

7.4 Optimal Control for Dual Algorithm

Having provided each user r with its route price that is obtained based on some prespecified rule, assume that the user is required to find the best updating policy to adjust its transmission rate x_r. This hypothesis implies that the dynamical system

$$\dot{p}_l(t) = h_l(p_l(t))(y_l(t) - c_l)^+_{p_l(t)}, \quad l \in \mathcal{L} \tag{7.6}$$

exists in the core of the network to generate link prices, where $\mathbf{p}(t)$ and $\mathbf{x}(t)$ are the state and the input of the system, respectively. It is desired to find a utility functional whose maximization leads to the local controllers

$$x_r(t) = U_r'^{-1}(q_r(t)), \quad r \in \mathcal{S}. \tag{7.7}$$

7.4.1 Simple Illustrative Example

Before handling the problem in the general case, let the main ideas be elucidated in a very simple example. As a trivial but illustrative case, assume that:

- The network has only one source and one link.

- The capacity of the link is equal to 1.

- The utility function $U(x)$ is equal to $-0.5(x-4)^2$ if $x \in [0, 3]$.

- The weighting function $h(p)$ is identical to 1.

Note that since $\mathcal{S}$ and $\mathcal{L}$ each have one element, the indices l and r are omitted. Moreover, although the utility function $U(x)$ is defined only on the interval of interest $[0,3]$, it can be extended smoothly to the entire interval $[0,\infty)$. For simplicity, suppose that the value of the initial price $p(0)$ is chosen so that the transmission rate $x(t)$ always stays in the interval $[0,3]$, and that the price $p(t)$ never hits zero. The problem now reduces to finding a utility functional whose maximization leads to the controller

$$x(t) = -q(t) + 4$$

for the system

$$\dot{q}(t) = x(t) - 1.$$

In order to eliminate the constant terms in the above equations, introduce the change of variables

$$\begin{aligned} \bar{x}(t) &= x(t) - 1, \\ \bar{q}(t) &= q(t) - 3. \end{aligned}$$

In the new coordinates, the system and the controller turn out to be $\dot{\bar{q}}(t) = \bar{x}(t)$ and $\bar{x}(t) = -\bar{q}(t)$, respectively. This control system has been studied in the toy example of the previous section (assuming $k = 1$), for which the utility functional

$$-\int_0^T \left(\bar{x}(t)^2 + \bar{q}(t)^2\right) dt - \bar{q}(T)^2$$

was found. One can rewrite the above expression in terms of the original variables to obtain

$$-\int_0^T \Big((x(t)-1)^2 + (q(t)-3)^2\Big)\, dt - (q(T)-3)^2. \tag{7.8}$$

To relate the terms in the above functional to the static utility maximization problem, notice that

$$\begin{aligned} 3 - q(t) &= U'^{-1}(q(t)) - 1 = \arg\max_v L(v, q(t)) \\ (q(T)-3)^2 &= 2\max_v L(v, q(T)) + 9. \end{aligned}$$

Substituting the above relations into (7.8), one can conclude that maximizing the utility functional given below leads to the dual controller:

$$-\frac{1}{2}\int_0^T \left((x(t)-c)^2 + \left(\arg\max_v L(v,q(t)) - c\right)^2\right) dt - \max_v L(v,q(T)).$$

As can be inferred from the toy example in Section 7.3, every other utility functional that is able to solve the underling inverse optimal problem includes the integrand of the above functional, in addition to some trivial terms, provided its terminal utility is chosen as above. This result will be generalized in the sequel, and the interpretation of the individual terms appearing in this utility functional will then be discussed in detail.

7.4.2 General Case

The next theorem extends the above-mentioned results to the general case.

Theorem 1 *Given $T > 0$, the decentralized controller given in (7.7) maximizes the utility functional*

$$\max_{\mathbf{x}(t)} \left\{ \frac{1}{2}\int_0^T \sum_{l\in\mathcal{L}} \{Y_l(y_l(t),p_l(t)) + Y_l(\tilde{y}_l(\mathbf{p}(t)),p_l(t))\}dt - \max_{\mathbf{v}(T)} L(\mathbf{v}(T),\mathbf{p}(T))\right\} \tag{7.9}$$

for the system (7.6), where

$$Y_l(\alpha,p_l(t)) := -(\alpha-c_l)h_l(p_l(t))(\alpha-c_l)^+_{p_l(t)}$$

for every $\alpha \in \Re$, $l \in \mathcal{L}$, and

$$\tilde{\mathbf{y}}(\mathbf{p}(t)) := R \times \arg\max_{\mathbf{v}(t)} L(\mathbf{v}(t),\mathbf{p}(t))$$

($\tilde{y}_l(\mathbf{p}(t))$ is equal to the l^{th} entry of $\tilde{\mathbf{y}}(\mathbf{p}(t))$).

Proof: Define the optimal cost-to-go function $J(\mathbf{p},t)$, $t\in[0,T]$, to be

$$\begin{aligned} J(\mathbf{p},t) := \max_{\mathbf{x}(s)} \Bigg\{ \frac{1}{2}\int_t^T \sum_{l\in\mathcal{L}} \{Y_l(y_l(s),p_l(s)) \\ + Y_l(\tilde{y}_l(\mathbf{p}(s)),p_l(s))\}ds - \max_{\mathbf{v}(T)} L(\mathbf{v}(T),\mathbf{p}(T))\Bigg\}, \end{aligned}$$

where the system starts at time t with an initial state $\mathbf{p}$ whose entries are all nonnegative. The Hamilton-Jacobi-Bellman (HJB) method [55] states that $J(\mathbf{p},t)$ satisfies the partial differential equation

$$\begin{aligned} 0 = \frac{\partial J(\mathbf{p},t)}{\partial t} &+ \max_{\mathbf{x}} \left\{ \frac{1}{2} \sum_{l\in\mathcal{L}} \left\{ Y_l(y_l,p_l) + Y_l(\tilde{y}_l(\mathbf{p}),p_l) \right\} \right. \\ &+ \left. \sum_{l\in\mathcal{L}} h_l(p_l)(y_l-c_l)^+_{p_l} \frac{\partial J(\mathbf{p},t)}{\partial p_l} \right\} \end{aligned} \tag{7.10}$$

with the boundary condition

$$J(\mathbf{p},T) = -\max_{\mathbf{v}} L(\mathbf{v},\mathbf{p}).$$

Solving the HJB differential equation is cumbersome in general. However, it is desired to show that this equation takes the simple solution $J(\mathbf{p},t) = J(\mathbf{p},T)$, $\forall t \in [0,T]$ in this problem. To this end, observe that

$$\tilde{y}_l(\mathbf{p}) = \sum_{r:\ l\in\mathcal{L}(r)} U_r'^{-1}(q_r).$$

Since $\mathbf{p}$ is a nonnegative vector, the maximum of the Lagrangian $L(\mathbf{v},\mathbf{p})$ (with respect to $\mathbf{v}$) is achieved when

$$v_r = U_r'^{-1}(q_r), \quad r \in \mathcal{S},$$

where v_r denotes the r^{th} entry of $\mathbf{v}$, for all $r\in\mathcal{S}$. For the above-mentioned choice of $J(\mathbf{p},t)$, it can be verified that

$$\begin{aligned} &\frac{\partial J(\mathbf{p},t)}{\partial t} = 0, \\ &\frac{\partial J(\mathbf{p},t)}{\partial p_l} = \tilde{y}_l(\mathbf{p}) - c_l, \quad \forall\, l \in \mathcal{L}. \end{aligned} \tag{7.11}$$

Using these equalities, one can also check that the input $\mathbf{x}$ given by

$$x_r = U_r'^{-1}(q_r), \quad r \in \mathcal{S} \tag{7.12}$$

maximizes the objective functional

$$\frac{1}{2}\sum_{l\in\mathcal{L}} \left\{ Y_l(y_l,p_l) + Y_l(\tilde{y}_l(\mathbf{p}),p_l) \right\} + \sum_{l\in\mathcal{L}} h_l(p_l)(y_l-c_l)^+_{p_l} \frac{\partial J(\mathbf{p},t)}{\partial p_l}$$

with respect to $\mathbf{x}$. By substituting the equations (7.11) and (7.12) into (7.10), it is straightforward to observe that the equation (7.10) is satisfied. Hence, the HJB method implies that the controller given in (7.12) (after replacing (x_r, q_r) with $(x_r(t), q_r(t))$) is an optimal controller for the underlying system. ■

The utility functional given in Theorem 1 has several interesting features that will be spelled out next. Consider the price vector $\mathbf{p}(t)$ at a time instant $t \in [0, T]$. The best transmission rates that the users may utilize at this time can be obtained by maximizing the term $L(\mathbf{v}(t), \mathbf{p}(t))$ over all possible $\mathbf{v}(t)$'s. In other words, $\arg\max_{\mathbf{v}(t)} L(\mathbf{v}(t), \mathbf{p}(t))$ is indeed the optimal instantaneous transmission rates that the system can accept given its current link prices. As a result, the terminal utility $\max_{\mathbf{v}(T)} L(\mathbf{v}(T), \mathbf{p}(T))$ resembles the static Lagrangian at time T, but is maximized over all possible transmission rates to evaluate the potential of the system given its final price $\mathbf{p}(T)$. In other words, a variant of the static utility maximization problem is mainly integrated into the final utility (and partially incorporated into the integrand to take care of the transient behavior). On the other hand, the integrand has two terms $Y_l(y_l(t), p_l(t))$ and $Y_l(\tilde{y}_l(\mathbf{p}(t), p_l(t))$, each of which has a physical interpretation. The term $Y_l(y_l(t), p_l(t))$ can be regarded as the actual l^{th} link utility at time t, by virtue of the following observations:

- If $p_l(t)$ is nonzero, then $Y_l(y_l(t), p_l(t))$ is proportional to the quadratic term $-(y_l(t) - c_l)^2$, which implies that in order not to over-utilize or under-utilize the network, the best strategy is to maintain the flow rate $y_l(t)$ precisely at the capacity of the link.

- If $p_l(t)$ is zero, then $Y_l(y_l(t), p_l(t))$ indicates that the optimal utilization of the link corresponds to employing a flow rate below the link capacity.

Furthermore, $Y_l(\tilde{y}_l(\mathbf{p}(t), p_l(t))$ can be envisaged as the virtual l^{th} link utility at time t due to the fact that $\tilde{y}_l(\mathbf{p}(t))$ is the optimal transmission rate over the l^{th} link given the current price $\mathbf{p}(t)$. To summarize the ideas, the proposed utility functional is natural in the sense it maximizes the sum of the actual and virtual link utilities over the time interval $[0, T)$ and a variant of the static utility function at the final time T.

Corollary 1 *For every time instant $T > 0$, the following relation holds:*

$$\max_{\mathbf{x}(t)} \left\{ \frac{1}{2} \int_0^T \sum_{l \in \mathcal{L}} \big\{ Y_l(y_l(t), p_l(t)) + Y_l(\tilde{y}_l(\mathbf{p}(t)), p_l(t)) \big\} dt \right. \\ \left. - \max_{\mathbf{v}(T)} L(\mathbf{v}(T), \mathbf{p}(T)) \right\} = - \max_{\mathbf{v}(0)} L(\mathbf{v}(0), \mathbf{p}(0)). \tag{7.13}$$

Proof: It follows from the proof of Theorem 1 and the HJB equation that the expression given in the left side of the equality (7.13) is identical to the optimal cost-to-go $J(\mathbf{p}(0), 0)$. On the other hand, it is shown in the proof of Theorem 1 that $J(\mathbf{p}(0), 0)$ is equal to the right side of the above equation. This completes the proof. ■

Theorem 1 and corollary 1 assert that there exists a natural utility functional whose maximization leads to the celebrated dual TCP controller, and that the maximum value of this functional is equal to $-\max_{\mathbf{v}(0)} L(\mathbf{v}(0), \mathbf{p}(0))$. As pointed out earlier, this term corresponds to the maximum source utility at time $t = 0$ under the given initial price $\mathbf{p}(0)$.

Evidently, there are some utility functionals that trivially solve the inverse optimal problem under study. For instance, one candidate is

$$- \int_0^\infty \sum_{r \in \mathcal{S}} \left(x_r(t) - U_r'^{-1}(q_r(t)) \right)^2 dt. \tag{7.14}$$

Nevertheless, this utility functional has nothing to do with the static utility maximization problem, and provides no extra information about the system such as its closed-loop stability. In contrast, Theorem 1 proposes a meaningful utility functional, which is somewhat involved. A question arises as to whether there exists a simpler utility functional which still conveys meaningful interpretations. To answer this question, notice that the terminal utility given in (7.9) is a suitable counterpart of the original static utility function. Therefore, it remains to show that the integrand of this functional is essentially required and cannot be simplified. For this purpose, assume that the controller (7.7) maximizes the utility functional

$$\max_{\mathbf{x}(t)} \left\{ \int_0^T g(\mathbf{p}(t), \mathbf{x}(t)) dt - \max_{\mathbf{v}(T)} L(\mathbf{v}(T), \mathbf{p}(T)) \right\}$$

for the system (7.6), where T is a positive time and $g(\mathbf{p}(t), \mathbf{x}(t))$ is some function. Suppose also that $g(\mathbf{p}, \mathbf{x})$ is continuously differentiable at every point $(\mathbf{p}, \mathbf{x})$ for which $\mathbf{p}$ is strictly positive. Define the optimal cost-to-go function $J(\mathbf{p}, t)$ as

$$J(\mathbf{p}, t) := \int_t^T g(\tilde{\mathbf{p}}(s), \tilde{\mathbf{x}}(s))ds - \max_{\mathbf{v}(T)} L(\mathbf{v}(T), \tilde{\mathbf{p}}(T)), \tag{7.15}$$

where $\tilde{\mathbf{p}}(s)$ and $\tilde{\mathbf{x}}(s)$ denote the state and the input of the system (7.6) under the controller (7.7) in the case when the system starts at time t with the initial state $\mathbf{p}$. Finally, assume that $J(\mathbf{p}, t)$ is continuously differentiable with respect to $\mathbf{p}$ and t.

Theorem 2 *Under the assumptions made above, there exist a function $\hat{g}(\mathbf{p}(t), \mathbf{x}(t))$ and a real number μ such that*

$$g(\mathbf{p}(t), \mathbf{x}(t)) = \mu + \hat{g}(\mathbf{p}(t), \mathbf{x}(t)) + \frac{1}{2}\sum_{l \in \mathcal{L}} \big\{Y_l(y_l(t), p_l(t)) + Y_l(\tilde{y}_l(\mathbf{p}(t)), p_l(t))\big\}, \tag{7.16}$$

where the function $\hat{g}(\mathbf{p}(t), \mathbf{x}(t))$ is identically zero along all trajectories of the optimal closed-loop system.

Proof: In light of the assumptions made right before Theorem 2, one can write the HJB equation for this system as

$$0 = \frac{\partial J(\mathbf{p}, t)}{\partial t} + \max_{\mathbf{x}} \left\{ g(\mathbf{p}, \mathbf{x}) + \sum_{l \in \mathcal{L}} h_l(p_l)(y_l - c_l)^+_{p_l} \frac{\partial J(\mathbf{p}, t)}{\partial p_l} \right\}, \tag{7.17}$$

where $J(\mathbf{p}, t)$ is given in (7.15). Consider a strictly positive vector $\mathbf{p}$. Taking the derivative of the above expression with respect to x_r, $r \in \mathcal{S}$, yields

$$\sum_{l \in \mathcal{L}(r)} h_l(p_l) \frac{\partial J(\mathbf{p}, t)}{\partial p_l} = -\frac{\partial g(\mathbf{p}, \mathbf{x})}{\partial x_r}.$$

Since R has full row rank, the quantities $\frac{\partial J(\mathbf{p},t)}{\partial p_l}$, $l \in \mathcal{L}$, can be uniquely solved in terms of $\frac{\partial g(\mathbf{p},\mathbf{x})}{\partial x_r}$, $r \in \mathcal{S}$. This result, together with the memoryless property of the controller (7.7), implies that $\frac{\partial J(\mathbf{p},t)}{\partial p_l}$ does not depend on time. Hence, it follows from the HJB equation that $\frac{\partial J(\mathbf{p},t)}{\partial t}$ does not depend on time either. As a result, there exist a scalar μ and a function $f(\mathbf{p})$ such that

$$J(\mathbf{p}, t) = f(\mathbf{p}) - \mu t.$$

On the other hand, the boundary condition on the HJB equation states that

$$J(\mathbf{p}, T) = -\max_{\mathbf{v}} L(\mathbf{v}, \mathbf{p}).$$

Thus, one can conclude that

$$J(\mathbf{p}, t) = -\max_{\mathbf{v}} L(\mathbf{v}, \mathbf{p}) - \mu(t - T), \quad \forall\ \mathbf{p} > 0.$$

It follows from the continuity of $J(\mathbf{p}, t)$ that

$$J(\mathbf{p}, t) = -\max_{\mathbf{v}} L(\mathbf{v}, \mathbf{p}) - \mu(t - T), \quad \forall\ \mathbf{p} \geq 0.$$

Having written $g(\mathbf{p}, \mathbf{x})$ in the form of (7.16), substituting the above equation into the HJB equation yields that the function $\hat{g}(\mathbf{p}(t), \mathbf{x}(t))$ is equal to zero along all trajectories of the optimal closed-loop system. This completes the proof. ■

Notice that the term $\hat{g}(\mathbf{p}(t), \mathbf{x}(t))$ in Theorem 2 is a trivial term, which provides no useful information. This quantity can be, for instance, equal to the integrand of the trivial utility functional (7.14). Ignoring the uninformative terms μ and $\hat{g}(\mathbf{p}(t), \mathbf{x}(t))$, the functional given in Theorem 2 reduces to the one provided in Theorem 1.

It can be observed that the utility functionals characterized in Theorem 2 closely parallel those provided in (7.4) and (7.5) for a simple toy example. More specifically:

- $Y_l(y_l(t), p_l(t))$ corresponds to $-\frac{x(t)^2}{k}$. This term depends much more weakly on the state, but strongly on the input.
- $Y_l(\tilde{y}_l(t), p_l(t))$ corresponds to $-kp(t)^2$, which only penalizes the state.
- The constant term μ exists in both utility functionals.
- $\hat{g}(\mathbf{p}(t), \mathbf{x}(t))$ corresponds to $\hat{g}(p(t), x(t))$, which is an uninformative term and specifies no trade-off between the state and the input.

7.4.3 Stability Proof

An application of the optimal control problem introduced in Theorem 1 is that the global asymptotic stability of the system (7.6) under the static controller (7.7) can be concluded automatically.

Theorem 3 *The controller (7.7) that maximizes the utility functional (7.9) for the system (7.6) makes the pair* $(\mathbf{x}(t), \mathbf{p}(t))$ *converge to the fixed point* $(\mathbf{x}^*, \mathbf{p}^*)$.

Proof: The main idea behind the proof is to observe that

$$\begin{aligned} Y_l(y_l(t), p_l(t)) &\leq 0, \quad \forall t \in [0, T],\ l \in \mathcal{L}, \\ Y_l(\tilde{y}_l(\mathbf{p}(t)), p_l(t)) &\leq 0, \quad \forall t \in [0, T],\ l \in \mathcal{L}, \\ -\max_{\mathbf{v}(T)} L(\mathbf{v}(T), \mathbf{p}(T)) &\leq -\mathbf{p}^*, \end{aligned} \tag{7.18}$$

and that the state and input of the closed-loop control system satisfy the equation (by Corollary 1)

$$\begin{aligned} &\frac{1}{2}\int_0^T \sum_{l\in\mathcal{L}} \big\{Y_l(y_l(t), p_l(t)) + Y_l(\tilde{y}_l(\mathbf{p}(t)), p_l(t))\big\} dt \\ &-\max_{\mathbf{v}(T)} L(\mathbf{v}(T), \mathbf{p}(T)) = -\max_{\mathbf{v}(0)} L(\mathbf{v}(0), \mathbf{p}(0)). \end{aligned} \tag{7.19}$$

By letting T go to infinity, the relations (7.18) and (7.19) can be combined to conclude that

$$\begin{aligned} Y_l(y_l(t), p_l(t)) &\to 0 \quad \text{as} \quad t \to \infty, \\ Y_l(\tilde{y}_l(\mathbf{p}(t)), p_l(t)) &\to 0 \quad \text{as} \quad t \to \infty, \end{aligned} \tag{7.20}$$

for every $l \in \mathcal{L}$, in light of the fact that the left side of the equation (7.19) must remain finite and cannot go to $-\infty$ due to the finiteness of its right side. As a result

$$\lim_{t\to\infty} (y_l(t) - c_l)^+_{p_l(t)} = 0, \quad \forall\ l \in \mathcal{L}. \tag{7.21}$$

The proof follows immediately from the above equation. ■

7.4.4 Joint Routing and Congestion Control

It is desired to accomplish both routing and resource allocation simultaneously. For this purpose, assume that each source has a fixed origin and destination, but an undetermined route. The objective is to find an optimal route for every source so that the utility of the network is maximized. Note that since the Lagrangian introduced in (7.1) depends on the unknown routing matrix R, it will be denoted by $L(\mathbf{x}, \mathbf{p}; R)$ henceforth. As far as the optimal routing with respect to the static utility function is concerned, one should solve the optimization problem

$$\max_{R}\min_{\mathbf{p}}\max_{\mathbf{x}} L(\mathbf{x},\mathbf{p};R) = \max_{R} L(\mathbf{x}^*(R),\mathbf{p}^*(R);R) \tag{7.22}$$

to find R, where $(\mathbf{x}^*(R),\mathbf{p}^*(R))$ is the saddle point of the Lagrangian in the case when the routing matrix of the network is R. Solving the above optimization problem in a distributed way is formidable, because it is NP-hard even at the centralized level [111]. Aside from this point, a static utility function may not be a good measure for optimal routing, as the transient behavior of the system should also be taken into account. In what follows, the problem of optimal routing with respect to the utility functional (7.9) is addressed.

Theorem 4 *Given $r \in \mathcal{S}$, find all possible simple paths in the network which starts from the origin of source r and ends at the destination of this source. For each of these paths, compute the initial route price, i.e., the route price based on the link price vector $\mathbf{p}(0)$. Among these paths, each one with the minimum initial price is an optimal route for source r under the problem of joint routing and congestion control with respect to the utility functional (7.9).*

Proof: The joint routing and resource allocation with respect to the utility functional (7.9) amounts to solving the optimization problem

$$\max_{R}\min_{\mathbf{x}(t)}\bigg\{\frac{1}{2}\int_0^T \sum_{l\in\mathcal{L}}\big\{-Y_l(y_l(t),p_l(t)) \\ - Y_l(\tilde{y}_l(\mathbf{p}(t)),p_l(t))\big\}dt + \max_{\mathbf{v}(T)} L(\mathbf{v}(T),\mathbf{p}(T);R)\bigg\},$$

which is tantamount to (by Corollary 1)

$$\max_{R}\max_{\mathbf{v}(0)} L(\mathbf{v}(0),\mathbf{p}(0);R). \tag{7.23}$$

Denote the optimal routing matrix with R^*. Besides, define $q_r(t;R)$, $r \in \mathcal{S}$, as the route price associated with source r at time t under the routing matrix R. It is evident that

$$\max_{\mathbf{v}(0)} L(\mathbf{v}(0),\mathbf{p}(0);R) = \sum_{r\in\mathcal{S}}\Big\{U_r\Big(U_r'^{-1}(q_r(0;R))\Big) - q_r(0;R)U_r'^{-1}(q_r(0;R))\Big\} \\ + \sum_{l\in\mathcal{L}} c_l p_l(0).$$

Note that the terms $q_{r_1}(0; R)$ and $q_{r_2}(0; R)$ are independent of each other for every $r_1, r_2 \in \mathcal{S}$ such that $r_1 \neq r_2$, due to the fact that they are contingent upon different columns of R. Hence, in order to maximize the expression given in the above relation over all possible routing matrices R, the following optimization problem can be solved alternatively:

$$\max_{R} \left\{ U_r\Big(U_r'^{-1}(q_r(0; R))\Big) - q_r(0; R) U_r'^{-1}(q_r(0; R)) \right\}. \tag{7.24}$$

For a scalar variable q, one can write

$$\begin{aligned} \frac{\partial\Big(U_r\left(U_r'^{-1}(q)\right) - qU_r'^{-1}(q)\Big)}{\partial q} &= U_r'\left(U_r'^{-1}(q)\right)\frac{\partial U_r'^{-1}(q)}{\partial q} \\ &- U_r'^{-1}(q) - q\frac{\partial U_r'^{-1}(q)}{\partial q} = -U_r'^{-1}(q) \leq 0. \end{aligned}$$

This means that the function

$$U_r\left(U_r'^{-1}(q)\right) - qU_r'^{-1}(q)$$

is non-increasing in the variable q. As a result, it can be concluded from (7.24) that $q_r(0; R^*)$ is equal to the minimum of $q_r(0; R)$ over all possible routing matrices R. This completes the proof. ■

Consider a network over which both routing and congestion control are to be performed. If the users of network were fixed and remained online for a very long time, it could be justified that an optimal route should be obtained based on the optimal equilibrium point of the system. However, since users in a real network join and leave, and most of the sources do not live long, it is reasonable to take the transient behavior of the system into account for optimal routing. Under this circumstance, Theorem 4 proves that the optimal routing is really simple and intuitive: each user who joins the network should find a route to its destination whose initial price is minimum. Note that it is commonly accepted in the literature that routing could be performed by assigning a cost to each link and then minimizing the route cost. The present work shows that this simple idea indeed leads to an optimal route taking care of the transient response of the system.

7.4.5 Another Meaningful Utility Functional

Roughly speaking, the utility functional proposed in Theorem 1 treats a variant of the static utility function as the terminal utility and defines dynamical utility functions on the links. Another idea would be to define dynamical utility functions on the sources. This idea has been exploited in the next theorem.

Theorem 5 *Assume that the weighting functions $h_l(p_l(t))$, $l \in \mathcal{L}$ are all equal to 1. Given $T > 0$, the decentralized controller (7.7) maximizes the utility functional*

$$\max_{\mathbf{x}(t)} \left\{ \int_0^T \left(\sum_{r \in \mathcal{S}} U_r(x_r(t)) - \max_{\mathbf{v}(t)} L(\mathbf{v}(t), \mathbf{p}(t)) \right) dt - \frac{1}{2}\mathbf{p}(T)^T \mathbf{p}(T) \right\}$$

for the system (7.6). Furthermore, the maximum of this utility functional is equal to $-\frac{1}{2}\mathbf{p}(0)^T\mathbf{p}(0)$.

Proof: The proof can be carried out in line with that of Theorem 1 after noticing that the optimal cost-to-go function for this control problem is equal to $J(\mathbf{p}, t) = -\frac{1}{2}\mathbf{p}^T\mathbf{p}$. The details are omitted here for brevity. ■

The utility functional proposed in Theorem 5 has an interesting interpretation. The quantity $\max_{\mathbf{v}(t)} L(\mathbf{v}(t), \mathbf{p}(t))$ is equal to the maximum *instantaneous* source utility that the system can provide based on the price $\mathbf{p}(t)$. Hence, the integrand $\sum_{r \in \mathcal{S}} U_r(x_r(t)) - \max_{\mathbf{v}(t)} L(\mathbf{v}(t), \mathbf{p}(t))$ can be regarded as the *relative* source utility function. Having assumed $h_l(p_l(t))$ to be equal to 1, each price $p_l(t)$ can be visualized as the queue size at the buffer of the l^{th} router. Thus, the utility functional provided in the theorem aims to maximize the relative utility function over the time interval $[0, T)$ and minimize the sum of routers' queue sizes at the final time T. The maximum of the utility functional, which is equal to the negative half of the sum of the squared queue sizes at $t = 0$, is independent of the route. This property, together with the negative term inside the integrand, does not allow for deducing the stability of the dual control system from this functional for free, or searching for the optimal route.

7.5 Optimal Control for Primal Algorithm

Let the utility maximization problem stated in Section 7.2 be modified as

$$\max_{\mathbf{x}} \left\{ \sum_{r\in\mathcal{S}} U_r(x_r) - \sum_{l\in\mathcal{L}} \int_0^{y_l} f_l(y)dy \right\},$$

where $f_l(y)$ is a barrier function that can be interpreted as the price for transferring data at the rate y on link l. Assume that $f_l(\cdot)$, $l \in \mathcal{L}$, is a non-decreasing, continuous function such that

$$\int_0^{y_l} f_l(y)dy \to \infty \quad \text{as} \quad y_l \to \infty.$$

Furthermore, assume that $U_r(x_r)$, $r \in \mathcal{S}$, goes to $-\infty$ as x_r approaches zero. Under these assumptions, the above utility maximization problem has a unique solution $\mathbf{x}^*$ at which the gradient of $V(\mathbf{x})$ vanishes, where

$$V(\mathbf{x}) = \sum_{r\in\mathcal{S}} U_r(x_r) - \sum_{l\in\mathcal{L}} \int_0^{y_l} f_l(y)dy. \tag{7.25}$$

To obtain the solution $\mathbf{x}^*$ in a distributed way, consider the interconnected system given by

$$\dot{x}_r(t) = k_r(x_r(t))(U_r'(x_r(t)) - q_r(t)), \quad \forall\ r \in \mathcal{S} \tag{7.26}$$

and

$$p_l(t) = f_l(y_l(t)), \quad \forall\ l \in \mathcal{L}, \tag{7.27}$$

where $k_r : \Re \to \Re^+$ is a non-decreasing continuous function. It is known that the point $(\mathbf{x}^*, \mathbf{p}^*)$ is the globally asymptotically stable fixed point of this interconnected system [101]. Thus, the above distributed system can be run to asymptotically solve the static utility maximization problem. The objective is to find the optimal control counterpart of this result. For this purpose, assume that the memoryless system (7.27) exists in the core of the network to generate the link prices, and that each user deploys a simple integrator to adjust its transmission rate as

$$\dot{x}_r(t) = u_r(t), \quad r \in \mathcal{S}, \tag{7.28}$$

where $u_r(t)$ is some input signal that needs to be determined. It is noteworthy that $p_l(t)$ is a measured output of this system. The goal is to derive a utility functional for the system (7.28) whose maximization leads to the decentralized controller

$$u_r(t) = k_r(x_r(t)) \left(U_r'(x_r(t)) - q_r(t)\right), \quad r \in \mathcal{S}. \tag{7.29}$$

Theorem 6 *Given a time instant $T > 0$, the decentralized controller (7.29) maximizes the utility functional*

$$\max_{\mathbf{u}(t)} \Bigg\{ -\frac{1}{2} \int_0^T \Big(\mathbf{u}(t)^T \mathbf{K}(\mathbf{x}(t))^{-1} \mathbf{u}(t) \\ + \nabla V(\mathbf{x}(t))^T \mathbf{K}(\mathbf{x}(t)) \nabla V(\mathbf{x}(t)) \Big) dt + V(\mathbf{x}(T)) \Bigg\}$$

for the system given by (7.27) and (7.28), where

- $\mathbf{K}(\mathbf{x}(t))$ *is a diagonal matrix with the (r, r) diagonal entry $k_r(x_r(t))$ for all $r \in \mathcal{S}$.*
- $\mathbf{u}(t)$ *is a vector with the r^{th} entry $u_r(t)$ for all $r \in \mathcal{S}$.*
- *The symbol ∇ denotes the gradient operator.*

Moreover, the maximum of this utility functional is equal to $V(\mathbf{x}(0))$.

Proof: One can adopt the technique used in Theorem 1 to prove this theorem, after considering the optimal cost-to-go function $J(\mathbf{x}, t)$ as $V(\mathbf{x})$. ■

As before, the utility functional proposed in the above theorem has some plausible intrinsic properties. For instance, this functional treats the static utility function as a terminal utility, and encompasses two terms accounting for the transient behavior of the system. The term $\nabla V(\mathbf{x}(t))^T \mathbf{K}(\mathbf{x}(t)) \nabla V(\mathbf{x}(t))$ penalizes the nonzero gradient of the objective function $V(\mathbf{x}(t))$ during the transient time (note that the optimal solution of the static utility maximization problem corresponds to the unique point at which the gradient of $V(\mathbf{x})$ vanishes). Besides, the term $\mathbf{u}(t)^T \mathbf{K}(\mathbf{x}(t))^{-1} \mathbf{u}(t)$ or equivalently $\dot{\mathbf{x}}(t)^T \mathbf{K}(\mathbf{x}(t))^{-1} \dot{\mathbf{x}}(t)$ is a measure of users' willingness to alter their transmission rates abruptly. Thus, $\mathbf{K}(\mathbf{x})$ is a weighting function representing the trade-off between the above penalty terms.

In analogy with Theorem 3, the stability of the system (7.28) under the control (7.29) is an immediate consequence of Theorem 6. More precisely, since the integrand of the proposed utility functional is always less than or equal to zero and its terminal utility is bounded from above by $V(\mathbf{x}^*)$, letting T grow towards infinity yields

$$\nabla V(\mathbf{x}(t))^T \mathbf{K}(\mathbf{x}(t)) \nabla V(\mathbf{x}(t)) \to 0 \quad \text{as} \quad t \to \infty$$

or equivalently

$$||\nabla V(\mathbf{x}(t))|| \to 0 \quad \text{as} \quad t \to \infty.$$

It results from the above relation that the state of the closed-loop system converges to the unique maximizer of the function $V(\mathbf{x})$.

7.5.1 Joint Routing and Congestion Control

It is desired to perform joint routing and congestion control for the primal controller similar to what was carried out in Section 7.4.4 for the dual controller. To this end, since the utility function $V(\mathbf{x})$ depends on the routing matrix, it will be denoted by $V(\mathbf{x}; R)$ henceforth. An optimal routing matrix with respect to the utility function (7.25) can be obtained by solving the optimization problem

$$\max_R \max_{\mathbf{x}} V(\mathbf{x}; R), \tag{7.30}$$

which may be a cumbersome distributed optimization problem. In contrast, an optimal routing matrix with respect to the utility functional given in Theorem 6 can be found by solving

$$\max_R V(\mathbf{x}(0); R)$$

or equivalently

$$\min_R \sum_{l \in \mathcal{L}} \int_0^{R_l \mathbf{x}(0)} f_l(y) dy,$$

where R_l, $l \in \mathcal{L}$, denotes the l^{th} row of R. It is evident that the above optimization problem is far simpler than the one given in (7.30). For instance, if $f_l(y)$ is equal to y for every $l \in \mathcal{L}$, then an optimal routing matrix R can be obtained by solving the optimization problem $\min_R \|R\mathbf{x}(0)\|_2$, where $\| \cdot \|_2$ denotes the 2-norm operator. Recall that the joint routing and congestion control for the dual controller causes each source to take a minimum-price route, which may not be a proper strategy as several sources could take the same route and some possible routes may remain empty. In contrast, the joint routing and congestion control for the primal controller makes every effort that each link is not over-utilized at the initial time.

7.5.2 Congestion Control and Multi-Path Routing

Assume that there could exist multiple routes between each source-destination pair. Denote the source of route $r \in \mathcal{S}$ with $s(r)$. The utility maximization problem in this case can be regarded as the maximization of the utility function

$$V(\mathbf{x}) = \sum_{i \in \mathcal{S}} U_i \Bigg(\sum_{r:s(r)=i} x_r \Bigg) - \sum_{l \in \mathcal{L}} \int_0^{y_l} f_l(y) dy. \tag{7.31}$$

It is known that the maximization of the above function could give rise to more than one solution [101]. In the case when there exists a unique maximizer, the optimal transmission rates can be obtained asymptotically using the distributed controller

$$\dot{x}_r(t) = u_r(t), \quad \forall r \in \mathcal{S}, \tag{7.32}$$

where

$$u_r(t) = k_r \left(U'_{s(r)} \Bigg(\sum_{p:s(p)=s(r)} x_p(t) \Bigg) - q_r(t) \right), \tag{7.33}$$

and k_r is an arbitrary positive number. The question arises as to what meaningful utility functional the above primal controller maximizes. To answer this question, it can be shown that one such a functional is the utility functional given in Theorem 6 for single-path routing, but with $V(\mathbf{x})$ provided in (7.31) as opposed to the one in (7.25). This shows that the multi-path routing case is a simple extension of the single-path routing case. Now, different properties, such as stability, can be deduced as before.

7.6 Optimal Control for Primal/Dual Algorithm

Consider an interconnected system consisting of the subsystem (7.6) in the core of the network to generate prices and the subsystem (7.28) at the edge of the network to adjust the transmission rates. The states of this system are $\mathbf{x}(t)$ and $\mathbf{p}(t)$, while its input (to be found) is $\mathbf{u}(t)$. The goal of this part is to obtain a utility functional whose maximization yields the distributed controller (7.29). The techniques developed earlier can be exploited to tackle this problem. It can be shown that one such optimal control problem can be defined as

$$\max_{\mathbf{u}(t)}\bigg\{-\frac{1}{2}\int_0^T\bigg(\nabla_x L(\mathbf{x}(t),\mathbf{p}(t))^T\mathbf{K}(\mathbf{x}(t))\nabla_x L(\mathbf{x}(t),\mathbf{p}(t)) \\ +\mathbf{u}(t)^T\mathbf{K}(\mathbf{x}(t))^{-1}\mathbf{u}(t)-2\sum_{l\in\mathcal{L}}Y_l(y_l(t),p_l(t))\bigg)dt+L(\mathbf{x}(T),\mathbf{p}(T))\bigg\},$$

where ∇_x denotes the gradient operator with respect to the first argument $\mathbf{x}$. The integrand of this functional is the difference between those given for the primal and dual algorithms (if $\tilde{y}_l(\mathbf{p}(t))$ is identified by $y_l(t)$). However, physical intuition suggests that a good utility functional for this case should be the sum of those obtained for the dual and primal algorithms separately (as opposed to their difference). Indeed, the term $\sum_l Y_l(y_l(t),p_l(t))$ in the above utility functional is a measure of link utility (as pointed out earlier) that is minimized, instead of being maximized. This phenomenon can be justified by noticing that the static utility maximization problem is a min-max optimization (as performed on the Lagrangian), whereas the above utility functional is only a max optimization. It is worth noting that the utility functional proposed in Theorem 1 is also a min-max optimization. Obtaining a better utility functional for this case is left for future research.

7.7 Summary

This work relates the optimal control theory to the Internet congestion control algorithms. The main motivation for investigating this relationship is that the existing algorithms solve the utility maximization problem only at the equilibrium point and ignore the transient behavior of the control system. Therefore, they cannot be modified systematically to incorporate other physical constraints, such as a real-time link capacity requirement. In order to substantiate that the optimal control theory provides the right tools to solve a constrained network utility problem in practice, it is shown that there exist natural, meaningful utility functionals whose maximization yields the distributed controllers proposed by the primal, dual, and primal/dual algorithms. These utility functionals provide useful insights into the optimal closed-loop system; for instance, they automatically conclude the closed-loop stability for free.

Chapter 8

Conclusions and Future Work

This dissertation is concerned with the analysis and synthesis of large-scale complex systems arising in different areas of electrical and computer engineering. The high-level objective is to study how the physical properties of such systems can be deployed to simplify their design. The systems of interest in this work are categorized into three groups: (i) power networks, (ii) circuits and systems, and (iii) distributed computation. The results of this dissertation are presented in three parts, where each part studies one of these groups of systems. In what follows, the contributions made in each part are first summarized and possible future directions are then outlined.

8.0.1 Part I: Power Networks

In this part, the operation planning of power networks is investigated. To this end, the optimal power flow (OPF) problem is first considered. The OPF problem can be regarded as a fundamental optimization in power systems, which aims to find an optimal operating point for a power grid. This nonconvex problem is NP-hard in the worst case and has been extensively studied for 50 years. Part I of this dissertation is motivated by the fact that a practical OPF problem is highly structured and therefore its structure might make the problem solvable in polynomial time. To study a given OPF problem, a convex optimization problem is derived, which can solve the OPF problem globally in polynomial time if a certain condition is satisfied. It is shown that this condition holds not only for IEEE benchmark systems but also for a large class of power networks. The reason for the successful convexification of practical OPF problems can be traced back to the natural properties of transmission lines and transformers. To extend the applicability of this result, it is also shown that many energy-related optimization problems can be convexified similarly due to the same reason. The results of Part I are useful for studying several long-standing open problems. Some of these problems, which are left for future research, are as follows:

- What is the exact shape of the feasible set for an OPF problem?
- Given the fact that a power flow problem often has multiple solutions, how can all those solutions be identified?
- How can a true pricing mechanism be designed based on which consumers are charged and generators are paid in a fair and optimal way?
- In presence of integer decision variables, how can an energy-related optimization problem (e.g., unit commitment) be solved efficiently?
- To what extent can the results of Part I be generalized if the models of loads and (renewable-based) generators are uncertain and stochastic?

8.0.2 Part II: Circuits and Systems

Motivated by a variety of applications in circuits, electromagnetics, optics, and power networks, the first objective of Part II is to optimize certain parameters of a linear circuit in order to meet given design specifications. To this end, it is shown that this problem is NP-hard, even in a very particular case. However, the problem can be solved efficiently as long as the circuit is passive and there are enough number of unknown parameters to be optimized. This result introduces a trade-off between the design simplicity and the implementation complexity for an important class of linear circuits. As future work, it would be interesting to extend the synthesis result developed here to nonlinear circuits as well as circuits with both passive and active components.

As an application of the aforementioned circuit design technique, the problem of optimizing the controllable parameters of a passive smart antenna is studied. To be more precise, since the existing smart antennas are either hard to program or hard to implement, a new type of smart antenna is designed in polynomial type which can be implemented fairly easily. To show the efficacy of this result, a wavelength-size smart antenna is designed as an example, which uses only one active radiating element but can make nulls in many directions by tuning its passive elements. There are two problems along this research direction, which are worth studying:

- The smart antenna designed here was assumed to be in free space and therefore communication issues such as multipath were ignored. For practical reasons, it is important to study the antenna in a more realistic environment.

- In this work, it was assumed that the structure of the antenna was fixed, while some of its elements were tunable and could be optimized. A more interesting problem is to design the structure of the smart antenna as well.

8.0.3 Part III: Distributed Computation

This part deals with two important problems in the area of distributed computation: (i) quantized consensus, and (ii) distributed network resource allocation. In the quantized consensus problem, the goal is to find the average of a group of numbers in a distributed way over digital communication channels. Due to the finite capacity of each communication channel, the quantization effect comes into play and makes the problem hard. In this work, it is shown that quantized consensus is reached by means of a stochastic gossip algorithm recently proposed in the literature. The convergence time of this algorithm is also studied. As a future research, it would be interesting to study the quantized consensus problem for multi-agent systems, where the behavior of each agent is governed by some differential equations. An application of this problem is in the coordination of a group of systems.

The second problem studied in Part III is the network resource allocation problem, where the objective is to design a distributed algorithm by means of which every user of a communication network can find its optimal transmission rate for efficiently utilizing the network resources. From the mathematical standpoint, the existing resource allocation algorithms, such as primal and dual algorithms, aim to optimize the operation of the network only in the steady state. Since the real-time (transient) performance of the network is of a great importance for avoiding packet drop and reducing the transmission delay, this work shows how to exploit tools from optimal and inverse optimal control theories to design a congestion control (resource allocation) protocol with a guaranteed real-time performance. Possible future work would be to include routing into the design problem so that an optimal joint routing and congestion control problem with a guaranteed real-time performance is devised. Moreover, it is useful to extend this design methodology to wireless communication networks for which an optimal scheduling protocol should be found.

Bibliography

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, “A survey on sensor networks,” *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[2] A. Babakhani, D. B. Rutledge and A. Hajimiri, “Transmitter architectures based on near-field direct antenna modulation,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 12, pp. 2674–2692, 2008.

[3] A. Babakhani, D. B. Rutledge and A. Hajimiri, “Near-field direct antenna modulation, *IEEE Microwave Magazine*, vol. 10, no. 1, pp. 36–46, 2009.

[4] X. Bai, H. Wei, K. Fujisawa and Y. Wang, “Semidefinite programming for optimal power flow problems,” *International Journal of Electric Power & Energy Systems*, vol. 30, no. 6-7, pp. 383–392, 2008.

[5] J. Bao and P. L. Lee, *Process control: the passive systems approach*, Springer, 2007.

[6] T. Baumeister, “Literature review on smart grid cyber security,” *Technical Report*, University of Hawaii, 2010.

[7] F. Benezit, A. G. Dimakis, P. Thiran and M. Vetterli, “Gossip along the way: Order-optimal consensus through randomized path averaging,” in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, 2007.

[8] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, Belmont, MA: Athena Scientific, 1997.

[9] G. Boudour, A. Lecointre, P. Berthou, D. Dragomirecu and T. Gayraud, “On designing sensor networks with smart antennas,” *IFAC International Conference on Fieldbuses and Networks in Industrial and Embedded Systems*, Toulouse, France, 2007.

[10] S. Boyd, L. EI Ghaoui, E. Feron and V. Balakrishnan, "Linear matrix inequalities in system and control theory," *SIAM*, Philadelphia, PA, 1994.

[11] S. Boyd, A. Ghosh, B. Prabhakar and D. Shah, "Analysis and optimization of randomized gossip algorithms," in *Proceedings of the 43rd IEEE Conference on Decision and Control*, 2004.

[12] S. Boyd, A. Ghosh, B. Prabhakar and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[13] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[14] L. S. Bramko and L. L. Peterson, "TCP Vegas: end-to-end congestion avoidance on a global Internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1465–1480, 1995.

[15] O. Brune, "Synthesis of a finite two terminal network whose driving-point impedance is a prescribed function of frequency," *Journal of Mathematics and Physics*, vol. 10, pp. 191–236, 1931.

[16] K. C. Budka, J. G. Deshpande, T. L. Doumi, M. Madden and T. Mew, "Communication network architecture and design principles for smart grids," *Bell Labs Technical Journal*, vol. 15, no. 2, pp. 205–227, 2010.

[17] H. R. Cai, C. Y. Chung and K. P. Wong, "Application of differential evolution algorithm for transient stability constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 719–728, 2008.

[18] E. H. Callaway, *Wireless sensor networks: architectures and protocols*, CRC Press, 2003.

[19] F. Capitanescu, M. Glavic, D. Ernst and L. Wehenkel, "Contingency filtering techniques for preventive security-constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1690–1697, 2007.

[20] R. Carli, F. Fagnani, A. Speranzon and S. Zampieri, "Communication constraints in the average consensus problem," *Automatica*, vol. 44, no. 3, pp. 671–684, 2008.

[21] J. Carpentier, "Contribution to the economic dispatch problem," *Bulletin Society Francaise Electriciens*, vol. 3, no. 8, pp. 431–447, 1962.

[22] A. Censi and R. M. Murray, "A biologically inspired approach to real-valued average consensus over quantized channels with arbitrary deterministic accuracy," in *Proceedings of the 2009 American Control Conference*, 2009.

[23] M. Chiang, S. H. Low, A. R. Calderbank and J. C. Doyle, "Layering as optimization decomposition," *Proceedings of IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[24] E. De Klerk, "Aspects of semidefinite programming: interior point algorithms and selected applications," Applied Optimization Series, Vol. 65, *Kluwer Academic Publishers*, 2002.

[25] D. V. Dollen, "Report to NIST on the smart grid interoperability standards roadmap," *Electric Power Research Institute*, 2009.

[26] D. M. Falcao, F. F. Wu and L. Murphy, "Parallel and distributed state estimation," *IEEE Transactions on Power Systems*, vol. 10, no. 2, pp. 724–730, 1995.

[27] M. Fazel, H. Hindi and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the 2001 American Control Conference*, 2001.

[28] P. Frasca, R. Carli, F. Fagnani and S. Zampieri, "Average consensus by gossip algorithms with quantized communication," in *Proceedings of the 47th IEEE Conference on Decision and Control*, 2008.

[29] D. Gan, R. J. Thomas and R. D. Zimmerman, "Stability-constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 535–540, 2000.

[30] L. C. Godara, *Smart antennas*, CRC Press, Boca Raton, 2004.

[31] C. Godsil and G. Royle, *Algebraic graph theory*, volume 207 of Graduate Texts in Mathematics, Springer, 2001.

[32] T. F. González, *Handbook of approximation algorithms and metaheurististics*, Chapman & Hall/CRC Computer & Information Science Series, 2007.

[33] M. Grant and S. Boyd, “Matlab software for disciplined convex programming (web page and software),” http://stanford.edu/∼boyd/cvx, 2009.

[34] K. M. Grigoriadis and E. B. Beran, “Alternating projection algorithms for linear matrix inequalities problems with rank constraints,” *Advances in Linear Matrix Inequality Methods in Control, SIAM*, 2000.

[35] R. F. Harrington, *Field computation by moment methods*, Piscataway, NJ: IEEE Press, 1993.

[36] J. Harrison, “Formal synthesis of circuits with minimum noise figure using linear matrix inequalities,” *IEEE Transactions on Circuits and Systems*, vol. 54, no. 4, pp. 855–862, 2007.

[37] B. Hayes, *Group theory in the bedroom and other mathematical diversions*, Hill & Wang, 2008.

[38] H. G. Hoang, H. D. Tuan and B. N. Vo, “Low-dimensional SDP formulation for large antenna array synthesis,” *IEEE Transactions on Antennas and Propagation*, vol. 55, no. 6, pp. 1716–1725, 2007.

[39] X. Huang, J. Wang and Y. Fang, “Achieving maximum flow in interference-aware wireless sensor networks with smart antennas,” *Ad Hoc Networks* , vol. 5, no. 6, pp. 885–896, 2007.

[40] M. Huneault and F. D. Galiana, “A survey of the optimal power flow literature,” *IEEE Transactions on Power Systems*, vol. 6, no. 2, pp. 762–770, 1991.

[41] IE3D electromagnetic simulation and optimization software, Zeland Software Inc., www.zeland.com.

[42] J. Iguchi-Cartigny, P. M. Ruiz, D. Simplot-Ryl, I. Stojmenovic and C. M. Yago, “Localized minimum-energy broadcasting for wireless multihop networks with directional antennas,” *IEEE Transactions on Computers*, vol. 58, no. 1, pp. 120–131, 2009.

[43] R. A. Jabr, “Radial distribution load flow using conic programming,” *IEEE Transactions on Power Systems*, vol. 21, no. 3, pp. 1458–1459, 2006.

[44] R. A. Jabr, “Optimal power flow using an extended conic quadratic formulation,” *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1000–1008, 2008.

[45] R. A. Jabr, A. H. Coonick and B. J. Cory, “A primal-dual interior point method for optimal power flow dispatching,” *IEEE Transactions on Power Systems*, vol. 17, no. 3, pp. 654–662, 2002.

[46] V. Jacobson and M. J. Karels, “Congestion avoidance and control,” *ACM Computer Communication Review*, vol. 18, no. 4, pp. 341–329, 1988.

[47] A. Jadbabaie, J. Lin and A. S. Morse, “Coordination of groups of mobile autonomous agents using nearest neighbor rules,” *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[48] Q. Y. Jiang, H. D. Chiang, C. X. Guo and Y. J. Cao, “Power-current hybrid rectangular formulation for interior-point optimal power flow,” *IET Generation, Transmission & Distribution*, vol. 3, no. 8, pp. 748–756, 2009.

[49] Q. Jiang, G. Geng, C. Guo and Y. Cao, “An efficient implementation of automatic differentiation in interior point optimal power flow,” *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 147–155, 2010.

[50] J. M. Jin, *The finite element method in electromagnetics*, New York: Wiley, 1993.

[51] N. Jin and Y. Rahmat-Samii, “Parallel particle swarm optimization and finite-difference time-domain (PSO/FDTD) algorithm for multiband and wide-band patch antenna designs, *IEEE Transactions on Antenna and Propagation*, vol. 53, no. 11, pp. 3459-3468, 2005.

[52] A. Kashyap, T. Basar and R. Srikant, “Quantized consensus,” *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.

[53] F. P. Kelly, “Charging and rate control for elastic traffic,” *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[54] F. P. Kelly, A. Maullo and D. Tan, “Rate control in communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[55] D. E. Kirk, *Optimal control theory: an introduction*, Dover Publications, 2004.

[56] K. S. Kunz and R. J. Luebbers, *The finite difference method for electromagnetics*, CRC Press, London, 1993.

[57] Y. Kuramoto, *Chemical oscillators, waves, and turbulance*, Springer-Verlag, Berlin, 1984.

[58] J. Kuruvila, A. Nayak and I. Stojmenovic, "Progress and location based localized power aware routing for ad hoc and sensor wireless networks," *International Journal of Distributed Sensor Networks*, vol. 2, no. 2, pp. 147–159, 2006.

[59] K. Langendoen and N. Reijers, "Distributed localization in wireless sensor networks: a quantitative comparison," *Computer Networks*, vol. 43, no. 4, pp. 499–518, 2003.

[60] J. Lavaei, "Zero duality gap for classical OPF problem convexifies fundamental nonlinear power problems," in *Proceedings of the 2011 American Control Conference*, 2011.

[61] J. Lavaei, A. Babakhani, A. Hajimiri and J. C. Doyle, "Solving Large-Scale Linear Circuit Problems via Convex Optimization," in *Proceedings of the 48th IEEE Conference on Decision and Control*, 2009.

[62] J. Lavaei, A. Babakhani, A. Hajimiri and J. C. Doyle, "Programmable antenna design using convex optimization," in *19th International Symposium on Mathematical Theory of Networks and Systems*, 2010.

[63] J. Lavaei and S. H. Low, "Convexification of optimal power flow problem," in *Forty-Eighth Annual Allerton Conference*, 2010.

[64] J. Lavaei and S. H. Low, "Relationship between power loss and network topology in power systems," in *Proceedings of the 49th IEEE Conference on Decision and Control*, Atlanta, Georgia, 2010.

[65] O. H. Lerma and J. B Lasserre, *Markov chains and invariant probabilities*, Birkhäuser, 2003.

[66] J. C. Liberti and T. S. Rappaport, *Smart antennas for wireless communications: IS-95 and third generation CDMA applications*, Prentice Hall, 1999.

[67] W. M. Lin, C. H. Huang and T. S. Zhan, “A hybrid current-power optimal power flow technique,” *IEEE Transactions on Power Systems*, vol. 23, no. 1, pp. 177–185, 2008.

[68] J. Löfberg, “A toolbox for modeling and optimization in MATLAB,” in *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[69] N. A. Lynch, *Distributed algorithms*, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1996.

[70] P. McDaniel and S. McLaughlin, “Security and privacy challenges in the smart grid,” *IEEE Security Privacy Magazine*, vol. 7, no. 3, pp. 75–77, 2009.

[71] F. Milano, “An open source power system analysis toolbox,” *IEEE Transactions on Power Systems*, vol. 20, no. 3, pp. 1199–1206, 2005.

[72] J. A. Momoh, *Electric power system applications of optimization*, Markel Dekker, New York, USA, 2001.

[73] J. A. Momoh M. E. El-Hawary and R. Adapa, “A review of selected optimal power flow literature to 1993. Part I: Nonlinear and quadratic programming approaches,” *IEEE Transactions on Power Systems*, vol. 14, no. 1, pp. 96–104, 1999.

[74] J. A. Momoh M. E. El-Hawary and R. Adapa, “A review of selected optimal power flow literature to 1993. Part II: Newton, linear programming and interior point methods,” *IEEE Transactions on Power Systems*, vol. 14, no. 1, pp. 105–111, 1999.

[75] P. Moylan and B. Anderson, “Nonlinear regulator theory and an inverse optimal control problem,” *IEEE Transactions on Automatic Control*, vol. 18, no. 5, pp. 460–465, 1973.

[76] N. Nagai, *Linear circuits, systems, and signal processing: advanced theory and applications*, Marcel Dekker, 1990.

[77] K. S. Narendra and A. M. Annaswamy, *Stable adaptive systems*, Dover, 2005.

[78] B. K. Natarajan, “Sparse approximate solutions to linear systems,” *SIAM Journal of Computing*, vol. 24, no. 2, pp. 227–234 , 1995.

[79] T. Ohira and K. Gyoda, “Electronically steerable passive array radiator antennas for low-cost analog adaptive beamforming,” *IEEE International Conference on Phased Array Systems and Technology*, Dana Point, CA, 2000.

[80] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 401–420, 2006.

[81] R. Olfati-Saber, J. A. Fax and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[82] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[83] K. S. Pandya and S. K. Joshi, "A survey of optimal power flow methods," *Journal of Theoretical and Applied Information Technology*, vol. 4, no. 5, pp. 450–458, 2008.

[84] P. A. Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, PhD dissertation, California Institute of Technology, 2000.

[85] I. C. Paschalidis and D. Guo, "Robust and distributed localization in sensor networks," in *Proceedings of the 46th IEEE Conference on Decision and Control*, New Orleans, LA, 2007.

[86] I. C. Paschalidis, W. Lai and D. Starobinski, "Asymptotically optimal transmission policies for large-scale low-power wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 15, no. 1, pp. 105–118, 2007.

[87] L. L. Peterson and B. S. Davie, *Computer networks: a systems approach*, Morgan Kaufman, 1999.

[88] S. Prajna, A. Papachristodoulou, P. Seiler and P. A. Parrilo, *SOSTOOLS sum of squares optimization toolbox for MATLAB*, Users guide, 2004.

[89] Y. Rabani, A. Sinclair and R. Wanka,"Local divergence of Markov chains and the analysis of iterative load-balancing schemes," in *Proceedings of IEEE Conference on Foundations of Computer Science*, 1998.

[90] Y. Rahmat-Samii and E. Michielssen, *Electromagnetic optimization by genetic algorithms*, Wiley & Sons, New York, 1999.

[91] B. Recht, M. Fazel and P. A. Parrilo, "Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[92] B. Recht, W. Xu and B. Hassibi, "Null space conditions and thresholds for rank minimization," *Mathematical Programming*, vol. 127, pp. 175–211, 2011.

[93] J. Robinson and Y. Rahmat-Samii, "Particle swarm optimization in electromagnetics," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 2, pp. 397–407, 2004.

[94] W. Rugh, "On an inverse optimal control problem," *IEEE Transactions on Automatic Control*, vol. 16, no. 1, pp. 87–88, 1971.

[95] C. Santivanez and J. Redi, "On the use of directional antennas for sensor networks," *IEEE Military Communications Conference*, Boston, MA, 2003.

[96] A. V. Savkin, "Coordinated collective motion of groups of autonomous mobile robots: Analysis of Vicsek's model," *IEEE Transactions on Automatic Control*, vol. 49, no. 6, pp. 981–982, 2004.

[97] R. Schlub, J. Lu and T. Ohira, "Seven-element ground skirt monopole ESPAR antenna design from a genetic algorithm and the finite element method," *IEEE Transactions on Antennas and Propagation*, vol. 51, no. 11, pp. 3033–3039, 2003.

[98] S. Shakkottai and R. Srikant, "Network optimization and control," *Foundations and Trends in Networking*, vol. 2, no. 3, pp. 271–379, 2008.

[99] L. Shi, A. Capponi, K. H. Johansson and R. Murray, "Resource optimization in a wireless sensor network with guaranteed estimator performance," *IET Control Theory & Applications*, vol. 4, no. 5, pp. 710–723, 2010.

[100] A. Speranzon, C. Fischione and K.H. Johansson, "Distributed and collaborative estimation over wireless sensor networks," in *Proceedings of the 45th IEEE Conference on Decision and Control*, 2006.

[101] R. Srikant, *The mathematics of Internet congestion control*, Birkhauser, 2004.

[102] I. Stojmenovic, *Handbook of sensor networks: algorithms and architectures*, John Wiley & Sons, 2005.

[103] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268–276, 2001.

[104] K. Sundaresan and R. Sivakumar, "A unified MAC layer framework for ad-hoc networks with smart antennas," *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, pp. 546–559, 2007.

[105] G. Tel, *Introduction to distributed algorithms*, Cambridge University Press, 2000.

[106] G. L. Torres and V. H. Quintana, "Optimal power flow by a nonlinear complementarity method," *IEEE Transactions on Power Systems*, vol. 15, no. 3, pp. 1028–1033, 2000.

[107] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[108] J. N. Tsitsiklis, D. P. Bertsekas and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[109] University of Washington, Power Systems Test Case Archive, http://www.ee.washington.edu/research/pstca.

[110] L. Vandenberghe, S. Boyd and A. E. Gamal, "Optimizing dominant time constant in RC circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 2, pp. 110–125, 1998.

[111] J. Wang, L. Li, S. H. Low and J. C. Doyle, "Cross-layer optimization in TCP/IP networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 582–268, 2005.

[112] H. Wang, C. E. Murillo-Sanchez, R. D. Zimmerman and R. J. Thomas, "On computational issues of market-based optimal power flow," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1185–1193, 2007.

[113] H. Wei, H. Sasaki, J. Kubokawa and R. Yokoyama, "An interior point nonlinear programming for optimal power flow problems with a novel data structure," *IEEE Transactions on Power Systems*, vol. 13, no. 3, pp. 870–877, 1998.

[114] J. Weiqing, V. Vittal and G. T. Heydt, "A distributed state estimator utilizing synchronized phasor measurements," *IEEE Transactions on Power Systems*, vol. 22, no. 2, pp. 563–571, 2007.

[115] H. Wolkowitz, R. Saigal and L. Vandenberghe, editors, *Handbook of semidefinite programming: theory, algorithms and applications*, Kluwer, 2000.

[116] Y. Xia and K. W. Chan, “Dynamic constrained optimal power flow using semi-infinite programming,” *IEEE Transactions on Power Systems*, vol. 21, no. 3, pp. 1455–1457, 2006.

[117] K. Xie and Y. H. Song, “Dynamic optimal power flow by interior point methods,” *IEE Proceedings–Generation, Transmission and Distribution*, vol. 148, no. 1, pp. 76–84, 2002.

[118] R. D. Zimmerman, C. E. Murillo-Sánchez and R. J. Thomas, “MATPOWER’s extensible optimal power flow architecture,” *IEEE Power and Energy Society General Meeting*, 2009.

[119] Y. Zhou, Y. Fang and Y. Zhang, “Securing wireless sensor networks: a survey,” *IEEE Communications Surveys & Tutorials*, vol. 10, no. 3, pp. 6–28, 2008.

[120] A. E. Zooghby, *Smart antenna engineering*, Artech House: Norwood, 2005.