Optimization of Distribution Power Networks: From Single-Phase to Multi-Phase

Thesis by
Fengyu Zhou

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended May 18th, 2022

Fengyu Zhou
ORCID: 0000-0002-2639-6491

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisor, Professor Steven Low. Steven is not only a great researcher who guides me in my field of research, but also an incredible advisor who generously helps me whenever I need. He gives me the complete freedom to explore the problems I enjoy, and also shares with me the big picture to help me progress in the right direction. I have enjoyed all my discussions with Steven over the past six years. Whenever I ask him questions, he is always so patient and provides very insightful feedback, and whenever I get stuck on some hard problems, he keeps encouraging me and helps me overcome my difficulties. I feel so fortunate to have Steven as my PhD advisor.

I would also like to thank all the members on my thesis committee: Professors Adam Wierman, Venkat Chandrasekaran, John Doyle, and James Anderson. I received so much help from them through discussions, classes, candidacy feedback, and collaborations. They helped me find the right mathematical tools, spot the weakness in my research that I could refine, and improve my writing and presentation skills.

Next, I would like to thank all my collaborators within and beyond Caltech: James Anderson, Ahmed Zamzam, Nicholas Sidiropoulos, Chen Liang, Yue Chen, Changhong Zhao, and Guannan Qu. Their input and help significantly contributed to this thesis as well as all my other works. I also really enjoyed working and studying with all my good friends and colleagues in Netlab and RSRG: Daniel Guo, Zach Lee, Tongxin Li, Chen Liang, John Pang, Yuanyuan Shi, Yu Su, Yujie Tang, Zhaojian Wang, Lucien Werner, and Pengcheng You. Also, I feel extremely lucky that in two of my internship experiences, I have had the chance to work with two incredible Caltech alumni: Mario Munich and Qiuyu Peng (Netlab alumnus as well!).

It has also been a wonderful and unique journey to study in EE and CMS departments. I am grateful to know so many good friends beyond my direct group: Jin Sima, Riley Murrey, Eric Zhan, Hao Zhou, Natalie Bernat, Sara Beery, Jian Xu, and so many more than I can list here. I also want to thank the EE/CMS staff including Christine Ortega, Tanya Owen, and Daniel Yoder from ISP for making my journey at Caltech so wonderful.

Finally, I would like to thank my parents, Ms. Peiming Bao and Mr. Bin Zhou, for their continuing love and support since my childhood, and my girlfriend Yujia

Huang, for her constant companionship and encouragement during my most helpless time. My family is always the ultimate source of my inspiration and the strongest supporter for my research. This thesis belongs to all of you!

ABSTRACT

Distributed energy resources play an important role in today's distribution power system. The Optimal Power Flow (OPF) problem is fundamental in power systems as many important applications such as economic dispatch, battery displacement, unit commitment, and voltage control can be formulated as an OPF. A paradoxical observation is the problem's complexity in theory but simplicity in practice. On the one hand, the problem is well known to be non-convex and NP-hard, so it is likely that no simple algorithms can solve all problem instances efficiently. On the other hand, there are many known algorithms which perform extremely well in practice for both standard test cases and real-world systems. This thesis attempts to reconcile this seeming contradiction.

Specifically, this thesis focuses on two types of properties that may underlie the simplicity in practice of OPF problems. The first property is the exactness of relaxations, meaning that one can find a convex relaxation of the original non-convex problem such that the two problems share the same optimal solution. This property would allow us to convexify the non-convex problem without altering the optimal solution and cost. The second property is that all locally optimal solutions of the non-convex problem are also globally optimal. This property allows us to apply local algorithms such as gradient descent without being trapped at some spurious local optima. We focus on distribution systems with radial networks (i.e., the underlying graphs are trees). We consider both single-phase models and unbalanced multi-phase models, since most real-world distribution systems are multi-phase unbalanced, and distributed energy resources (DERs) can be connected in either wye or delta configurations.

The main results of this thesis are two-fold. In the first half, we propose a class of sufficient conditions for a non-convex problem to simultaneously have exact relaxation and no spurious local optima. Then we apply the result to single-phase system and conclude that if all buses have no injection lowerbounds, then both properties (exactness and global optimality) can be achieved. While the same condition is already known to be sufficient for exactness, our work is the first to extend it to global optimality. In the second half, we focus on the exactness property for multi-phase systems. For systems without delta connections, the exactness can be guaranteed if 1) the binding constraints are sparse in the network at optimality; or 2) all nodal prices fall within a narrow range. Using the DC model as an

approximation, we further analyze the OPF sensitivity and explain why nodal prices tend to be close to each other. In the presence of delta connections, we conclude that the inexactness can be resolved by either postprocessing an optimal solution, or adding a new regularization term in the cost function. Both methods achieve global optimality for IEEE standard test cases.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Fengyu Zhou and Steven Low. "Conditions for Exact Convex Relaxation and No Spurious Local Optima". In: *IEEE Transactions on Control of Network Systems* (2021). DOI: 10.1109/TCNS.2021.3112758.

[2] Fengyu Zhou et al. "Exactness of OPF relaxation on three-phase radial networks with delta connections". In: *IEEE Transactions on Smart Grid* 12.4 (2021), pp. 3232–3241. DOI: 10.1109/TSG.2021.3066530.

[3] James Anderson, Fengyu Zhou, and Steven H Low. "Worst-case sensitivity of DC optimal power flow problems". In: *2020 American Control Conference (ACC)*. IEEE. 2020, pp. 3156–3163. DOI: 10.23919/ACC45564.2020.9147770.

[4] Fengyu Zhou, James Anderson, and Steven H Low. "The optimal power flow operator: Theory and computation". In: *IEEE Transactions on Control of Network Systems* 8.2 (2020), pp. 1010–1022. DOI: 10.1109/TCNS.2020.3044258.

[5] Fengyu Zhou and Steven H Low. "A note on branch flow models with line shunts". In: *IEEE Transactions on Power Systems* 36.1 (2020), pp. 537–540. DOI: 10.1109/TPWRS.2020.3029732.

[6] Fengyu Zhou and Steven H Low. "A sufficient condition for local optima to be globally optimal". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 1684–1690. DOI: 10.1109/CDC42340.2020.9303868.

[7] Fengyu Zhou, James Anderson, and Steven H Low. "Differential privacy of aggregated DC optimal power flow data". In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 1307–1314. DOI: 10.23919/ACC.2019.8815257.

[8] Fengyu Zhou, Yue Chen, and Steven H Low. "Sufficient conditions for exact semi-definite relaxation of optimal power flow in unbalanced multiphase radial networks". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 6227–6233. DOI: 10.1109/CDC40024.2019.9029827.

[9] James Anderson, Fengyu Zhou, and Steven H Low. "Disaggregation for networked power systems". In: *2018 Power Systems Computation Conference (PSCC)*. IEEE. 2018, pp. 1–7. DOI: 10.23919/PSCC.2018.8442521.

F. Zhou contributed to the conception of the project, developed the underlying theory, designed and conducted the experiments, and participated in the writing of the manuscripts.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

The rapid development in distributed energy resources (DERs) over the past decades is now pushing the power grid through a historic transformation. Compared to traditional power generators, renewable energy resources such as wind turbines and solar panels supply clean and sustainable energy to the power grid. For example, during the COVID-19 pandemic, it was reported that renewables were the only energy source for which demand increased in 2020 despite the pandemic [39]. It is also projected that we will transition to having 100% renewable energy by 2050 [66]. On the demand side, the large-scale adoption of electrical vehicles (EVs) also has a significant impact, given that there are 10 million EVs worldwide by the end of 2020 and the industry is growing rapidly [78]. While the rapid growth of DERs brings numerous convenience to our daily life and improves global sustainability, there are also huge challenges for the control and optimization of distribution power networks. The first challenge is the impact of highly volatile distributed renewable energy resources on the optimization of power systems. Traditional power systems have a small number of large power plants with controllable generation [54]. As a result, distribution networks are typically maintained and operated in a more centralized way. However, with more and more renewable sources available, we can expect that more consumers will have the potential ability to supply the grid as well. Secondly, power is carried by three-phase transmission lines and while the grid operates traditionally in a *balanced mode* where the power in the three phases is equal in magnitude and equally spaced in phase angles, future distribution grids will not be balanced across the three phases because renewable generations fluctuate randomly and frequently, and DERs including electric vehicles are generally not phase-balanced. Most existing work on OPF assume the grid is balanced and their results cannot be directly applied to unbalanced three-phase systems.

This thesis focuses on a specific class of optimization problems for distribution networks, known as Optimal Power Flow (OPF) problems. An OPF problem is a mathematical program that computes optimal operating point of a power system, subject to power flow equations and operational constraints. The problem was first proposed in [21] and has been intensively studied since it has numerous applications in power systems, such as economical dispatch, battery displacement, unit commit-

ment, voltage control, and so on. We refer to [18] for a detailed review of the history of OPF problems. Specifically, we are interested in the computational properties of OPF problems. On the one hand, Alternating Current (AC) OPF problems are non-convex, and it has been proved in [11, 46] that they are NP-hard even for radial networks, and thus the problem in general can be very difficult in theory. On the other hand, the problem is also tractable in practice as researchers have found many algorithms that can yield globally optimal solutions for real-world test-cases [2, 41, 59]. In this thesis, we want to understand why practical OPF problems tend to have nice computational properties. We divide this research area into four quadrants as shown in Table 1.1. Network-wise, we consider both single-phase and multi-phase models of distribution networks. The macro structure of the network is assumed to be radial unless otherwise specified. Topic-wise, we focus on two types of questions: 1) is the convex relaxation exact? 2) are there any spurious locally optimal solutions?

Table 1.1: Four quadrants on OPF computational properties.

	Relaxation exactness	Global optimality
Single-phase network	Existing literature	This thesis
Multi-phase network	This thesis	To be explored

The first quadrant is on the relaxation exactness of OPF problems for single-phase networks. This topic has been extensively studied since the first proposal of semi-definite relaxation in [2]. In [45], it has been proven that the semi-definite relaxation has the same dual problem as the primal problem, and having exact relaxation is equivalent to the primal problem having zero duality gap. It also shows that zero duality gap can be guaranteed if the power marginal prices are non-negative. In subsequent works, many conditions that can be checked *a priori* have been discovered to guarantee exactness. For example, [15, 27] proposed the condition that all injections have no lower bound; the condition in [31] requires the problem to have no voltage upper-bounds. Work in [74] provides a geometric illustration of the injection region and shows that Pareto-optimal points of the injection region remains unchanged when taking the convex hull. Those works all apply to single-phase radial networks, or meshed networks with phase shifters. In [56], the authors extend partial results to a class of single-phase meshed network named weakly-cyclic graphs, and also lossless networks. More detailed surveys on this topic can be found in [52, 57]. Those works pioneered the research on the computational properties of

OPF problems and many of them also provide us with the building blocks to extend results to multi-phase networks.

The second quadrant studies whether there are spurious local optimal solutions of the non-convex OPF problems for single-phase networks. The idea is motivated by the observation that local algorithms such as Newton-Raphson or interior-point methods proposed in [41] often (though not always) yield the same solution as convex relaxation, indicating that local algorithms also tend to converge to globally optimal solutions. To the best of our knowledge, no existing work has studied this topic for OPF problems from a theoretical perspective. However, many similar non-convex problems in other areas such as machine learning and signal processing have been found to have no spurious local optima under certain conditions [1, 17, 33, 69, 70]. Some of those problems, such as matrix completion problem and low-rank semi-definite programs, may have exact relaxation and no spurious local optima at the same time under very similar conditions. We also have two interesting observations. First, traditional proof techniques on relaxation exactness and global optimality are very different though they sometimes can be applied to the same classes of problems. Second, most proof techniques in the literature analyze the optimization landscape through the gradient and curvature of the cost function and usually require the feasible set to have a simple structure, e.g., the spherical surface. In contrast, we develop a new perspective that can characterize problems that simultaneously have exact relaxation and no spurious local optima, and with more complicated feasible sets. Specifically, we consider an optimization problem with a convex cost function but a non-convex compact feasible set $\mathcal{X}$, and its relaxation with a compact and convex feasible set $\hat{\mathcal{X}} \supset \mathcal{X}$. We prove that if from any point $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$ there is a path connecting x to $\mathcal{X}$ along which both the cost function and a Lyapunov-like function are improvable, then any local optimum in $\mathcal{X}$ for the original non-convex problem is a global optimum. This result helps explain the widespread empirical experience that both local algorithms and convex relaxation for OPF problems often work extremely well.

As we move to the third quadrant, we try to extend the exactness results to multi-phase networks. Real-world distribution networks typically have more than one phase, and those phases are usually unbalanced at the operational point. Single-phase models can be poor approximations of such systems. Additionally, multi-phase networks may have both wye and delta connected devices. Previous studies on multi-phase power flow models can be found in [6, 9, 30, 80]. Specifically, the simulation results

in [76] show that if we relax the multi-phase OPF problem as a convex semi-definite program or second-order cone program, then the variables that are associated with wye connected devices are still exact. However, the variables that are associated with delta connected devices may no longer be exact. This result suggests that delta devices may require additional treatment in the problem formulation. In this thesis, we first study the multi-phase OPF problems without delta connections and provide sufficient conditions for exactness. Our result suggests that if the binding constraints are sparse, or if the nodal prices fall within a narrow range, then the problem tends to be exact. In the presence of delta connected devices, we find that the inexactness is due to under-specification of the relaxation formulation and numerical errors. Inspired by this finding, we then propose two algorithms which yield exact optimal solutions up to an acceptable numerical precision in simulations of IEEE 13-, 37-, and 123-bus systems.

The fourth quadrant is not explored in this thesis. However, we refer to the literature [8, 25]. These works apply primal-dual algorithms to multi-phase OPF problems and their applications to real systems perform very well in terms of both convergence and tracking capabilities. It will be interesting to analyze the performance of local algorithms for multi-phase distribution networks.

The rest of this thesis is organized as follows. In Chapter 2, we propose a condition that is sufficient for a non-convex optimization problem to have both exact relaxation and no spurious local optima. Then we apply the result to single-phase OPF problems and conclude that as long as there is no lower bounds for the power injections, all local optima should be globally optimal as well. In Chapter 3, we prove that a multi-phase OPF problem is exact if the corresponding dual matrix is $\mathcal{G}$-invertible. Then we consider two cases: 1) binding constraints are very sparse in the network, and 2) nodal prices fall within a narrow range. We prove that both cases guarantee the $\mathcal{G}$-invertibility of the dual matrix and hence the exactness of the primal solution. In Chapter 4, we investigate the multi-phase OPF problems with delta connections, and we conclude that the inexactness associated with delta devices can be resolved by either postprocessing the optimal solution, or penalizing a new regularization term in the cost function. Simulations on new algorithms achieve high accuracy for IEEE test cases. In Chapter 5, we revisit a question left in Chapter 3 and explain why nodal prices tend to fall within a narrow range, using DC model as an approximation. Further, we study the sensitivity of the OPF operator, and some results are also extended to classes of meshed networks.

Chapter 2

GLOBAL OPTIMALITY FOR SINGLE-PHASE NETWORKS

2.1 Background

Non-convex optimization problems in general are computationally challenging. However, many heuristics tend to work well for real-world problems. Those approaches include convex relaxations and local algorithms. It is usually hoped that relaxations yields exact solutions and local optima are also globally optimal. In this chapter, we derive the conditions, sufficient or necessary, for these two properties to hold simultaneously. Our focus is specifically on the optimization formulation with convex cost and non-convex constraints.

Related Works

Many problems have been proved to have exact relaxation and no spurious local optima (such as matrix completion [19, 20, 33] and low rank semidefinite program [5, 17, 61]); the proofs for those two properties are usually based on different types of certificates. In this subsection, we review some widely-used certificates for each property.

One type of certificates to exhibit relaxation exactness is by showing that any relaxed (and infeasible) point maps to a feasible solution with lower cost. This asserts that relaxed points cannot be the optimal solution. For instance, [27, 31] prove that optimal power flow problems can be solved via second-order cone relaxation under certain conditions, using the argument that any solution in the interior of the second-order cone can always be moved towards the boundary to further reduce the cost. In [5, 61], it is proved that if a semi-definite program has a solution with sufficiently large rank, then one can always reduce the rank without increasing the cost or violating the constraints. Another type of certificates involve studying the dual variables and KKT conditions. The underlying idea is a pair of primal and dual solutions satisfying KKT conditions that certify their optimality for both the primal and dual problems. Thus constructing dual variables with certain structures can also certify the optimality of primal solutions. In [19] for instance, the dual variable is related to the subgradient of the cost function at a desired matrix and therefore it helps certify the optimality of that desired matrix. Another example is [45], which proves the primal matrix should be of rank 1 through the argument that the null

space of its dual matrix has dimension at most 1. Similar techniques are also used in [42, 49, 53].

There are also considerable literature establishing the global optimality of local optima. We refer to [32, 70] and references therein. In [70], the authors focus on a class of problems with twice continuously differentiable function as the cost and Riemannian manifold as the feasible set. The values of Riemannian gradient and Hessian at a certain point then help certify properties such as *strong gradient*, *negative curvature* or *local convexity* in its neighborhood. It then eliminates spurious local optima and saddle points, where local algorithms can be trapped. This technique or idea were also used in [69] for dictionary recovery problem and [16] for phase synchronization. In both problems, the Riemannian manifold is some n-sphere or the Cartesian product of n-spheres. The framework summarized in [32] also leverages the landscape of the cost function, and the problem is usually reformulated into an unconstrained form. Instead of explicitly computing the gradient and Hessian matrix, the paper shows it suffices to find a single direction of improvement. For certain symmetric positive definite problems, the paper shows that the decision variable will always get closer to the global optimizer when the cost is reduced. A similar idea was also applied in [1], where the main result is built upon a correlation condition which states that the gradient (or any updating rule) is correlated with the direction from the current location towards the global optimizer. Therefore the underlying algorithm such as gradient descent can always produce a solution closer to the global optimizer as the algorithm progresses.

Contribution

The brief review above shows most works study exact relaxation and local optimality separately. It is unclear what might be the common feature of non-convex problems that possess both properties. Many real-world non-convex problems, however, seem to possess both properties, either provably or empirically, and it is hard to explain why these nice properties, though seemingly different, often occur simultaneously. Besides, most literature on local optimality focuses on problems without constraints or with tractable constraints. This is usually the case for problems in the learning area. However, for problems arising in cyber-physical systems, the constraints could include non-convex functions enforced by physical laws, as we will see in power systems. In these cases, either the feasible set is not a Riemannian manifold, or the Riemannian gradient and Hessian are very hard to derive. These questions motivate us to study conditions, sufficient or necessary, for problems to simultaneously have

exact relaxation and no spurious local optima. These conditions also help us study local optimality using properties of its relaxation, instead of its landscape.

Our conditions have two parts. The first part is on the sufficient condition. Roughly, if for any relaxed point, there exists a path connecting it to the non-convex feasible set and the path satisfies the following:

- along the path the cost is non-increasing,
- along the path the 'distance' to the non-convex feasible set is non-increasing,

then the problem must have exact relaxation and no spurious local optima simultaneously. Here the 'distance' can be any properly constructed function, as we will define later as a Lyapunov-like function (Definition 10). The second part is on the necessary condition, which says that if a problem does have exact relaxation and no spurious local optima simultaneously, then there must exist such a Lyapunov-like function and paths satisfying the requirements above. [1]

Though Lyapunov-like functions and paths are guaranteed to exist, for specific problems it could still be difficult to construct them. We then derive certain rules to construct a Lyapunov-like function and paths of a new problem from primitive problems with known Lyapunov-like functions and paths. This process allows us to reuse and extend known results as the problem changes and grows. Finally, we apply the proposed approach to two specific problems, optimal power flow (OPF) and low rank SDP. Our work proves the first known condition (that can be checked *a priori*) for OPF to have no spurious local optima, and it helps explain the widespread empirical experience that local algorithms for OPF problems often work extremely well.

Background for Power Systems

As one of the applications and main motivation of this work, OPF is a core problem in power systems. First proposed in [21], OPF is a class of optimization problems that minimizes a certain cost subject to nonlinear physical laws and operational constraints. It is known to be non-convex and NP-hard in its AC formulation [45, 46, 72]. Therefore, there is no known efficient algorithm that can solve all problem instances in polynomial time. Traditional approaches to solving OPF are usually

[1] The necessary condition is based upon some stronger assumptions so the second part is not the exact converse of the first part.

based on local algorithms such as Newton-Raphson, see [41, 59, 60] for examples. Over the past decade, techniques on convex relaxation have also been introduced to solve OPF [2, 40]. A surprising empirical finding in the literature shows that despite the non-convexity, both local algorithms and convex relaxations very often yield global optimum of the original non-convex problem [2, 15, 40, 45]. In recent years, there have been considerable analytical works on provable conditions for the relaxation exactness, which are summarized in the reviews [51, 57] and references therein. However, few analytical results are known on the performance guarantee of local algorithms. In this chapter, we show that a known sufficient condition for exact relaxation is also sufficient for local optima to be globally optimal. To the best of our knowledge, this is the first analytical result of its kind, and we hope that the approaches developed in this chapter can help derive more sufficient conditions along this direction.

2.2 Preliminaries

In this chapter, we will use $\mathbb{K}$ to denote the set $\mathbb{R}$ of real numbers or the set $\mathbb{C}$ of complex numbers. For any finite positive integer n, $\mathbb{K}^n$ is a Banach space.

Consider a (potentially non-convex) optimization problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) && \text{(2.1a)} \\ \text{s.t.} \quad & x \in \mathcal{X} && \text{(2.1b)} \end{aligned}$$

and its convex relaxation

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) && \text{(2.2a)} \\ \text{s.t.} \quad & x \in \hat{\mathcal{X}}. && \text{(2.2b)} \end{aligned}$$

Here $\mathcal{X}$ is a nonempty compact subset of $\mathbb{K}^n$, not necessarily convex, while $\hat{\mathcal{X}} \subseteq \mathbb{K}^n$ is an arbitrary compact and convex superset of $\mathcal{X}$. The cost function $f : \hat{\mathcal{X}} \to \mathbb{R}$ is convex and continuous over $\hat{\mathcal{X}}$. We do not require the relaxation $\hat{\mathcal{X}}$ to be efficiently represented.

Definition 1. *A point $x^{\text{lo}} \in \mathcal{X}$ is called a* local optimum *of* (2.1) *if there exists a $\delta > 0$ such that $f(x^{\text{lo}}) \leq f(x)$ for all $x \in \mathcal{X}$ with $\|x - x^{\text{lo}}\| < \delta$.*

Definition 2 (Strong Exactness)**.** *We say the relaxation* (2.2) *is* exact *with respect to* (2.1) *if every optimal point of* (2.2) *is feasible, and hence globally optimal, for* (2.1).

Unless otherwise specified, we will always use the term *exact* to refer to such strong exactness. Definition 2 implies in particular that, if (2.2) is exact, then $\forall \hat{x} \in \hat{\mathcal{X}} \setminus \mathcal{X}$, $f(\hat{x}) > \min_{x \in \hat{\mathcal{X}}} f(x)$.

Definition 3. *A* path *in $\mathcal{S} \subseteq \mathbb{K}^n$ connecting point a to point b is a continuous function $h : [0, 1] \to \mathcal{S}$ such that $h(0) = a$ and $h(1) = b$.*

We may refer to a path as the corresponding function h in the remainder of the chapter.

Lemma 1. *The following are equivalent:*

(A) Problem (2.2) *is exact with respect to* (2.1).

(B) For any $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, there is a path h in $\hat{\mathcal{X}}$ such that $h(0) = x, h(1) \in \mathcal{X}$, $f(h(t))$ is non-increasing for $t \in [0, 1]$ and $f(h(0)) > f(h(1))$.

Proof. (A) $\implies$ (B): Let x^* be any optimal point of (2.2). By (A), $x^* \in \mathcal{X}$, thus for $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, we could choose the path as the line segment from x to x^* since $\hat{\mathcal{X}}$ is convex.

(B) $\implies$ (A): Condition (B) implies that no point $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$ can be optimal for (2.2). □

Lemma 1 is not surprising, and in fact many works in the literature proving exact relaxations of Optimal Power Flow problems can be interpreted as using (B) to prove (A) by implicitly finding such a path h for each $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$ [51].

Condition (B) does not say anything about the local optima in $\mathcal{X}$ for (2.1). In the next section we will strengthen (B) by equipping the path with a Lyapunov-like function and show that the stronger condition implies that all local optima of (2.1) are globally optimal. We start by classifying local minima.

Definition 4. *We classify each local optimum x^{lo} of* (2.1) *into three disjoint classes: x^{lo} is a*

- *Global optimum (g.o.) if $f(x^{\text{lo}}) \leq f(x)$ for all the feasible $x \in \mathcal{X}$.*
- *Pseudo local optimum (p.l.o.) if there is a path $h : [0, 1] \to \mathcal{X}$ such that $h(0) = x^{\text{lo}}$, $f(h(t)) \equiv f(x^{\text{lo}})$ for all $t \in [0, 1]$ and $h(1)$ is not a local optimum.*

- *Genuine local optimum (g.l.o.) if it is neither a global optimum nor a pseudo local optimum.*

Examples of all three classes are shown in Fig. 2.1.

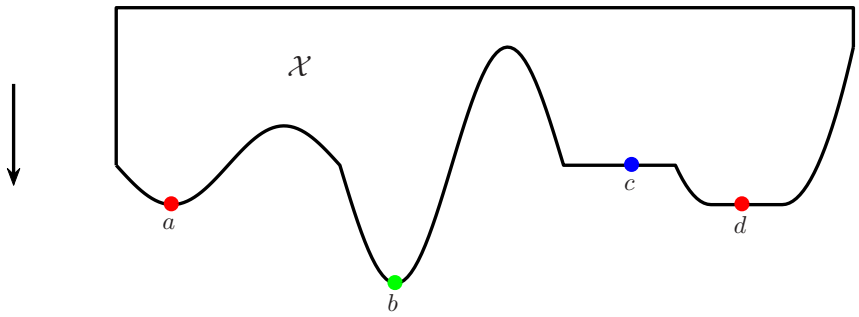

Figure 2.1: Examples for three classes of local optima. The arrow indicates the direction along which the cost function decreases. Point b is a global optimum point c is a pseudo local optimum, while points a and d are genuine local optima.

Definition 5. *A point x is improvable in $\mathcal{X}$ if there is a path $h : [0,1] \to \mathcal{X}$ such that*

- $h(0) = x$;
- $f(h(t))$ *is non-increasing for* $t \in [0,1]$;
- $h(1)$ *is not a local optimum or* $f(h(1)) < f(x)$.

Remark 1. *A local optimum is a pseudo local optimum if and only if it is improvable in $\mathcal{X}$.*

Definition 6. *A set $\{h_i : i \in \mathcal{I}\}$ of paths indexed by i is said to be* uniformly bounded *if there is a finite number M such that $\|h_i(t)\|_\infty \leq M$ for every $i \in \mathcal{I}$ and $t \in [0,1]$.*

Definition 7. *A set $\{h_i : i \in \mathcal{I}\}$ of paths indexed by i is said to be* uniformly equicontinuous *if for any $\epsilon > 0$, there exists a $\delta > 0$ such that $\|h_i(t_1) - h_i(t_2)\|_\infty < \epsilon$ for every $i \in \mathcal{I}$ whenever $|t_1 - t_2| < \delta$.*

Remark 2. *The index set $\mathcal{I}$ could be empty or uncountably infinite. An empty path set (i.e., when $\mathcal{I} = \emptyset$) is considered to be both uniformly bounded and uniformly equicontinuous.*

Let $\Pi|_a^b$ be the family of all the finite ordered subsets of $[a, b]$. We use Π as a shorthand for $\Pi|_0^1$. For $\pi = (t_0, \cdots, t_N) \in \Pi$ and a path h, define

$$L_\pi(h) := \sum_{i=1}^{N} \|h(t_{i-1}) - h(t_i)\|_{\ell_2}.$$

Clearly, $L_\pi(h)$ is always finite for given π and h.

Definition 8 ([71])**.** *For path h, define the function $L(h) := \sup_{\pi\in\Pi} L_\pi(h)$. We say h is rectifiable iff $L(h)$ is finite. When h is rectifiable, $L(h)$ is also referred to as its length.*

Definition 9 ([71])**.** *For a rectifiable path $h : [0, 1] \to \mathbb{K}^n$, let its arc-length reparameterization be $\bar{h} : [0, 1] \to \mathbb{K}^n$ and*

$$\begin{cases} \bar{h}\left(\frac{1}{L(h)} \sup_{\pi\in\Pi|_0^t} L_\pi(h)\right) := h(t), & \text{if } L(h) > 0 \\ \bar{h} := h, & \text{if } L(h) = 0 \end{cases}$$

One could see $L(\bar{h}) = L(h) < \infty$ and they have the same function image, i.e., $\{\bar{h}(t)|t \in [0, 1]\} = \{h(t)|t \in [0, 1]\}$. For $0 \le t_1 \le t_2 \le 1$, $\bar{h}$ has the property that $\sup_{\pi\in\Pi|_{t_1}^{t_2}} L_\pi(\bar{h}) = (t_2 - t_1)L(\bar{h})$.

Lemma 2. *For a set of rectifiable paths $h_i, i \in \mathcal{I}$, if the values of $L(h_i)$ are uniformly bounded, then the set of $\bar{h}_i, i \in \mathcal{I}$ is uniformly equicontinuous.*

Proof. Assume $L(h_i) \le M$ for all $i \in \mathcal{I}$, then for any $0 \le t_1 \le t_2 \le 1$, we have for any i,

$$\begin{aligned} &\|\bar{h}_i(t_1) - \bar{h}_i(t_2)\|_\infty \le \|\bar{h}_i(t_1) - \bar{h}_i(t_2)\|_{\ell_2} \\ \le &\sup_{\pi\in\Pi|_{t_1}^{t_2}} L_\pi(\bar{h}_i) = (t_2 - t_1)L(h_i) \le M|t_1 - t_2|. \end{aligned}$$

Setting $\delta = \epsilon/M$, the equicontinuity is proved. □

Corollary 1. *If $\mathcal{S}$ is compact in $\mathbb{K}^n$ and all paths in a set $\mathcal{H} = \{h_i : i \in \mathcal{I}\}$ are $[0, 1] \to \mathcal{S}$ and consist of at most N linear segments, then $\{\overline{h_i} : i \in \mathcal{I}\}$ must be both uniformly bounded and uniformly equicontinuous. Here N is a finite constant for all paths in $\mathcal{H}$.*

2.3 Sufficient Conditions

In this section, we first study the sufficient conditions under which (2.2) is exact w.r.t. (2.1) and all the local optima of (2.1) are also globally optimal. Those sufficient conditions will be proposed by strengthening Condition (B). Note that (B) has already implied (2.2) is exact w.r.t. (2.1), so our strategy is to strengthen (B) in order to rule out the possibility of genuine local optima and pseudo local optima.

Ruling Out Genuine Local Optima

Definition 10. *A* Lyapunov-like function [2] *associated with* (2.1) *and* (2.2) *is a continuous function* $V : \hat{\mathcal{X}} \to \mathbb{R}^+$ *such that* $V(x) = 0$ *for* $x \in \mathcal{X}$ *and* $V(x) > 0$ *for* $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$.

A strengthened version of (B) is as follows.

(C) There exists a Lyapunov-like function V associated with (2.1) and (2.2) such that:

(C1) For any $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, there is a path h_x in $\hat{\mathcal{X}}$ such that $h_x(0) = x$, $h_x(1) \in \mathcal{X}$, both $f(h_x(t))$ and $V(h_x(t))$ are non-increasing for $t \in [0, 1]$ and $f(h_x(0)) > f(h_x(1))$.

(C2) The set $\{h_x\}_{x \in \hat{\mathcal{X}} \setminus \mathcal{X}}$ is uniformly bounded and uniformly equicontinuous.

Theorem 1. *If (C) holds, then (A) also holds and any local optimum in* $\mathcal{X}$ *for* (2.1) *is either a global optimum or a pseudo local optimum.*

Proof. (C) $\implies$ (A) is because (C) is stronger than (B). As for the second part of the argument, we include an illustrative sketch of the notations in Fig. 2.2. Suppose $x \in \mathcal{X}$ is a local but not global optimum for (2.1). We will prove that x must be improvable in $\mathcal{X}$ (and thus a pseudo local optimum).

Let $x^* \neq x$ be a global optimum of (2.1), so $f(x^*) < f(x)$. Let $\ell : [0, 1] \to \hat{\mathcal{X}}$ be the linear function characterizing the line segment from x to x^*, i.e., $\ell(t) = (1-t)x + tx^*$ with $f(\ell(1)) = f(x^*) < f(x)$. Note that $f(\ell(t))$ is non-increasing in t. To see this, consider any $t \geq 0, \epsilon > 0$ with $t + \epsilon \leq 1$, $x_1 = \ell(t)$, $x_2 = \ell(t + \epsilon)$. Setting

[2] In contrast to a standard Lyapunov function, we do not require V to be differentiable here.

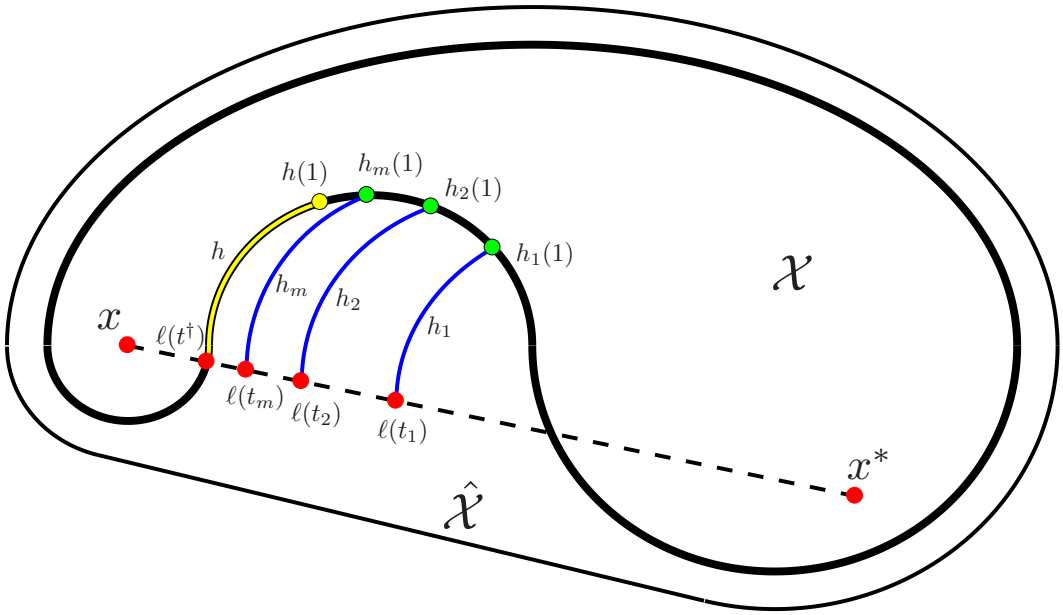

Figure 2.2: Sketch of notations for the proof of Theorem 1. Point x and $\ell(t^\dagger)$ will be later proved to be identical to each other.

$s := \epsilon/(1-t)$, we have $x_2 = (1-s)x_1 + sx^*$. Since f is convex and x^* is also a global optimum of (2.2) over $\hat{\mathcal{X}}$, we have

$$f(x_2) \;\le\; (1-s)f(x_1) + sf(x^*) \;\le\; f(x_1).$$

Define

$$t^\dagger := \sup_{t\in[0,1]} \; t \;\text{ s. t. }\ell(\tau) \in \mathcal{X} \;\; \forall \tau \le t.$$

As $\mathcal{X}$ is closed, $\ell(t^\dagger)$ is also in $\mathcal{X}$. We first prove $\ell(t^\dagger)$ must be x (i.e., $t^\dagger = 0$). Otherwise, as x is a local optimum, we could find $\delta \in (0, t^\dagger)$ such that $f(\ell(t)) \ge f(\ell(0)) = f(x)$ for all $t \in [0, \delta)$. Since $f(\ell(t))$ is non-increasing in t, we must have $f(\ell(t)) \equiv f(\ell(0)) = f(x)$ for all $t \in [0, \delta)$. It contradicts the fact that $f(\ell(t))$ is convex and $f(\ell(1)) = f(x^*) < f(x) = f(\ell(0))$ for the same reason f is non-increasing in t.

Therefore $\ell(t^\dagger) = x$ and $f(\ell(t^\dagger)) = f(x)$. It is sufficient to show $\ell(t^\dagger)$ is improvable in $\mathcal{X}$. That is to say, it is sufficient to find some function $h : [0,1] \to \mathcal{X}$ such that $h(0) = \ell(t^\dagger)$, $f(h(t))$ is non-increasing in $t \in [0,1]$, and either $f(h(1)) < f(\ell(t^\dagger))$ or $h(1)$ is not a local optimum in $\mathcal{X}$ for (2.1).

By the definition of $t^\dagger$, there is a decreasing sequence $t_m \to t^\dagger$ such that $t_m \in (t^\dagger, 1]$ and $\ell(t_m) \in \hat{\mathcal{X}} \setminus \mathcal{X}$ for all m. Since $f(\ell(t))$ is non-increasing in t, the sequence

$f(\ell(t_m))$ is non-decreasing in m and $f(\ell(t_m)) < f(\ell(t^\dagger))$. [3] For each $\ell(t_m)$ we take the function $h_m : [0,1] \to \hat{\mathcal{X}}$ guaranteed by Condition (C). As the sequence h_m is uniformly bounded and uniformly equicontinuous, a subsequence must uniformly converge to a limit h by Arzelà-Ascoli theorem. Without loss of generality, we denote this subsequence as h_m as well. Next we prove this h satisfies all the properties in Definition 5, implying the improvability of x.

To show $h(t) \in \mathcal{X}$ for any fixed $t \in [0,1]$, we consider the sequence $(V(h_m(t)) : m \in \mathbb{Z})$. As $\hat{\mathcal{X}}$ is closed, we have $h(t) = \lim_{m\to\infty} h_m(t) \in \hat{\mathcal{X}}$. Further consider V is continuous and $V(h_m(t)) \le V(h_m(0))$, thus

$$\begin{aligned} 0 \le & V(h(t)) = V(\lim_{m\to\infty} h_m(t)) = \lim_{m\to\infty} V(h_m(t)) \\ & \le \lim_{m\to\infty} V(h_m(0)) = \lim_{m\to\infty} V(\ell(t_m)) = V(\ell(t^\dagger)) = 0. \end{aligned}$$

Hence $V(h(t)) = 0$ and $h(t) \in \mathcal{X}$.

To show $h(0) = \ell(t^\dagger)$, we consider

$$h(0) = \lim_{m\to\infty} h_m(0) = \lim_{m\to\infty} \ell(t_m) = \ell(t^\dagger).$$

To show $f(h(t))$ is non-increasing, we take any $s, t \in [0,1]$ such that $s < t$. As f is continuous, we have

$$\begin{aligned} f(h(s)) &= \lim_{m\to\infty} f(h_m(s)) \\ f(h(t)) &= \lim_{m\to\infty} f(h_m(t)) \end{aligned}$$

and by Condition (C) we have $f(h_m(s)) \ge f(h_m(t))$ for each m. Therefore $f(h(s)) \ge f(h(t))$.

Finally, we will show if $f(h(1)) = f(\ell(t^\dagger))$ then $h(1)$ must not be a local minimal in $\mathcal{X}$ for (2.1). For each m,

$$f(h_m(1)) \le f(h_m(0)) = f(\ell(t_m)) < f(\ell(t^\dagger)) = f(h(1))$$

and $h_m(1) \in \mathcal{X}$. Since the sequence $h_m(1)$ converges to $h(1)$ as $m \to \infty$, within any open neighborhood of $h(1)$ in $\mathcal{X}$, we could always find some $h_m(1)$ with strictly smaller cost value. Thus $h(1)$ cannot be a local minimum in $\mathcal{X}$. □

[3] The strict inequality is due to the convexity of $f(\ell(t))$ and the fact that $f(\ell(1)) < f(\ell(t^\dagger))$.

Ruling Out Pseudo Local Optima

So far, Condition (C) has eliminated the possibility of having genuine local optima, and in this subsection we further strengthen the condition to also rule out pseudo local optima.

Consider the following lemma and its corollaries.

Lemma 3. *If* (2.1) *is exact with respect to* (2.2) *and* (2.1) *has no genuine local optima, then the feasible set of* (2.1) *is connected.*

Proof. If $\mathcal{X}$ is not connected, then by definition $\mathcal{X}$ can be partitioned into two disjoint non-empty closed sets $\mathcal{X}_1$ and $\mathcal{X}_2$ with $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, which are hence both compact. Further we let x_i be any global optimum of $\min_{x \in \mathcal{X}_i} f(x)$ for $i = 1, 2$. Clearly $x_1 \neq x_2$ and they are both local optima of (2.1).

If $f(x_1) = f(x_2)$, then any convex combination of x_1, x_2 must be a global optimum to (2.2). Since there is no path in $\mathcal{X}$ that connects x_1 and x_2, there must be some convex combination that is outside $\mathcal{X}$. This contradicts the exactness of relaxation.

If $f(x_1) \neq f(x_2)$, without loss of generality we assume $f(x_1) < f(x_2)$, i.e., x_2 is not a global optimum of (2.1). But x_2 is not a pseudo local optimum of (2.1) either, contradicting Theorem 1. To see this, note that any point $x' \in \mathcal{X}$ which is connected to x_2 via a path in $\mathcal{X}$ must also be a point in $\mathcal{X}_2$ and if $f(x') = f(x_2)$ then x' must be a local optimum of (2.1) as well. □

Corollary 2. *Condition (C) implies that the feasible set of* (2.1) *is connected.*

Now we are in a good position to discuss some conditions that rule out pseudo local optima and therefore guarantee that any local optimum must be a global optimum.

Corollary 3. *If all local optima of* (2.1) *are isolated, then Condition (C) implies that any local optimum of* (2.1) *is a global optimum.*

Here, local optima being isolated means any local optimum of (2.1) has an open neighborhood which contains no other local optimum. The proof is straightforward as by definition isolated local optimum could not be pseudo local optimum. In fact, in this case the optimum can be proved to also be unique.

Another way to eliminate pseudo local optima is by strengthening the monotonicity of $f(h_x(t))$ in Condition (C). Consider the following condition which is slightly stronger than (C).

(C') Condition (C) holds, and there exists $k > 0$ such that $\forall x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, $\forall 0 \leq t < s \leq 1$ we have

$$f(h_x(t)) - f(h_x(s)) \geq k\|h_x(t) - h_x(s)\|. \tag{2.3}$$

In Condition (C'), $\|\cdot\|$ could be any norm on $\mathbb{K}^n$. As a caveat, ℓ_0-"norm" is *not* allowed here as it is not a norm since it does not satisfy $\|\alpha x\| = |\alpha|\|x\|$. Note that Condition (C) already implies $f(h_x(t)) - f(h_x(s)) \geq 0$, while (C') strengthens this condition by enforcing a positive lower bound depending on h_x.

Theorem 2. *If (C') holds, then any local optimum of* (2.1) *must be a global optimum.*

Proof. Following the proof of Theorem 1, suppose $x \in \mathcal{X}$ is a local but not global optimum for (2.1). Then we have $x = \ell(t^\dagger)$ and could obtain a limit point of the sequence h_m, denoted as h. Since both sides of (2.3) are continuous in $h_m(t)$ and $h_m(s)$, and the limits of $h_m(t)$ and $h_m(s)$ are $h(t)$ and $h(s)$, we must have whenever $h(t) \neq h(s)$,

$$f(h(t)) - f(h(s)) \geq k\|h(t) - h(s)\| > 0.$$

Taking $t = 0$ we can conclude that $h(0)$ (which is the same point as x) is not a local optimum of (2.1). □

2.4 Necessary Conditions

In this section we will study the necessary conditions for a non-convex problem to have exact relaxation and no spurious local optima simultaneously. It turns out the results are not exactly the converses of Theorem 1 or Theorem 2, but in a slightly weaker sense. Specifically, we show that if a non-convex problem is known to have exact relaxation and no spurious local optima simultaneously, then the Lyapunov-like function and paths satisfying Condition (C) are guaranteed to exist. However, it still may or may not be easy to find those functions or paths in practice for a specific problem.

Results

Assumption 1. *The feasible set $\mathcal{X}$ is semianalytic and the cost function f is analytic.*

We refer to [13] for more detailed definitions and properties of semianalytic sets. This assumption is not restrictive for most engineering problems. If $\mathbb{K}$ is chosen as

$\mathbb{C}$, then we suggest to view all the complex functions as functions of real variables by separating the real and imaginary parts, and the space of $\mathbb{C}^n$ can be viewed as a shorthand for $\mathbb{R}^{2n}$ in this section.

Theorem 3 (necessary condition)**.** *If* (2.2) *is exact with respect to* (2.1) *and any local optimum of* (2.1) *is globally optimal, then there exists a Lyapunov-like function V and a corresponding family of paths $\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ satisfying (C1) and (C2).*

Remark 3. *Note that Theorem 3 is NOT the converse of Theorem 1 in a strict sense. There are a few differences in their settings.*

- *Theorem 1 allows pseudo local optimum (in the conclusion) of the theorem, while Theorem 3 disallows it (in the premise).*
- *Theorem 3 relies on Assumption 1 while Theorem 1 does not.*

Proof Setup

In the rest of the section, we will prove Theorem 3. From now on, we assume (2.2) is exact with respect to (2.1) and any local optimum of (2.1) is also globally optimal. We first have the following definition and lemmas, which are the main reasons we introduced Assumption 1.

Definition 11 (Whitney regularity [12, 13, 36])**.** *For a compact set $\mathcal{U} \subset \mathbb{K}^n$ and a positive integer p, we say $\mathcal{U}$ is p-regular if there exists $C > 0$ such that $\forall x, y \in \mathcal{U}$, x, y can be joined by a rectifiable curve h in $\mathcal{U}$ satisfying $L(h) \le C\|x - y\|^{1/p}$.*

Lemma 4 (Theorem 6.10 in [13])**.** *If $\mathcal{U}$ is a compact connected subanalytic subset of $\mathbb{K}^n$, then there is a positive integer p such that $\mathcal{U}$ is p-regular and the curves can always be chosen semianalytic.*

The proof of Lemma 4 can be found in [13]. Note that any semianalytic set is also subanalytic.

Lemma 5. *For any $x_0 \in \mathcal{X}$ that is not a local optimum of* (2.1) *and for any $\epsilon > 0$, there exists a path h in $\mathcal{X}$ such that $h(0) = x_0$, $f(h(t))$ is non-increasing in t, $f(h(1)) < f(h(0))$ and $L(h) < \epsilon$.*

Proof. Consider the set $\mathcal{U} := \{x \in \mathcal{X} : f(x) \le f(x_0)\}$, which by definition is also semi-analytic. Since $x_0 \in \mathcal{X}$ is not an optimum of (2.1), the problem $\min_{x\in\mathcal{U}} f(x)$

must also be exact with respect to (2.2) and it does not introduce new local optima compared to (2.1). By Lemma 3, $\mathcal{U}$ must be connected.

According to Lemma 4, there is a rectifiable and semianalytic curve h_0 in $\mathcal{U}$ such that $h_0(0) = x_0$, $L(h_0) < \epsilon$, $f(h_0(1)) < f(x_0)$ and $f(h_0(t)) \leq f(x_0)$ for all $t \in [0, 1]$. Here $f(h_0(1))$ can be chosen as any point in $\mathcal{U}$ which has a strictly smaller cost value than x_0 and is sufficiently close to x_0 in Euclidean distance. [4] It is known that a semianalytic curve is analytic except for a finite number of points [29]. Assume $h_0(t)$ is not analytic at $0 = a_0 < a_1 < \cdots < a_k = 1$ where $k \geq 1$. By Theorem on the parametrization of a semi-analytic arc in [50] and the assumption that f is analytic, the value of $f(h_0(t))$ within any interval $[a_{\ell-1}, a_\ell]$ should be equal to some analytic function defined over an open superset of $[a_{\ell-1}, a_\ell]$. Since $f(h_0(1)) < f(h_0(0))$, the function $f(h_0(t))$ cannot be a constant function over $[0, 1]$. Let $[a_{\ell-1}, a_\ell]$ be the first interval within which $f(h_0(t))$ is not constant, then $f(h_0(a_{\ell-1})) = f(h_0(0))$. As $f(h_0(t))$ within $[a_{\ell-1}, a_\ell]$ equals to a analytic function defined over an open superset of $[a_{\ell-1}, a_\ell]$, there must be a small subinterval $[a_{\ell-1}, a_{\ell-1} + \delta)$ for some $\delta > 0$ within which we always have

$$f(h_0(t)) = f(h_0(a_{\ell-1})) + \sum_{i=0}^{\infty} c_i (t - a_{\ell-1})^i,$$

where the right hand side is the Taylor expansion of $f(h_0(t))$ at $a_{\ell-1}$. Since $f(h_0(t))$ is not constant over $[a_{\ell-1}, a_\ell]$, the coefficients c_i cannot all be zeros by the identity theorem. Suppose c_{i_0} is the first nonzero coefficient in the sequence $\{c_i\}_{i=0}^{\infty}$, we have two cases. If $c_{i_0} > 0$, then $f(h_0(t))$ is strictly increasing within $[a_{\ell-1}, a_{\ell-1} + \delta')$ for some small positive $\delta' < \delta$. It contradicts to the facts that $f(h_0(a_{\ell-1})) = f(x_0)$ and $f(h_0(t)) \leq f(x_0)$ for all $t \in [0, 1]$. If $c_{i_0} < 0$, then $f(h_0(t))$ is strictly decreasing within $[a_{\ell-1}, a_{\ell-1} + \delta')$ for some small positive $\delta' < \delta$. Then we can construct a new path h such that $h(t) = h_0(t \cdot (a_{\ell-1} + \delta'))$ for all $t \in [0, 1]$. It is easy to check that h satisfies all the requirements in Lemma 5. □

Now we consider weaker versions of (C1) and (C2).

(C3) For any $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, there is a path h_x in $\hat{\mathcal{X}}$ such that $h_x(0) = x, h_x(1) \in \mathcal{X}$, both $f(h_x(t))$ and $V(h_x(t))$ are non-increasing for $t \in [0, 1]$.

(C4) All the $\{L(h_x)\}_{x \in \hat{\mathcal{X}} \setminus \mathcal{X}}$ are finite and uniformly bounded.

[4] We can always do so because x_0 is not a local optimum of (2.1). The inequality $L(h_0) < \epsilon$ is satisfied because of the p-regularity of $\mathcal{U}$.

Compared to (C1), (C3) does not require $f(h_x(0)) > f(h_x(1))$ to strictly hold. Then we have a weaker version of Theorem 3 as follows.

Lemma 6 (weaker necessary condition)**.** *If* (2.2) *is exact to* (2.1) *and any local optimum of* (2.1) *is also globally optimal, then there always exists a Lyapunov-like function V and a corresponding family of paths* $\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ *satisfying (C3) and (C4).*

We now show that Lemma 6, though weaker in its statement, actually implies Theorem 3, so later on we will only focus on the proof of Lemma 6. To see this, we suppose $V^\dagger$ and $\{h_x^\dagger\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ are the Lyapunov-like function and paths guaranteed by Lemma 6.

For each $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, if $h_x^\dagger(1)$ is a local optimum (so it is also a global optimum) of (2.1) then we must have $f(h_x(1)) < f(h_x(0))$ since the relaxation is exact. We construct $h_x^\ddagger = h_x^\dagger$.

If $h_x^\dagger(1)$ is not a local optimum, then by Lemma 5, there exists a path $h_x^\wr$, which satisfies $h_x^\wr(0) = h_x^\dagger(1)$, $h_x^\wr(t) \in \mathcal{X}$, $f(h_x^\wr(t))$ is non-increasing, $f(h_x^\wr(1)) < f(h_x^\wr(0))$ and $L(h_x^\wr) < \epsilon$. Here we choose ϵ as a fixed positive value for all x. We then construct

$$h_x^\ddagger(t) := \begin{cases} h_x^\dagger(2t), & \text{if } t \in [0, 1/2] \\ h_x^\wr(2t-1), & \text{if } t \in [1/2, 1]. \end{cases}$$

We now let $V = V^\dagger$ and $h_x = \overline{h_x^\ddagger}$ for all $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$. Recall that $\overline{h_x^\ddagger}$ is the arc-length reparameterization of $h_x^\ddagger$. Clearly, such construction satisfies (C1) as we strictly reduce the cost at the end of each path unless the path has already reached the global optimum. Besides, $\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ is uniformly bounded as $\hat{\mathcal{X}}$ is compact, and $\{L(h_x)\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ also has the uniform upperbound as both $L(h_x^\dagger)$ and $L(h_x^\wr)$ are uniformly bounded for all x. By Lemma 2 $\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ is uniformly equicontinuous, so (C2) is also satisfied. To summarize, when Lemma 6 is correct, one can always revise the Lyapunov-like function and paths provided by Lemma 6 to make Theorem 3 hold as well. Therefore, in the rest of the section, we will only prove Lemma 6 and the correctness of Theorem 3 just follows.

Let x^* be any global optimum of (2.1), then it is also an optimum of (2.2). Define

$$\begin{aligned} \mathcal{H} :=& \{h \mid h : [0,1] \to \hat{\mathcal{X}} \text{ is continuous and } L(h) < \infty\} \\ \bar{\mathcal{H}} :=& \{\bar{h} \mid h \in \mathcal{H}\} \\ \hat{\mathcal{H}} :=& \{h \mid h \in \bar{\mathcal{H}},\ f(h(t)) \geq f(h(1)) \text{ for all } t \in [0,1]\}. \end{aligned}$$

An immediate observation is if a continuous function $h : [0,1] \to \hat{\mathcal{X}}$ satisfies $L(h) < \infty$ and $f(h(t)) \geq f(h(1))$ for all t, then $\bar{h} \in \hat{\mathcal{H}}$.

Construction

We construct V as

$$V(x) = \inf_{\substack{h \in \hat{\mathcal{H}} \\ h(0)=x \\ h(1) \in \mathcal{X}}} L(h) \tag{2.4}$$

Lemma 7. *For a sequence $(h_i)_{i=1}^{\infty}$ where $h_i \in \hat{\mathcal{H}}$, if both $(h_i)_{i=1}^{\infty}$ and $(L(h_i))_{i=1}^{\infty}$ are uniformly bounded, then there must be a subsequence which uniformly converges to some h^* such that its arc-length reparameterization, denoted as $\overline{h^*}$, is in $\hat{\mathcal{H}}$. Furthermore, $L(h^*) = L(\overline{h^*}) \leq \limsup_i L(h_i)$.*

Proof. By Lemma 2, $(h_i)_{i=1}^{\infty}$ is both uniformly bounded and uniformly equicontinuous. By Arzelà-Ascoli theorem, a subsequence of $(h_i)_{i=1}^{\infty}$ uniformly converges to a limit h^*. Without loss of generalization, we denote this subsequence as $(h_i)_{i=1}^{\infty}$ as well. By uniform limit theorem and the compactness of $\hat{\mathcal{X}}$, $\overline{h^*}$ is a continuous function mapping from $[0,1]$ to $\hat{\mathcal{X}}$. To show $\overline{h^*} \in \hat{\mathcal{H}}$, it is sufficient to show $f(h^*(t)) \geq f(h^*(1))$ for all $t \in [0,1]$ and $L(h^*) < \infty$. If $f(h^*(t)) = f(h^*(1)) - \epsilon$ for some $t \in [0,1]$ and $\epsilon > 0$, then for sufficiently large i, we would have $|f(h_i(t)) - f(h^*(t))| < \epsilon/3$ and $|f(h_i(1)) - f(h^*(1))| < \epsilon/3$. Thus $f(h_i(t)) \leq f(h_i(1)) - \epsilon/3$ and it contradicts to $h_i \in \hat{\mathcal{H}}$.

Instead of showing $L(h^*) < \infty$, we directly prove $L(h^*) \leq \limsup_i L(h_i)$. Otherwise, there exists $\pi = (t_1, \cdots, t_N) \in \Pi$ such that $L_\pi(h^*) = \limsup_i L(h_i) + \epsilon$ for $\epsilon > 0$. For sufficiently large i, we have

$$\begin{aligned}
&|L_\pi(h_i) - L_\pi(h^*)| \\
=&\Big| \sum_{j=1}^{N} \|h_i(t_{j-1}) - h(t_j)\|_{\ell_2} - \sum_{j=1}^{N} \|h^*(t_{j-1}) - h^*(t_j)\|_{\ell_2} \Big| \\
\leq& \sum_{j=1}^{N} \Big(\|h_i(t_{j-1}) - h^*(t_{j-1})\|_{\ell_2} + \|h_i(t_j) - h^*(t_j)\|_{\ell_2} \Big) \leq \frac{\epsilon}{2}.
\end{aligned}$$

Thus, $L(h_i) \geq L_\pi(h_i) \geq \limsup_i L(h_i) + \epsilon/2$ holds for sufficiently large i. It contradicts the definition of lim sup. As a result, we must have $L(h^*) \leq \limsup_i L(h_i)$. □

Lemma 8. *The optimization in* (2.4) *is feasible and the optimal cost can be achieved.*

Proof. We fix some $x \in \hat{\mathcal{X}}$. To show the feasibility, consider $h_{\text{fea}}(t) := (1-t)x + tx^*$, which is feasible to (2.4). Let $L_{\text{fea}} = L(h_{\text{fea}})$. Since L_{fea} is finite and $L(h)$ is non-negative, $V(x)$ must be finite. To show the achievability of the optimal cost, we prove by contradiction. If not, then there must be a sequence of feasible $(h_i)_{i=1}^{\infty}$ such that

$$\begin{aligned} &L_{\text{fea}} > L(h_i) \geq L(h_{i+1}) > V(x) \text{ for all } i \geq 1 \\ &\lim_{i\to\infty} L(h_i) = V(x). \end{aligned}$$

The compactness of $\hat{\mathcal{X}}$ implies $(h_i)_{i=1}^{\infty}$ is uniformly bounded as well. By Lemma 7, a subsequence of $(h_i)_{i=1}^{\infty}$, denoted as $(h_i)_{i=1}^{\infty}$ as well, uniformly converges to a limit h^* and $L(h^*) = L(\overline{h^*}) \leq V(x)$. Moreover,

$$\begin{aligned} \overline{h^*}(0) = h^*(0) = \lim_{i\to\infty} h_i(0) = x \\ \overline{h^*}(1) = h^*(1) = \lim_{i\to\infty} h_i(1) \in \mathcal{X}. \end{aligned}$$

Above all, we proved $\overline{h^*}$ is feasible to (2.4), and the cost $L(\overline{h^*})$ is not worse than $V(x)$. It contradicts the non-achievability assumption. □

For each $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, we construct h_x as

$$h_x = \arg\min_{\substack{h\in\hat{\mathcal{H}} \\ h(0)=x \\ h(1)\in\mathcal{X}}} L(h). \tag{2.5}$$

If there are multiple minimizers then h_x can be chosen as any one of them.

Lemma 9. *For $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, the function h_x is injective.*

Proof. Otherwise, for some x, there exist $t_1 < t_2$ such that $h_x(t_1) = h_x(t_2)$. Since $h_x \in \hat{\mathcal{H}} \subseteq \bar{\mathcal{H}}$, we have

$$\sup_{\pi\in\Pi|_{t_1}^{t_2}} L_\pi(h_x) = (t_2 - t_1)L(h_x) = (t_2 - t_1)V(x) > 0.$$

Consider a new path defined as

$$h^*(t) := \begin{cases} h_x(t), & \text{if } t \in [0,1] \setminus [t_1, t_2] \\ h_x(t_1), & \text{if } t \in [t_1, t_2] \end{cases}.$$

It is easy to check h^* is continuous and entirely within $\hat{\mathcal{X}}$. For any $t \in [0,1]$, $h_x \in \hat{\mathcal{H}}$ implies $f(h^*(t)) \geq f(h_x(1)) = f(h^*(1))$. Further, we have

$$\begin{aligned}
L(h^*) =& \sup_{\pi\in\Pi} L_\pi(h^*) \\
=& \sup_{\pi\in\Pi|_0^{t_1}} L_\pi(h^*) + \sup_{\pi\in\Pi|_{t_1}^{t_2}} L_\pi(h^*) + \sup_{\pi\in\Pi|_{t_2}^{1}} L_\pi(h^*) \\
=& \sup_{\pi\in\Pi|_0^{t_1}} L_\pi(h_x) + 0 + \sup_{\pi\in\Pi|_{t_2}^{1}} L_\pi(h_x) \\
<& \sup_{\pi\in\Pi|_0^{t_1}} L_\pi(h_x) + \sup_{\pi\in\Pi|_{t_1}^{t_2}} L_\pi(h_x) + \sup_{\pi\in\Pi|_{t_2}^{1}} L_\pi(h_x) \\
=& \sup_{\pi\in\Pi} L_\pi(h_x) = L(h_x).
\end{aligned}$$

Above all, the arc-length reparameterization of h^*, denoted as $\overline{h^*}$, is feasible to (2.5) but achieves a strictly lower cost than h_x. This contradicts to the optimality of h_x. □

Corollary 4. *For distinctive $t_1, t_2, t_3 \in [0,1]$, if $f(h_x(t_1)) \geq f(h_x(t_2))$ and $f(h_x(t_1)) > f(h_x(t_3))$, then $\|h_x(t_1) - h_x(t_2)\|_{\ell_2} + \|h_x(t_1) - h_x(t_3)\|_{\ell_2} > \|h_x(t_2) - h_x(t_3)\|_{\ell_2}$.*

Proof. It is sufficient to show $h_x(t_1)$ is not the convex combination of $h_x(t_2)$ and $h_x(t_3)$. Otherwise, we assume $h_x(t_1) = \lambda h_x(t_2) + (1-\lambda)h_x(t_3)$ for some $\lambda \in [0,1]$. First, Lemma 9 implies $\lambda \neq 0, 1$. For $\lambda \in (0,1)$, the convexity of f implies

$$\begin{aligned}
f(h_x(t_1)) =& f(\lambda h_x(t_2) + (1-\lambda)h_x(t_3)) \\
\leq& \lambda f(h_x(t_2)) + (1-\lambda) f(h_x(t_3)) \\
<& \lambda f(h_x(t_1)) + (1-\lambda) f(h_x(t_1)) = f(h_x(t_1)).
\end{aligned}$$

This contradiction shows $h_x(t_1)$ is not the convex combination of $h_x(t_2)$ and $h_x(t_3)$. Then the triangle inequality implies this corollary. □

Lemma 10. *For each h_x defined in* (2.5)*, we have $f(h_x(t))$ is non-increasing in t for $t \in [0,1]$.*

Proof. We fix an $x_0 \in \hat{\mathcal{X}} \setminus \mathcal{X}$ and prove the result for h_{x_0} defined above. If not, then there exist $0 \leq t_1 < t_2 \leq 1$ such that $f(h_{x_0}(t_1)) < f(h_{x_0}(t_2))$. Now define

$$t^\dagger = \arg\max_{t\in[t_1,1]} f(h_{x_0}(t)), \quad t^\ddagger = \max_{\substack{t\in[t^\dagger,1] \\ f(h_{x_0}(t))=f(h_{x_0}(t^\dagger))}} t.$$

In other words, $t^\dagger$ is an arbitrary maximizer of $f(h_{x_0}(t))$, while $t^\ddagger$ is the largest maximizer. Clearly, both $t^\dagger$ and $t^\ddagger$ are well defined (due to the continuity and closedness) and are strictly between t_1 and 1. We also have that $f(h_{x_0}(t^\ddagger)) > f(h_{x_0}(1))$ strictly holds. By the continuity of $h_{x_0}(\cdot)$ and $f(h_{x_0}(\cdot))$, there exist $r, \delta > 0$ such that $[t^\ddagger - \delta, t^\ddagger + \delta] \subseteq (t_1, 1)$ and

- for $x \in \mathcal{B}(h_{x_0}(t^\ddagger), r) \cap \hat{\mathcal{X}}$, $f(h_{x_0}(1)) \leq f(x)$.
- for $t \in [t^\ddagger - \delta, t^\ddagger)$, $h_{x_0}(t) \in \mathcal{B}(h_{x_0}(t^\ddagger), r)$. Therefore $f(h_{x_0}(1)) \leq f(h_{x_0}(t)) \leq f(h_{x_0}(t^\ddagger))$.
- for $t \in (t^\ddagger, t^\ddagger + \delta]$, $h_{x_0}(t) \in \mathcal{B}(h_{x_0}(t^\ddagger), r)$. Therefore $f(h_{x_0}(1)) \leq f(h_{x_0}(t)) < f(h_{x_0}(t^\ddagger))$.

Now we construct another h^* as

$$h^*(t) = \begin{cases} h_{x_0}(t), & \text{if } t \in [0,1] \setminus [t^\ddagger - \delta, t^\ddagger + \delta] \\ \frac{t^\ddagger + \delta - t}{2\delta} h_{x_0}(t^\ddagger - \delta) + \frac{t - t^\ddagger + \delta}{2\delta} h_{x_0}(t^\ddagger + \delta), & \\ & \text{if } t \in [t^\ddagger - \delta, t^\ddagger + \delta]. \end{cases}$$

It is easy to verify h^* is continuous. For $t \in [t^\ddagger - \delta, t^\ddagger + \delta]$, $h^*(t)$ is the convex combination of $h_{x_0}(t^\ddagger - \delta)$ and $h_{x_0}(t^\ddagger + \delta)$, and must be within $\mathcal{B}(h_{x_0}(t^\ddagger), r) \cap \hat{\mathcal{X}}$, which is convex. Therefore, h^* is entirely within $\hat{\mathcal{X}}$ and $f(h^*(t)) \geq f(h_{x_0}(1)) = f(h^*(1))$ holds for all t.

Next, we show $L(h^*) < L(h_{x_0})$ by (2.6). The strict inequality in (2.6) is because of Corollary 4.

Above all, the arc-length reparameterization of h^*, denoted as $\overline{h^*}$, is feasible to (2.5) but achieves a strictly lower cost than h_{x_0}. This contradicts the optimality of h_{x_0}. □

Verification

To show V satisfies Definition 10

It is sufficient to show V is continuous in x. The proof is twofold. To abuse the notations a little bit, we let $h_x(t) \equiv x$ for $x \in \mathcal{X}$, so that h_x is the unique minimizer of (2.4) and $L(h_x) = V(x) = 0$ for $x \in \mathcal{X}$.

First we show for $x_0 \in \hat{\mathcal{X}}$ and $\epsilon > 0$, there exists $\delta_+ > 0$ such that $\forall x \in \mathcal{B}(x_0, \delta_+) \cap \hat{\mathcal{X}}$, $V(x) \leq V(x_0) + \epsilon$. There are two scenarios. If $h_{x_0}(1)$ is a global optimum of (2.1),

$$
\begin{aligned}
& L(h^*) \\
=& \sup_{\pi\in\Pi} L_\pi(h^*) = \sup_{\pi\in\Pi|_0^{t^\ddagger-\delta}} L_\pi(h^*) + \sup_{\pi\in\Pi|_{t^\ddagger-\delta}^{t^\ddagger+\delta}} L_\pi(h^*) + \sup_{\pi\in\Pi|_{t^\ddagger+\delta}^{1}} L_\pi(h^*) \\
=& \sup_{\pi\in\Pi|_0^{t^\ddagger-\delta}} L_\pi(h_{x_0}) + \|h_{x_0}(t^\ddagger-\delta) - h_{x_0}(t^\ddagger+\delta)\|_{\ell_2} + \sup_{\pi\in\Pi|_{t^\ddagger+\delta}^{1}} L_\pi(h_{x_0}) \\
<& \sup_{\pi\in\Pi|_0^{t^\ddagger-\delta}} L_\pi(h_{x_0}) + \|h_{x_0}(t^\ddagger-\delta) - h_{x_0}(t^\ddagger)\|_{\ell_2} + \|h_{x_0}(t^\ddagger) - h_{x_0}(t^\ddagger+\delta)\|_{\ell_2} \\
& + \sup_{\pi\in\Pi|_{t^\ddagger+\delta}^{1}} L_\pi(h_{x_0}) \\
\le& \sup_{\pi\in\Pi|_0^{t^\ddagger-\delta}} L_\pi(h_{x_0}) + \sup_{\pi\in\Pi|_{t^\ddagger-\delta}^{t^\ddagger+\delta}} L_\pi(h_{x_0}) + \sup_{\pi\in\Pi|_{t^\ddagger+\delta}^{1}} L_\pi(h_{x_0}) \\
=& \sup_{\pi\in\Pi} L_\pi(h_{x_0}) = L(h_{x_0}).
\end{aligned} \tag{2.6}
$$

---

then we could set $\delta_+ = \epsilon$. For any $x \in \mathcal{B}(x_0, \delta_+) \cap \hat{\mathcal{X}}$, construct

$$
h^*(t) = \begin{cases} (1-2t)x + 2tx_0, & t \in [0, \frac{1}{2}] \\ h_{x_0}(2t-1), & t \in (\frac{1}{2}, 1]. \end{cases}
$$

Its arc-length reparameterization $\overline{h^*}$ is feasible to (2.4) (w.r.t. x) and $V(x) \le L(\overline{h^*}) = |x - x_0| + L(h_{x_0}) \le V(x_0) + \epsilon$. Next we focus on the scenario that $h_{x_0}(1)$ is not a global optimum of (2.1), so it is not a local optimum neither. By Lemma 5, there is a path $h^\wr$ in $\mathcal{X}$ such that $h^\wr(0) = h_{x_0}(1)$, $f(h^\wr(t))$ is non-increasing in t, $f(h^\wr(1)) < f(h^\wr(0))$ and $L(h^\wr) < \epsilon/2$. Suppose $f(h^\wr(0)) - f(h^\wr(1)) = \tau > 0$. Since f is continuous, there must be some $\gamma > 0$ such that for any $x \in \mathcal{B}(x_0, \gamma) \cap \hat{\mathcal{X}}$, we have $|f(x) - f(x_0)| < \tau$. Now we choose δ_+ as $\min(\gamma, \epsilon/2)$. For any $x \in \mathcal{B}(x_0, \delta_+) \cap \hat{\mathcal{X}}$, construct

$$
h^*(t) = \begin{cases} (1-3t)x + 3tx_0, & t \in [0, \frac{1}{3}] \\ h_{x_0}(3t-1), & t \in (\frac{1}{3}, \frac{2}{3}] \\ h^\wr(3t-2), & t \in (\frac{2}{3}, 1]. \end{cases}
$$

Its arc-length reparameterization $\overline{h^*}$ is feasible to (2.4) (w.r.t. x) and $V(x) \le L(\overline{h^*}) = |x - x_0| + L(h_{x_0}) + L(h^\wr) \le \delta_+ + V(x_0) + \epsilon/2 \le V(x_0) + \epsilon$.

Second we show for $x_0 \in \hat{\mathcal{X}}$ and $\epsilon > 0$, there exists $\delta_- > 0$ such that $\forall x \in \mathcal{B}(x_0, \delta_-) \cap \hat{\mathcal{X}}$, $V(x) \ge V(x_0) - \epsilon$. If not, then there must be a sequence $(x_i)_{i=1}^\infty$ such that $\lim_{i\to\infty} x_i = x_0$ but $V(x_i) < V(x_0) - \epsilon$ for all $i \ge 1$. Let $h_i := h_{x_i}$ for $i \ge 0$, then both $(h_i)_{i=1}^\infty$ and $(L(h_i))_{i=1}^\infty$ are uniformly bounded. By Lemma 7, a subsequence of

$(h_i)_{i=1}^{\infty}$ uniformly converges to a limit h^* and

$$\begin{aligned} L(h^*) =& L(\overline{h^*}) \le \limsup_i L(h_i) \\ =& \limsup_i V(x_i) \le V(x_0) - \epsilon. \end{aligned}$$

Lemma 7 also indicates $\overline{h^*} \in \hat{\mathcal{H}}$ and $h^*(0) = \lim_{i\to\infty} h_i(0) = \lim_{i\to\infty} x_i = x_0$, $h^*(1) = \lim_{i\to\infty} h_i(1) \in \mathcal{X}$ (as $\mathcal{X}$ is closed). Therefore, $\overline{h^*}$ is feasible to (2.4) but its cost is strictly lower than $V(x_0)$. It leads to the contradiction.

To show (C3) holds

By our construction (2.5), h_x is entirely within $\hat{\mathcal{X}}$, and $h_x(0) = x$, $h_x(1) \in \mathcal{X}$. Lemma 10 shows $f(h_x(t))$ is non-increasing for $t \in [0, 1]$. It is sufficient to show $V(h_x(t))$ is also non-increasing for $t \in [0, 1]$. Consider the following lemma.

Lemma 11. *For fixed $x_0 \in \hat{\mathcal{X}}$ and $t_0 \in [0, 1]$,*

$$V(h_{x_0}(t_0)) = \sup_{\pi \in \Pi|_{t_0}^1} L_\pi(h_{x_0}).$$

Proof. Let $x_1 = h_{x_0}(t_0)$. We have $L(h_{x_1}) = V(h_{x_0}(t_0))$. If the lemma does not hold, then we have two cases.

First, if $L(h_{x_1}) < \sup_{\pi\in\Pi|_{t_0}^1} L_\pi(h_{x_0})$, then let

$$h^*(t) = \begin{cases} h_{x_0}(2t_0 t), & t \in [0, \frac{1}{2}] \\ h_{x_1}(2t-1), & t \in (\frac{1}{2}, 1] \end{cases}.$$

It is easy to check h^* is continuous and entirely within $\hat{\mathcal{X}}$, and $h^*(0) = x_0$, $h^*(1) = h_{x_1}(1) \in \mathcal{X}$. For $t \in [0, 1/2]$,

$$\begin{aligned} f(h^*(t)) =& f(h_{x_0}(2t_0 t)) \ge f(h_{x_0}(t_0)) = f(x_1) \\ =& f(h_{x_1}(0)) \ge f(h_{x_1}(1)) = f(h^*(1)). \end{aligned}$$

For $t \in [1/2, 1]$,

$$f(h^*(t)) = f(h_{x_1}(2t-1)) \ge f(h_{x_1}(1)) = f(h^*(1)).$$

Further, we have

$$\begin{aligned}
L(h^*) &= \sup_{\pi\in\Pi} L_\pi(h^*) \\
&= \sup_{\pi\in\Pi|_0^{0.5}} L_\pi(h^*) + \sup_{\pi\in\Pi|_{0.5}^{1}} L_\pi(h^*) \\
&= \sup_{\pi\in\Pi|_0^{t_0}} L_\pi(h_{x_0}) + \sup_{\pi\in\Pi} L_\pi(h_{x_1}) \\
&= \sup_{\pi\in\Pi|_0^{t_0}} L_\pi(h_{x_0}) + L(h_{x_1}) \\
&< \sup_{\pi\in\Pi|_0^{t_0}} L_\pi(h_{x_0}) + \sup_{\pi\in\Pi|_{t_0}^{1}} L_\pi(h_{x_0}) = L(h_{x_0}).
\end{aligned}$$

Above all, the arc-length reparameterization of h^*, denoted as $\overline{h^*}$, is feasible to (2.5) (w.r.t. x_0) but achieves a strictly lower cost than h_{x_0}. This contradicts the optimality of h_{x_0}.

Second, if $L(h_{x_1}) > \sup_{\pi\in\Pi|_{t_0}^{1}} L_\pi(h_{x_0})$, then let

$$h^*(t) = \begin{cases} h_{x_0}(t_0), & t \in [0, t_0] \\ h_{x_0}(t), & t \in (t_0, 1]. \end{cases}$$

It is easy to check that h^* is continuous and entirely within $\hat{\mathcal{X}}$, and $h^*(0) = h_{x_0}(t_0) = x_1, h^*(1) = h_{x_0}(1) \in \mathcal{X}$. For $t \in [0, 1]$, $f(h_{x_0}(t)) \geq f(h_{x_0}(1))$ implies $f(h^*(t)) \geq f(h_{x_0}(1)) = f(h^*(1))$. Further, we have

$$\begin{aligned}
L(h^*) = \sup_{\pi\in\Pi} L_\pi(h^*) &= \sup_{\pi\in\Pi|_0^{t_0}} L_\pi(h^*) + \sup_{\pi\in\Pi|_{t_0}^{1}} L_\pi(h^*) \\
=0 + \sup_{\pi\in\Pi|_{t_0}^{1}} L_\pi(h_{x_0}) &< L(h_{x_1}).
\end{aligned}$$

Above all, the arc-length reparameterization of h^*, denoted as $\overline{h^*}$, is feasible to (2.5) (w.r.t. x_1) but achieves a strictly lower cost than h_{x_1}. This contradicts the optimality of h_{x_1}. □

Using this lemma, we are in a good position to show $V(h_x(t))$ is non-increasing for $t \in [0, 1]$. For any $t_1 < t_2$, we have

$$\begin{aligned}
V(h_x(t_1)) &= \sup_{\pi\in\Pi|_{t_1}^{1}} L_\pi(h_x) = \sup_{\pi\in\Pi|_{t_1}^{t_2}} L_\pi(h_x) + \sup_{\pi\in\Pi|_{t_2}^{1}} L_\pi(h_x) \\
&\geq \sup_{\pi\in\Pi|_{t_2}^{1}} L_\pi(h_x) = V(h_x(t_2)).
\end{aligned}$$

To show (C4) holds

The set $\{L(h_x)\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ is uniformly bounded by $\max_{x\in\hat{\mathcal{X}}}|x-x^*|$, which is finite.

To summarize, we have verified that such construction is well defined and satisfies both (C3) and (C4), so Lemma 6 is proved. Since we have shown that Lemma 6 implies Theorem 3, the latter is also proved.

2.5 Other Properties

Constructing from Primitives

Though the previous section guarantees the existence of the Lyapunov-like function and paths under certain conditions, it is not clear how to systematically find or construct them. In this subsection, we show that if one can find the Lyapunov-like function and paths for some primitive problems, then there are natural ways to construct the Lyapunov-like function and paths for new problems built up from those primitives in certain ways. To streamline the notations, we will use the tuple $(f,\mathcal{X})$ to refer to (2.1) and the tuple $(f,\mathcal{X},\hat{\mathcal{X}})$ to refer to the problem pair (2.1), (2.2). Assume $(V,\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}})$ is a valid construction of Lyapunov-like function and paths for $(f,\mathcal{X},\hat{\mathcal{X}})$. In this subsection, when we say V and h_x are valid, it means they not only are valid by definition, but also satisfy (C1) and (C2).

Function Composition

Suppose $g:\mathbb{R}\to\mathbb{R}$ is non-decreasing and convex. Then $(V,\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}})$ is also a valid construction of Lyapunov-like function and paths for $(g\circ f,\mathcal{X},\hat{\mathcal{X}})$. This result is trivial as $g\circ f$ preserves the convexity over $\hat{\mathcal{X}}$ and monotonicity over any path.

Union of Feasible Sets

Suppose for two pairs of problems $(f_1,\mathcal{X}_1,\hat{\mathcal{X}}_1)$, for which $(V^1,\{h^1_x\}_{x\in\hat{\mathcal{X}}_1\setminus\mathcal{X}_1})$ is valid, and $(f_2,\mathcal{X}_2,\hat{\mathcal{X}}_2)$, for which $(V^2,\{h^2_x\}_{x\in\hat{\mathcal{X}}_2\setminus\mathcal{X}_2})$ is valid. We consider a new problem $(f,\mathcal{X},\hat{\mathcal{X}})$ where $\mathcal{X}:=(\mathcal{X}_1\cup\mathcal{X}_2)\cap\hat{\mathcal{X}}_1\cap\hat{\mathcal{X}}_2$ and $\hat{\mathcal{X}}:=\hat{\mathcal{X}}_1\cap\hat{\mathcal{X}}_2$. The formulation of f will be provided later. If for any $x\in\hat{\mathcal{X}}\setminus\mathcal{X}$, we have $h^1_x\equiv h^2_x$, then construct $\tilde{V}:\hat{\mathcal{X}}\to\mathbb{R}$ such that $\tilde{V}(x):=V^1(x)\cdot V^2(x)$ and $\tilde{h}_x=h^1_x$ for all $x\in\hat{\mathcal{X}}\setminus\mathcal{X}$. We have the following two results.

Corollary 5. *For any $\lambda\in(0,1)$, define $f:\hat{\mathcal{X}}\to\mathbb{R}$ as $f(x):=\lambda f_1(x)+(1-\lambda)f_2(x)$. Then $(\tilde{V},\{\tilde{h}_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}})$ is valid for $(f,\mathcal{X},\hat{\mathcal{X}})$.*

Table 2.1: A summary on constructing V and h_x from primitives.

Operation	Primitive problem	New problem	V, h_x for new problem	Additional requirements
Function Composition	$(f, \mathcal{X}, \hat{\mathcal{X}})$: V, h_x	$(g \circ f, \mathcal{X}, \hat{\mathcal{X}})$	$\tilde{V} := V$ $\tilde{h}_x := h_x$	g is non-decreasing and convex
Union of feasible sets (with cost as the sum)	$(f_1, \mathcal{X}_1, \hat{\mathcal{X}}_1)$: V^1, h_x^1 $(f_2, \mathcal{X}_2, \hat{\mathcal{X}}_2)$: V^2, h_x^2	$(f, \mathcal{X}, \hat{\mathcal{X}})$ where $f := \lambda f_1 + (1-\lambda) f_2$ $\mathcal{X} := (\mathcal{X}_1 \cup \mathcal{X}_2) \cap \hat{\mathcal{X}}_1 \cap \hat{\mathcal{X}}_2$ $\hat{\mathcal{X}} := \hat{\mathcal{X}}_1 \cap \hat{\mathcal{X}}_2$	$\tilde{V} := V^1 \times V^2$ $\tilde{h}_x := h_x^1$	h_x^1 and h_x^2 coincide for $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$
Union of feasible sets (with cost as the maximum)	$(f_1, \mathcal{X}_1, \hat{\mathcal{X}}_1)$: V^1, h_x^1 $(f_2, \mathcal{X}_2, \hat{\mathcal{X}}_2)$: V^2, h_x^2	$(f, \mathcal{X}, \hat{\mathcal{X}})$ where $f := \max(f_1, f_2)$ $\mathcal{X} := (\mathcal{X}_1 \cup \mathcal{X}_2) \cap \hat{\mathcal{X}}_1 \cap \hat{\mathcal{X}}_2$ $\hat{\mathcal{X}} := \hat{\mathcal{X}}_1 \cap \hat{\mathcal{X}}_2$	$\tilde{V} := V^1 \times V^2$ $\tilde{h}_x := h_x^1$	h_x^1 and h_x^2 coincide for $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$
Intersection of feasible sets (with cost as the sum)	$(f_1, \mathcal{X}_1, \hat{\mathcal{X}})$: V^1, h_x^1 $(f_2, \mathcal{X}_2, \hat{\mathcal{X}})$: V^2, h_x^2 $(u, v) =: x \in \hat{\mathcal{X}}$	$(f, \mathcal{X}, \hat{\mathcal{X}})$ where $f := \lambda f_1 + (1-\lambda) f_2$ $\mathcal{X} := \mathcal{X}_1 \cap \mathcal{X}_2$	$\tilde{V} := V^1 + V^2$ $\tilde{h}_x$ is defined as in (2.7)	$f_i(x), V^i(x)$ depend on $\mathcal{P}_i x$ only and $\mathcal{P}_{1-i} h_x^i$ is constant
Intersection of feasible sets (with cost as the maximum)	$(f_1, \mathcal{X}_1, \hat{\mathcal{X}})$: V^1, h_x^1 $(f_2, \mathcal{X}_2, \hat{\mathcal{X}})$: V^2, h_x^2 $(u, v) =: x \in \hat{\mathcal{X}}$	$(f, \mathcal{X}, \hat{\mathcal{X}})$ where $f := \max(f_1, f_2)$ $\mathcal{X} := \mathcal{X}_1 \cap \mathcal{X}_2$	$\tilde{V} := V^1 + V^2$ $\tilde{h}_x$ is defined as in (2.7)	$f_i(x), V^i(x)$ depend on $\mathcal{P}_i x$ only and $\mathcal{P}_{1-i} h_x^i$ is constant

Corollary 6. *Define function* $f : \hat{\mathcal{X}} \to \mathbb{R}$ *as* $f(x) := \max(f_1(x), f_2(x))$. *Then* $(\tilde{V}, \{\tilde{h}_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}})$ *is valid for* $(f, \mathcal{X}, \hat{\mathcal{X}})$.

Proof for Corollary 5 and Corollary 6. The function $\tilde{V}$ is still continuous and vanishes if and only if $x \in \mathcal{X}$ (since $V(x) = 0 \Leftrightarrow V^1(x) = 0$ *or* $V^2(x) = 0$). By construction, $\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ is a subset of $\{h^1_x\}_{x\in\hat{\mathcal{X}}_1\setminus\mathcal{X}_1}$ so (C2) is naturally satisfied. To see (C1) holds, we fix any $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$. Then $h_x(0) = h^1_x(x) = x$ and $h_x(1) = h^1_x(1) \in \mathcal{X}_1 \subseteq \mathcal{X}$. Further, $V(h_x(t)) = V^1(h_x(t))V^2(h_x(t)) = V^1(h^1_x(t))V^2(h^2_x(t))$ as h^1_x and h^2_x conincide when $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$. Because both $V^1(h^1_x(t))$ and $V^2(h^2_x(t))$ are non-negative and non-increasing, so is $V(h_x(t))$. Finally, as $f_1(h_x(t))$ and $f_2(h_x(t))$ are both non-increasing over $[0, 1]$, their convex-combination or maximum (i.e., $f(h_x(t))$) must be non-increasing as well. A similar argument can also be applied to show $f(h_x(1)) < f(h_x(0))$. Thus (C1) holds and it completes the proof. □

Intersection of Feasible Sets

We still consider two pairs of problems $(f_1, \mathcal{X}_1, \hat{\mathcal{X}})$, for which $(V^1, \{h^1_x\}_{x\in\hat{\mathcal{X}}_1\setminus\mathcal{X}_1})$ is valid, and $(f_2, \mathcal{X}_2, \hat{\mathcal{X}})$, for which $(V^2, \{h^2_x\}_{x\in\hat{\mathcal{X}}_2\setminus\mathcal{X}_2})$ is valid. Different from the previous setting, two pairs are required to share the same relaxation set $\hat{\mathcal{X}}$. Further, we view each $x \in \hat{\mathcal{X}}$ as a tuple with two parts $x := (u, v)$. Define $\mathcal{P}_1$ and $\mathcal{P}_2$ as two projection operators such that $\mathcal{P}_1 x = u$ and $\mathcal{P}_2 x = v$.

We consider a new problem $(f, \mathcal{X}, \hat{\mathcal{X}})$ where $\mathcal{X} := \mathcal{X}_1 \cap \mathcal{X}_2$. The formulation of f will be provided later.

If f_i, V^i and h^i_x are completely separated with respect to u and v in the sense that for $i = 1, 2$, $f_i(x), V^i(x)$ depend on $\mathcal{P}_i x$ only and $\mathcal{P}_{1-i}(h^i_x(t))$ is constant, then we can construct $\tilde{V}$ as $\tilde{V}(x) := V^1(x) + V^2(x)$. For $x \in \hat{\mathcal{X}} \setminus \mathcal{X}$, the path $\tilde{h}_x$ is constructed in three ways depending on the values of $V^1(x)$ and $V^2(x)$.

$$\text{If } V^1(x) = 0 \text{ then } \tilde{h}_x := h^2_x, \tag{2.7a}$$

$$\text{If } V^2(x) = 0 \text{ then } \tilde{h}_x := h^1_x, \tag{2.7b}$$

$$\text{If } V^1(x), V^2(x) > 0 \text{ then}$$
$$\tilde{h}_x(t) := \begin{cases} h^1_x(2t), & t \in [0, \frac{1}{2}) \\ h^2_{h^1_x(1)}(2t-1), & t \in [\frac{1}{2}, 1] \end{cases}. \tag{2.7c}$$

Corollary 7. *For any* $\lambda \in (0, 1)$, *define* $f : \hat{\mathcal{X}} \to \mathbb{R}$ *as* $f(x) := \lambda f_1(x) + (1-\lambda) f_2(x)$. *Then* $(\tilde{V}, \{\tilde{h}_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}})$ *is valid for* $(f, \mathcal{X}, \hat{\mathcal{X}})$.

Corollary 8. *Define function* $f : \hat{\mathcal{X}} \to \mathbb{R}$ *as* $f(x) := \max(f_1(x), f_2(x))$. *Then* $(\tilde{V}, \{\tilde{h}_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}})$ *is valid for* $(f, \mathcal{X}, \hat{\mathcal{X}})$.

Proof for Corollary 7 and Corollary 8. The function $\tilde{V}$ is still continuous and vanishes if and only if $x \in \mathcal{X}$ (since $V(x) = 0 \Leftrightarrow V^1(x) = 0$ *and* $V^2(x) = 0$). The set $\{\tilde{h}_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ satisfies (C2) as each path is constructed either as h^i_x or the concatenation of h^1_x and $h^2_{h^1_x(1)}$. Next we are showing $\tilde{h}_x(1)$ is in $\mathcal{X}$. If $\tilde{h}_x$ is constructed by (2.7a), then we have $\tilde{V}(\tilde{h}_x(1)) = V^1(h^2_x(1)) + V^2(h^2_x(1)) = V^1(h^2_x(1))$. Since $V^1(h^2_x(1))$ only depends on $\mathcal{P}_1 h^2_x(1)$ and $\mathcal{P}_1 h^2_x(1) = \mathcal{P}_1 h^2_x(0) = \mathcal{P}_1 x$, there must be $V^1(h^2_x(1)) = V^1(x) = 0$. Thus $\tilde{V}(\tilde{h}_x(1)) = 0$ and $\tilde{h}_x(1) \in \mathcal{X}$. It is similar if $\tilde{h}_x$ is constructed by (2.7b). When $\tilde{h}_x$ is constructed by (2.7c), then

$$\begin{aligned}\tilde{V}(\tilde{h}_x(1)) =& V^1(h^2_{h^1_x(1)}(1)) + V^2(h^2_{h^1_x(1)}(1))\\ =& V^1(h^2_{h^1_x(1)}(1)) = V^1(h^2_{h^1_x(1)}(0))\\ =& V^1(h^1_x(1)) = 0,\end{aligned}$$

so $\tilde{h}_x(1) \in \mathcal{X}$ as well. The monotonicity properties of $\tilde{V}(\tilde{h}_x(t))$ and $f(\tilde{h}_x(t))$ are also the direct consequence of the fact that f_i, V^i and h^i_x are completely separated. □

A summary of this subsection has been provided in Table 2.1.

Weak Exactness

One observation from the proof of Theorem 1 is we do not actually need $f(h_x(0)) > f(h_x(1))$ to eliminate genuine local optima. However, such strict inequality is required to show the exactness. We can consider a weaker version of exactness defined as follows.

Definition 12 (Weak Exactness)**.** *We say the relaxation* (2.2) *is* weakly exact *with respect to* (2.1) *if at least one optimum of* (2.2) *is feasible, and hence globally optimal, for* (2.1).

Theorem 4. *If there exists a Lyapunov-like function V associated with* (2.1) *and* (2.2) *such that (C3) and (C2) hold, then* (2.2) *is weakly exact with respect to* (2.1) *and any local optimum in* $\mathcal{X}$ *for* (2.1) *is either a global optimum or a pseudo local optimum.*

The argument on weak exactness follows from the fact that the path connects any global optimum of (2.2) must determine an endpoint in $\mathcal{X}$ with the same cost, which

by definition must be a global optimum as well. The argument on local optimality follows directly from the proof of Theorem 1.

2.6 Applications

In this sections we will use two examples to show for specific problems, what V and $\{h_x\}$ might look like. The first example is Optimal Power Flow (OPF) problem in power systems with tree structres, which is also the motivating problem for us to develop this theory. By finding the Lyapunov-like function and paths, we show the first known condition (that can be checked *a priori*) for OPF to have no spurious local optima. The same condition was only known to guarantee exact relaxation before our work.

In the second example, we study the Low Rank Semidefinite Program (LRSDP) problem, which was known to have weakly exact relaxation [5, 61] and no spurious local optima [17] in existing literatures. Specifically, we show that part of the results proved in [17] can also be proved by finding appropriate V and $\{h_x\}$. They exemplify the usage of Theorem 1, Theorem 2, and Theorem 4 in practice.

Optimal Power Flow

Consider a radial power network with an underlying connected directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Let $\mathcal{V} := \{0, 1, \cdots, N-1\}$ be the set of buses (i.e., nodes), and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ be the set of power lines (i.e., edges). We will refer to a power line from bus j to bus k by $j \to k$ or (j, k) interchangeably. For each power line (j, k), its series admittance is denoted by $y_{jk} \in \mathbb{C}$, and its series impedance is hence $z_{jk} := y_{jk}^{-1}$. Both the real and imaginary parts of z_{jk} are assumed to be positive.

As we assume $\mathcal{G}$ is a tree, we can adopt the DistFlow Model [3, 4] to formulate power flow equations. For each bus j, let $V_j \in \mathbb{C}, s_j = p_j + \boldsymbol{i} q_j \in \mathbb{C}$ denote its voltage and bus injection respectively. For line (j, k), let S_{jk} and $I_{jk} \in \mathbb{C}$ denote the branch power flow and current from bus j to k, both at the sending end. Let $v_j := |V_j|^2 \in \mathbb{R}$ and $\ell_{jk} := |I_{jk}|^2 \in \mathbb{R}$. We will denote the conjugate of a complex number a by a^{H}.

The power flow equations are:

$$v_j = v_k + 2\mathrm{Re}(z_{jk} S_{jk}^{\mathsf{H}}) - |z_{jk}|^2 \ell_{jk}, \qquad \forall (j,k) \in \mathcal{E} \tag{2.8a}$$

$$v_j = \frac{|S_{jk}|^2}{\ell_{jk}}, \qquad \forall (j,k) \in \mathcal{E} \tag{2.8b}$$

$$s_j = \sum_{k:j\to k} S_{jk} - \sum_{i:i\to j} (S_{ij} - z_{ij}\ell_{ij}), \qquad \forall j \in \mathcal{V}. \tag{2.8c}$$

Given a cost function $f(s) : \mathbb{C}^N \to \mathbb{R}$, we are interested in the following OPF problem:

$$\begin{aligned}
\underset{x=(s,v,\ell,S)}{\text{minimize}} \quad & f(s) && \text{(2.9a)}\\
\text{subject to} \quad & (2.8) && \text{(2.9b)}\\
& \underline{v}_j \le v_j \le \overline{v}_j && \text{(2.9c)}\\
& \underline{s}_j \le s_j \le \overline{s}_j && \text{(2.9d)}\\
& \ell_{jk} \le \overline{\ell}_{jk}. && \text{(2.9e)}
\end{aligned}$$

All the inequalities for complex numbers in this section are enforced for both the real and imaginary parts.

Definition 13. *A function $g : \mathbb{R} \to \mathbb{R}$ is* strongly increasing *if there exists real $c > 0$ such that for any $a > b$, we have*

$$g(a) - g(b) \ge c(a - b).$$

We now make the following assumptions on OPF:

(i) The underlying graph $\mathcal{G}$ is a tree.

(ii) The cost function f is convex, and is strongly increasing in $\mathrm{Re}(s_j)$ (or $\mathrm{Im}(s_j)$) for each $j \in \mathcal{V}$ and non-decreasing in $\mathrm{Im}(s_j)$ (or $\mathrm{Re}(s_j)$).

(iii) The problem (2.9) is feasible.

(iv) The line current limit satisfies $\overline{\ell}_{jk} \le \underline{v}_j |y_{jk}|^2$.

Assumption (i) is generally true for distribution networks and assumption (iii) is typically mild. As for (ii), f is commonly assumed to be convex and increasing in $\mathrm{Re}(s_j)$ and $\mathrm{Im}(s_j)$ in the literature (e.g., [31, 74]). Assumption (ii) is only

slightly stronger since one could always perturb any increasing function by an arbitrarily small linear term to achieve strong monotonicity. Assumption (iv) is not common in the literature but is also mild because of the following reason. Typically $V_j = (1+\epsilon_j)e^{\boldsymbol{i}\theta_j}$ in per unit where $\epsilon \in [-0.1, 0.1]$ and the angle difference $\theta_{jk} := \theta_j - \theta_k$ between two neighboring buses j, k typically has a small magnitude. Thus the maximum value of $|V_j - V_k|^2 = |(1+\epsilon_j)e^{\boldsymbol{i}\theta_{jk}} - (1+\epsilon_k)|^2$, which is equivalent to $\overline{\ell}_{jk}/|y_{jk}|^2$, should be much smaller than $\underline{v}_j$ which is ≈ 1 per unit.

Problem (2.9) is non-convex, as constraint (2.8b) is not convex. Denote by $\mathcal{X}$ the set of (s, v, ℓ, S) that satisfy (2.9b)-(2.9e), so (2.9) is in the form of (2.1). We can relax (2.9) by convexifying (2.8b) into a second-order cone [27]:

$$\begin{aligned} \underset{x=(s,v,\ell,S)}{\text{minimize}} \quad & f(s) && \text{(2.10a)} \\ \text{subject to} \quad & \text{(2.8a), (2.8c), (2.9c) – (2.9e)} && \text{(2.10b)} \\ & |S_{jk}|^2 \le v_j \ell_{jk}. && \text{(2.10c)} \end{aligned}$$

One can similarly regard $\hat{\mathcal{X}}$ as the set of (s, v, ℓ, S) that satisfy (2.10b), (2.10c). It is proved in [27] that if $\underline{s}_j = -\infty - \boldsymbol{i}\infty$ for all $j \in \mathcal{V}$, then (2.10) is exact, meaning any optimal solution of (2.10) is also feasible and hence globally optimal for (2.9). Now we show that the same condition also guarantees that any local optimum of (2.9) is also globally optimal. This implies that a local search algorithm such as the primal-dual interior point method can produce a global optimum as long as it converges.

Theorem 5. *If $\underline{s}_j = -\infty - \boldsymbol{i}\infty$ for all $j \in \mathcal{V}$, then any local optimum of* (2.9) *is a global optimum.*

Proof. Our strategy is to construct appropriate V and $\{h_x\}$ and then prove such construction satsify both Condition (C) and Condition (C'). Let

$$V(x) := \sum_{(j,k)\in\mathcal{E}} v_j \ell_{jk} - |S_{jk}|^2. \tag{2.11}$$

Clearly, V is a valid Lyapunov-like function satisfying Definition 10.

For each $x = (s, v, \ell, S) \in \hat{\mathcal{X}} \setminus \mathcal{X}$, let $\mathcal{M}$ be the set of $(j, k) \in \mathcal{E}$ such that $|S_{jk}|^2 < v_j \ell_{jk}$. For $(j, k) \in \mathcal{M}$, the quadratic function

$$\phi_{jk}(a) := \frac{|z_{jk}|^2}{4} a^2 + \big(v_j - \mathrm{Re}(z_{jk} S_{jk}^{\mathsf{H}})\big) a + |S_{jk}|^2 - v_j \ell_{jk}$$

must have a unique positive root as $\phi_{jk}(0) < 0$. We define Δ_{jk} to be this positive root if $(j,k) \in \mathcal{M}$ and 0 otherwise.

Assumption (iv) implies $\ell_{jk} \le v_j |y_{jk}|^2$, and therefore

$$\begin{aligned} v_j - \mathrm{Re}(z_{jk} S_{jk}^{\mathsf{H}}) \;&\ge\; v_j - |z_{jk}||S_{jk}| \\ \ge v_j - |z_{jk}|\sqrt{v_j \ell_{jk}} \;&\ge\; v_j - |z_{jk}|\sqrt{v_j^2 |y_{jk}|^2} \;=\; 0. \end{aligned}$$

It further implies $\phi_{jk}(a)$ is strictly increasing for $a \in [0, \Delta_{jk}]$.

Now consider the path $h_x(t) := (\tilde{s}(t), \tilde{v}(t), \tilde{\ell}(t), \tilde{S}(t))$ for $t \in [0,1]$, where

$$\tilde{s}_j(t) = s_j - \frac{t}{2} \sum_{i:i\to j} z_{ij}\Delta_{ij} - \frac{t}{2} \sum_{k:j\to k} z_{jk}\Delta_{jk}, \tag{2.12a}$$

$$\tilde{v}_j(t) = v_j, \tag{2.12b}$$

$$\tilde{\ell}_{jk}(t) = \ell_{jk} - t\Delta_{jk}, \tag{2.12c}$$

$$\tilde{S}_{jk}(t) = S_{jk} - \frac{t}{2} z_{jk}\Delta_{jk}. \tag{2.12d}$$

Clearly we have $h_x(0) = x$. It can be easily checked that $h_x(t)$ is feasible for (2.10) for $t \in [0,1]$ and $h_x(1)$ is feasible for (2.9). Therefore, h_x is indeed $[0,1] \to \hat{\mathcal{X}}$ and $h_x(1) \in \mathcal{X}$.

Since $z_{jk} > 0$, both real and imaginary parts of $\tilde{s}_j(t)$ are strictly decreasing for $(j,k) \in \mathcal{M}$ and stay unchanged otherwise. By assumption (ii), $f(\tilde{s}(t))$ is also strictly decreasing. To show $V(h_x(t))$ is also decreasing, we notice that $V(h_x(t))$ equals

$$\begin{aligned} &\sum_{(j,k)\in\mathcal{E}} \tilde{v}_j(t)\tilde{\ell}_{jk}(t) - |\tilde{S}_{jk}(t)|^2 \\ =\; &\sum_{(j,k)\in\mathcal{M}^{\mathrm{c}}} v_j \ell_{jk} - |S_{jk}|^2 + \sum_{(j,k)\in\mathcal{M}} \tilde{v}_j(t)\tilde{\ell}_{jk}(t) - |\tilde{S}_{jk}(t)|^2 \\ =\; &\sum_{(j,k)\in\mathcal{M}^{\mathrm{c}}} v_j \ell_{jk} - |S_{jk}|^2 - \sum_{(j,k)\in\mathcal{M}} \phi_{jk}(t\Delta_{jk}). \end{aligned}$$

As $\phi_{jk}(a)$ is strictly increasing for $a \in [0, \Delta_{jk}]$, we conclude that $V(h_x(t))$ is strictly decreasing for $t \in [0,1]$.

By Corollary 1, the set $\{h_x\}_{x\in\hat{\mathcal{X}}\setminus\mathcal{X}}$ is uniformly bounded and uniformly equicontinuous as all $h_x(t)$ are linear functions in t. In summary, Condition (C) is satisfied.

Finally, we show Condition (C') also holds. By assumption (ii), there exists some real $c > 0$ independent of x such that for any $0 \le a < b \le 1$,

$$\begin{aligned} & f(\tilde{s}(a)) - f(\tilde{s}(b)) \\ \ge\ & c \sum_{j \in \mathcal{V}} \mathrm{R}e(\tilde{s}_j(a) - \tilde{s}_j(b)) + \mathrm{I}m(\tilde{s}_j(a) - \tilde{s}_j(b)) \\ =\ & c\|\tilde{s}(a) - \tilde{s}(b)\|_{\mathrm{m}}. \end{aligned}$$

where $\|\cdot\|_{\mathrm{m}}$ is defined as $\|\mathbf{a}\|_{\mathrm{m}} := \sum_i |\mathrm{R}e(a_i)| + |\mathrm{I}m(a_i)|$ over the complex vector space. It is easy to check $\|\cdot\|_{\mathrm{m}}$ is a valid norm.

On the other hand, by (2.12) we have $\|\tilde{v}(a) - \tilde{v}(b)\|_{\mathrm{m}} \equiv 0$ and

$$\begin{aligned} \|\tilde{\ell}(a) - \tilde{\ell}(b)\|_{\mathrm{m}} &\le \frac{1}{\min\limits_{(j,k)\in\mathcal{E}} \{\|z_{jk}\|_{\mathrm{m}}\}} \|\tilde{s}(a) - \tilde{s}(b)\|_{\mathrm{m}}, \\ \|\tilde{S}(a) - \tilde{S}(b)\|_{\mathrm{m}} &\le \frac{1}{2}\|\tilde{s}(a) - \tilde{s}(b)\|_{\mathrm{m}}. \end{aligned}$$

Therefore,

$$\|h_x(a) - h_x(b)\|_{\mathrm{m}} \le \Big(\frac{3}{2} + \frac{1}{\min\limits_{(j,k)\in\mathcal{E}} \{\|z_{jk}\|_{\mathrm{m}}\}}\Big)\|\tilde{s}(a) - \tilde{s}(b)\|_{\mathrm{m}}$$

and there exists $\hat{c} > 0$ independent of x, a, b such that

$$f(\tilde{s}(a)) - f(\tilde{s}(b)) \ge \hat{c}\|h_x(a) - h_x(b)\|_{\mathrm{m}}.$$

Therefore Condition (C') is also satisfied, and by Theorem 2, any local optimum of (2.9) is a global optimum. □

The results in this subsection only apply to radial networks, which serve as the underlying network for balanced distribution power systems. For transmission systems and unbalanced distribution systems, networks are usually highly meshed. It has been found that for most of meshed networks, both convex relaxation and local search algorithms can also yield the global optimum for most of testcases [34, 41]. Thus Theorem 3 suggests that there may also exist similar Lyapunov-like function and paths for meshed networks. Finding such Lyapunov-like function and paths would be an interesting future work to extend our results in this chapter.

Low Rank Semidefinite Program

This subsection proves a known result in [17] but using a different approach. Adopting the same notations as [17], we have the following problem.

$$\begin{aligned} \underset{X\geq 0}{\text{minimize}} \quad & \operatorname{tr}(CX) && \text{(2.13a)} \\ \text{subject to} \quad & \operatorname{tr}(A_i X) = b_i, \quad i = 1, \cdots, m && \text{(2.13b)} \\ & \operatorname{rank}(X) \leq r. && \text{(2.13c)} \end{aligned}$$

Here, C, A_i, X are all n-by-n matrices. We assume the problem is feasible and $\{X \geq 0 \mid (2.13b)\}$ is compact.

Theorem 6. *If* $(r+1)(r+2)/2 > m+1$*, then any local optimum of* (2.13) *is either a global optimum or a pseudo local optimum.*

Before proving Theorem 6, we consider the convex relaxation of (2.13) as

$$\begin{aligned} \underset{X\geq 0}{\text{minimize}} \quad & \operatorname{tr}(CX) && \text{(2.14a)} \\ \text{subject to} \quad & \operatorname{tr}(A_i X) = b_i, \quad i = 1, \cdots, m. && \text{(2.14b)} \end{aligned}$$

As a side note, the results in [5, 61] show that if $(r+1)(r+2)/2 > m$, then (2.14) is weakly exact to (2.13). While our theorem is the same as in [17], some insights to find V and $\{h_X\}$ are also from the structures first raised in [5, 61].

Proof. Clearly, (2.13) can be reformulated in the form of (2.1) by setting $f(X) = \operatorname{tr}(CX)$, $\mathcal{X} = \{X \geq 0 \mid (2.13b), (2.13c)\}$ and $\hat{\mathcal{X}} = \{X \geq 0 \mid (2.13b)\}$. Define V as

$$V(X) := \sum_{i=r+1}^{n} \lambda_i(X),$$

where $\lambda_i(X)$ is the i^{th} eigenvalue of X (in decreasing order). This function V satisfies Definition 10 and is concave.

For fixed $X \in \hat{\mathcal{X}} \setminus \mathcal{X}$, we denote $\operatorname{rank}(X)$ as $r_0 > r$. We first construct $r_0 - r$ paths labeled as $h_1, h_2, \cdots, h_{r_0-r}$. When we construct h_i, if $i > 1$ then we assume path h_{i-1} has already been constructed and let $X_{i-1} := h_{i-1}(1)$. We let $X_0 = X$. For $i \geq 1$, if $\operatorname{rank}(X_{i-1}) \leq r_0 - i$ then we let $h_i(t) \equiv X_{i-1}$ for $t \in [0, 1]$. Otherwise, we decompose X_{i-1} as $U\Sigma U^{\mathsf{H}}$ where Σ is a k-by-k positive definite diagonal matrix with $k = \operatorname{rank}(X_{i-1}) > r_0 - i$. The linear system

$$\begin{aligned} \operatorname{tr}(CUYU^{\mathsf{H}}) &= 0 \\ \operatorname{tr}(A_i UYU^{\mathsf{H}}) &= 0, \ i = 1, \cdots, m \end{aligned} \tag{2.15}$$

must have a non-zero solution for Hermitian matrix $Y \in \mathbb{C}^{k\times k}$. To see this, we have $k \geq r_0 - i + 1 \geq r + 1$, and thus $k(k+1)/2 \geq (r+1)(r+2)/2 > m + 1$. As a result, (2.15) has more unknown variables than equations. We simply denote this non-zero solution as Y and for any $\alpha \in \mathbb{R}$, αY is also a solution to (2.15). The concavity of V also implies that $V(U(\Sigma + \alpha Y)U^{\mathsf{H}})$ is concave in α when U and Σ are fixed. Since $\Sigma > 0$, one of the following two scenarios must be true.

- $\exists a < 0$ such that $V(U(\Sigma+\alpha Y)U^{\mathsf{H}})$ is non-decreasing, $\operatorname{rank}(U(\Sigma+\alpha Y)U^{\mathsf{H}}) \leq k$ for $\alpha \in [a, 0]$ and $\operatorname{rank}(U(\Sigma + aY)U^{\mathsf{H}}) \leq k - 1$.
- $\exists b > 0$ such that $V(U(\Sigma+\alpha Y)U^{\mathsf{H}})$ is non-increasing, $\operatorname{rank}(U(\Sigma+\alpha Y)U^{\mathsf{H}}) \leq k$ for $\alpha \in [0, b]$ and $\operatorname{rank}(U(\Sigma + bY)U^{\mathsf{H}}) \leq k - 1$.

Without loss of generality, we suppose $V(U(\Sigma + \alpha Y)U^{\mathsf{H}})$ is non-increasing for $\alpha \in [0, b]$ (otherwise we take $-Y$ instead). We then construct h_i as $h_i(t) = U(\Sigma+tbY)U^{\mathsf{H}}$ for $t \in [0, 1]$. By construction, $V(h_i(t))$ is non-increasing and $f(h_i(t))$ stays a constant.

Finally, we construct h_X as the concatenation of paths $h_1, \cdots, h_{r_0-r}$. That is to say,

$$h_X(t) := h_i((r_0 - r)t - i + 1) \text{ for } t \in \left[\frac{i-1}{r_0-r}, \frac{i}{r_0-r}\right].$$

It is easy to see h_X is continuous and $h_X(0) = h_1(0) = X$. To see $h_X(1) \in \mathcal{X}$, we prove that $\operatorname{rank}(X_i) \leq r_0 - i$. We first have $\operatorname{rank}(X_0) = \operatorname{rank}(X) = r_0$. For $i \geq 1$, we have $\operatorname{rank}(X_i) = \operatorname{rank}(X_{i-1})$ if $\operatorname{rank}(X_{i-1}) \leq r_0 - i$ and $\operatorname{rank}(X_i) \leq \operatorname{rank}(X_{i-1}) - 1$ otherwise. By induction, we can prove $\operatorname{rank}(X_i) \leq r_0 - i$ always holds. As a result, $\operatorname{rank}(h_X(1)) = \operatorname{rank}(h_{r_0-r}(1)) \leq r$ and thus $h_X(1) \in \mathcal{X}$. By construction, $h_i(t)$ never violates (2.13b) and thus is in $\hat{\mathcal{X}}$, so is $h_X(t)$ for all t. Functions $V(h_i(t))$ and $f(h_i(t))$ being non-increasing implies that $V(h_X(t))$ and $f(h_X(t))$ are also non-increasing. Therefore, (C3) is satisfied. By Corollary 1 (C2) also holds for $\{\overline{h_X}\}$. It completes the proof (by Theorem 4). □

Remark 4. *In [17], Theorem 3.4 claims that any local optimum of* (2.13) *should also be globally optimal, unless it is harbored in some positive-dimensional face of SDP. The result in this chapter further asserts that if it is indeed harbored in such a face, then there must be some point on the edge of this face whose cost can be further reduced in its neighborhood (i.e., the local optimum is in the same situation as point c rather than d as in Fig. 2.1).*

Table 2.2: Sufficient and necessary conditions

Condition	Relaxation exactness	Local optimality
Sufficient conditions: $\Rightarrow$		
(C1), (C2)	Strong exactness	l.o. is p.l.o. or g.o.
(C3), (C2)	Weak exactness	l.o. is p.l.o. or g.o.
(C')	Strong exactness	l.o. is g.o.
Necessary condition: $\Leftarrow$		
(C1), (C2)	Strong exactness	l.o. is g.o.

2.7 Conclusion and Discussion

Table 2.2 summaries both sufficient and necessary conditions for non-convex problem (2.1) to simultaneously have exact (weak or strong) relaxation and no spurious local optima (allowing or not allowing pseudo local optima). The necessary condition relies on Assumption 1, which is usually true for real-world problems. Those results provide a new perspective to certify a non-convex problem is computationally easy to solve. Furthermore, whenever the problem is indeed computationally easy, the certificates (Lyapunov-like functions and paths) are guaranteed to exist. We also provide a hierarchical framework which shows how such certificates for a complicated problem can be constructed from primitive problems. Our results have been applied to OPF and LRSDP problems.

Based on the examples shown in Section 2.6, a natural way to apply this approach is to first look at existing results on exact relaxation, and then construct V and $\{h_x\}$ according to the hidden structure underlying the exactness. Once V and $\{h_x\}$ are appropriately constructed, our result can help extend existing results on relaxation exactness to new results on local optimality.

Compared to some existing techniques to study local optimality, our results do not require differentiating or analyzing the curvature of feasible sets. It allows the feasible sets to incorporate more complicated and possibly non-convex constraints. Those non-convex constraints are common for problems arising in cyber physical systems which are generally governed by physical laws.

Chapter 3

RELAXATION EXACTNESS FOR MULTI-PHASE NETWORKS: WITHOUT DELTA CONNECTIONS

3.1 Background

In previous chapters, we have introduced and studied the Optimal Power Flow problems in single-phase networks. The single-phase model is widely used to describe both single-phase and balanced multi-phase networks [2, 40]. Most existing results on semi-definite relaxation [15, 28, 31, 45, 47, 67, 74] are also based on such single-phase model and assume that the underlying network topology is a tree.

Most radial distribution networks are, however, unbalanced multiphase, e.g., [43, 68]. SDP relaxation has recently been applied to unbalanced multiphase radial networks [24, 30, 73, 76]. Simulation results in these papers suggest that SDP relaxation is often exact even though no sufficient condition for exact relaxation is known to the best of our knowledge. Indeed, it has been observed in [7, 22, 44] that a multiphase unbalanced network has an equivalent single-phase circuit model where each bus-phase pair in the multiphase network is identified with a single bus in the equivalent model. The single-phase equivalent model is then a meshed network and therefore existing guarantees on exact SDP relaxation are not applicable. Most distribution systems are unbalanced multiphase networks [23] and hence the performance of SDP relaxation of OPF on these networks is important.

In this chapter, we generalize the sufficient conditions for single-phase network proposed in [15] to the multiphase setting. The result shows that the exactness of the primal problem can be guaranteed if its dual variable is $\mathcal{G}$-invertible (to be defined later). Informally speaking, it requires every block matrix in the dual variable that correspond to an edge in the network to be invertible. Then we provide two perspectives to study this $\mathcal{G}$-invertibility property. One perspective requires the binding constraints in the primal problem to be sparse, and the other perspective relates to the condition that the marginal prices of the power should span over a narrow range.

3.2 System Model

Network Structure

We use a similar model as in [30, 76].We assume that all buses have the same number of phases and all generations and loads are wye connected. Let the underlying simple undirected graph be $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{0, 1, \ldots, n-1\}$ denotes the set of buses and $\mathcal{E}$ the set of edges. Throughout this chapter, we will use (graph, vertex, edge) and (power network, bus, line) interchangeably. Without loss of generality, we let bus 0 be the slack bus where the voltage is specified. Assume all buses have m phases for $m \in \mathbb{Z}^+$. We will use (j, k) and $j \sim k$ interchangeably to denote an edge connecting bus j and k. Consider an m-phase line (j, k) characterized by the admittance matrix $y_{jk} \in \mathbb{C}^{m\times m}$, we assume y_{jk} is invertible. The admittance matrix $\boldsymbol{Y} \in \mathbb{C}^{mn\times mn}$ for the entire network can be divided into $n \times n$ number of $m \times m$ block matrices. Let $\boldsymbol{Y}_{jk} \in \mathbb{C}^{m\times m}$ denote the block matrix corresponding to the admittance between bus j and k, and then we have

$$\boldsymbol{Y}_{jj} = \sum_{k:j\sim k} y_{jk},\ j \in \mathcal{V}$$
$$\boldsymbol{Y}_{jk} = \begin{cases} -y_{jk} & , j \sim k \\ 0 & , j \nsim k \end{cases}.$$

For each bus j, let the voltages of all m phases at bus j be the vector $\boldsymbol{V}_j \in \mathbb{C}^m$. We use $\boldsymbol{V}_j^\phi$ for $\phi \in \mathcal{M} := \{1, 2, \ldots, m\}$ to indicate the voltage for phase ϕ. Let $\boldsymbol{V} = [\boldsymbol{V}_0^\mathsf{T}, \boldsymbol{V}_1^\mathsf{T}, \ldots, \boldsymbol{V}_{n-1}^\mathsf{T}]^\mathsf{T}$ be the voltage vector for the entire network. Similarly, we use s_j^ϕ to denote the bus injection for phase ϕ at bus j. We will refer to the ϕ $^{\text{th}}$ diagonal entry of y_{jk} as y_{jk}^ϕ. Let $\boldsymbol{e}_j^\phi \in \mathbb{R}^{mn}$ be the base vector which has 1 at the $(jm + \phi)^{\text{th}}$ entry and 0 elsewhere. Let $\boldsymbol{E}_j^\phi = \boldsymbol{e}_j^\phi(\boldsymbol{e}_j^\phi)^\mathsf{T}$, then we define

$$Y_j^\phi := \boldsymbol{E}_j^\phi \boldsymbol{Y} \in \mathbb{C}^{mn\times mn}$$

and

$$\boldsymbol{\Phi}_j^\phi := \frac{1}{2}\big((Y_j^\phi)^\mathsf{H} + Y_j^\phi\big)$$
$$\boldsymbol{\Psi}_j^\phi := \frac{1}{2\boldsymbol{i}}\big((Y_j^\phi)^\mathsf{H} - Y_j^\phi\big)$$
$$\boldsymbol{\Theta}_{jk}^\phi := (y_{jk}^\phi)^2(\boldsymbol{e}_j^\phi - \boldsymbol{e}_k^\phi)(\boldsymbol{e}_j^\phi - \boldsymbol{e}_k^\phi)^\mathsf{T}.$$

All $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Theta}$ are Hermitian matrices. The relationship between bus voltages

and injections can be expressed as

$$\begin{aligned} \mathrm{Re}(s_j^\phi) &= \boldsymbol{V}^{\mathsf{H}} \boldsymbol{\Phi}_j^\phi \boldsymbol{V}, \\ \mathrm{Im}(s_j^\phi) &= \boldsymbol{V}^{\mathsf{H}} \boldsymbol{\Psi}_j^\phi \boldsymbol{V} \end{aligned} \tag{3.1}$$

and the squared current is

$$\ell_{jk}^\phi = \boldsymbol{V}^{\mathsf{H}} \boldsymbol{\Theta}_{jk}^\phi \boldsymbol{V}.$$

Optimal Power Flow

Optimal power flow problems minimize certain cost functions subject to constraints involving voltages, injections, and currents. Here we consider problems that take the linear combination of bus injections as the cost function and are subject to operational constraints for voltage magnitudes, real/reactive injections, and line currents. For problems with nonlinear cost functions, see Section 3.7. Suppose the bounds $\underline{V}$ and $\overline{V}$ for the voltage magnitudes are always positive and finite, but the bounds for real/reactive injections can be $\pm\infty$ if there are no such constraints.

$$\begin{aligned}
\underset{\boldsymbol{V},s}{\text{minimize}} \quad & \sum_{j,\phi} c_{j,\mathrm{re}}^\phi \mathrm{Re}(s_j^\phi) + c_{j,\mathrm{im}}^\phi \mathrm{Im}(s_j^\phi) && \text{(3.2a)} \\
\text{subject to} \quad & (3.1) && \text{(3.2b)} \\
& \underline{V}_j^\phi \le |\boldsymbol{V}_j^\phi| \le \overline{V}_j^\phi, \quad \forall j, \phi && \text{(3.2c)} \\
& \underline{p}_j^\phi \le \mathrm{Re}(s_j^\phi) \le \overline{p}_j^\phi, \quad \forall j, \phi && \text{(3.2d)} \\
& \underline{q}_j^\phi \le \mathrm{Im}(s_j^\phi) \le \overline{q}_j^\phi, \quad \forall j, \phi && \text{(3.2e)} \\
& \ell_{jk}^\phi \le \overline{\ell}_{jk}^\phi, \quad \forall j \sim k, \phi && \text{(3.2f)} \\
& \boldsymbol{V}_0 = \boldsymbol{V}_{\mathrm{ref}} && \text{(3.2g)}
\end{aligned}$$

Here, $\boldsymbol{V}_{\mathrm{ref}} \in \mathbb{C}^m$ denotes the reference voltage for m phases at the slack bus. Substituting the decision variables s and $\boldsymbol{V}$ with $\boldsymbol{W} := \boldsymbol{V}\boldsymbol{V}^{\mathsf{H}}$, the following equivalent formulation of (3.2) is obtained:

$$\begin{aligned}
\underset{\boldsymbol{W} \succeq 0}{\text{minimize}} \quad & \mathrm{tr}(\boldsymbol{C}_0 \boldsymbol{W}) && \text{(3.3a)} \\
\text{subject to} \quad & \underline{v}_j^\phi \le \mathrm{tr}(\boldsymbol{E}_j^\phi \boldsymbol{W}) \le \overline{v}_j^\phi, \quad \forall j, \phi && \text{(3.3b)} \\
& \underline{p}_j^\phi \le \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}) \le \overline{p}_j^\phi, \quad \forall j, \phi && \text{(3.3c)} \\
& \underline{q}_j^\phi \le \mathrm{tr}(\boldsymbol{\Psi}_j^\phi \boldsymbol{W}) \le \overline{q}_j^\phi, \quad \forall j, \phi && \text{(3.3d)} \\
& \mathrm{tr}(\boldsymbol{\Theta}_{jk}^\phi \boldsymbol{W}) \le \overline{\ell}_{jk}^\phi, \quad \forall j \sim k, \phi && \text{(3.3e)} \\
& [\boldsymbol{W}]_{00} = \boldsymbol{v}_{\mathrm{ref}} && \text{(3.3f)} \\
& \mathrm{rank}(\boldsymbol{W}) = 1. && \text{(3.3g)}
\end{aligned}$$

Here, $\underline{v}_j^\phi = |\underline{V}_j^\phi|^2$, $\overline{v}_j^\phi = |\overline{V}_j^\phi|^2$, $\boldsymbol{v}_{\text{ref}} = \boldsymbol{V}_{\text{ref}}\boldsymbol{V}_{\text{ref}}^{\mathsf{H}}$, and $[\boldsymbol{W}]_{00}$ stands for the upper left $m \times m$ submatrix of $\boldsymbol{W}$. The cost matrix $\boldsymbol{C}_0 = \sum_{j,\phi} c_{j,\text{re}}^\phi \boldsymbol{\Phi}_j^\phi + c_{j,\text{im}}^\phi \boldsymbol{\Psi}_j^\phi$. Dropping the rank-1 constraint in (3.3g) yields the semidefinite relaxation:

$$\begin{aligned} \underset{\boldsymbol{W}\geq 0}{\text{minimize}} \quad & \text{tr}(\boldsymbol{C}_0\boldsymbol{W}) && \text{(3.4a)} \\ \text{subject to} \quad & \text{(3.3b)} - \text{(3.3f)}. && \text{(3.4b)} \end{aligned}$$

We use the following exactness definition.

Definition 14. *A relaxation problem* (3.4) *is exact if at least one of its optimal solutions* $\boldsymbol{W}^*$ *is of rank 1.*

Given a rank-1 solution $\boldsymbol{W}^*$ of (3.4), a $\boldsymbol{V}^*$ can be uniquely determined, which is feasible, and hence optimal, for (3.3).

We first make the assumption that (3.4) has a unique optimal solution. In Section 3.7, we discuss the case when multiple optimal solutions exist.

3.3 Perturbation Analysis

We first study a perturbed version of (3.4). Fix a nonzero Hermitian matrix $\boldsymbol{C}_1$, and consider the following perturbed problem for $\varepsilon \geq 0$:

$$\begin{aligned} \underset{\boldsymbol{W}\geq 0}{\text{minimize}} \quad & \text{tr}((\boldsymbol{C}_0 + \varepsilon\boldsymbol{C}_1)\boldsymbol{W}) && \text{(3.5a)} \\ \text{subject to} \quad & \text{(3.3b)} - \text{(3.3f)}. && \text{(3.5b)} \end{aligned}$$

We say that (3.5) is *exact* if one of its optimal solution is of rank 1.

Lemma 12. *For any nonzero* $\boldsymbol{C}_1$*, if there exists a sequence* $\{\varepsilon_l\}_{l=1}^\infty$ *with* $\lim_{l\to\infty} \varepsilon_l = 0$ *such that* (3.5) *is exact for all* ε_l*, then* (3.4) *is exact.*

Proof. Suppose the rank-1 optimal solution to (3.5) for ε_l is $\boldsymbol{W}_l$. If the rank-1 optimal solution is non-unique, then pick any one as $\boldsymbol{W}_l$. As all the $\overline{v}_j^\phi$ are finite, we assume they are upper bounded by a constant α. Hence the constraint (3.3b) implies all the diagonal elements of $\boldsymbol{W}$ are upper bounded by α. Since $\boldsymbol{W}$ is positive semidefinite, the norms of all their entries can be upper bounded by α as well. Consider the set

$$\mathcal{S} = \{\boldsymbol{W} \geq 0 : \text{(3.3b)} - \text{(3.3g)}\}. \tag{3.6}$$

The set $\{\boldsymbol{W} : \text{rank}(\boldsymbol{W}) \leq 1\}$ is closed [38] and all other constraints (3.3b)-(3.3f) also prescribe closed sets. Further, the zero matrix is not in $\mathcal{S}$ and we have shown

that for any $\boldsymbol{W} \in \mathcal{S}$, its max norm must be upper bounded by α, so $\mathcal{S}$ is compact. The infinite set $\{\boldsymbol{W}_l\}_{l=1}^{\infty}$ is a subset in $\mathcal{S}$ and hence has a limit point $\boldsymbol{W}_{\lim} \in \mathcal{S}$ [64]. For any ε_l, (3.5) has the same feasible set as (3.4), and hence the rank-1 matrix $\boldsymbol{W}_{\lim}$ is also feasible for (3.4). Next we show that $\boldsymbol{W}_{\lim}$ is also an optimal point for (3.4).

If there exists another feasible $\boldsymbol{W}_{\text{opt}} \neq \boldsymbol{W}_{\lim}$ such that $\text{tr}(\boldsymbol{C}_0\boldsymbol{W}_{\lim}) - \text{tr}(\boldsymbol{C}_0\boldsymbol{W}_{\text{opt}}) = \nu > 0$. Clearly $\forall \boldsymbol{W}$ feasible for (3.4), $|\text{tr}(\boldsymbol{C}_1\boldsymbol{W})| \leq m^2n^2\|\boldsymbol{C}_1\|_\infty\|\boldsymbol{W}\|_\infty \leq m^2n^2\alpha\|\boldsymbol{C}_1\|_\infty$. For sufficiently large l such that

$$\begin{aligned} \varepsilon_l &< \frac{\nu}{4m^2n^2\alpha\|\boldsymbol{C}_1\|_\infty} \\ \|\boldsymbol{W}_l - \boldsymbol{W}_{\lim}\|_\infty &< \frac{\nu}{4m^2n^2\|\boldsymbol{C}_0\|_\infty}, \end{aligned}$$

we have

$$\begin{aligned} \text{tr}(\boldsymbol{C}_0(\boldsymbol{W}_l - \boldsymbol{W}_{\lim})) &\geq -\frac{\nu}{4} && \text{(3.7a)} \\ \text{tr}(\varepsilon_l\boldsymbol{C}_1\boldsymbol{W}_l) &\geq -\frac{\nu}{4} && \text{(3.7b)} \\ \text{tr}(\boldsymbol{C}_0\boldsymbol{W}_{\lim}) &= \text{tr}(\boldsymbol{C}_0\boldsymbol{W}_{\text{opt}}) + \nu && \text{(3.7c)} \\ \frac{\nu}{4} &\geq \text{tr}(\varepsilon_l\boldsymbol{C}_1\boldsymbol{W}_{\text{opt}}). && \text{(3.7d)} \end{aligned}$$

Summing up (3.7a)-(3.7d) gives

$$\text{tr}((\boldsymbol{C}_0 + \varepsilon_l\boldsymbol{C}_1)\boldsymbol{W}_l) > \text{tr}((\boldsymbol{C}_0 + \varepsilon_l\boldsymbol{C}_1)\boldsymbol{W}_{\text{opt}}),$$

contradicting the optimality of $\boldsymbol{W}_l$ for ε_l. □

3.4 Duality

The dual problem of (3.4) is as follows.

$$\begin{aligned} \underset{\substack{\overline{\lambda}_j^\phi, \underline{\lambda}_j^\phi, \overline{\mu}_j^\phi, \underline{\mu}_j^\phi, \\ \overline{\eta}_j^\phi, \underline{\eta}_j^\phi, \nu_{jk}^\phi, \kappa}}{\text{maximize}} \quad & -\sum_{j,\phi}(\overline{\lambda}_j^\phi\overline{v}_j^\phi - \underline{\lambda}_j^\phi\underline{v}_j^\phi + \overline{\mu}_j^\phi\overline{p}_j^\phi - \underline{\mu}_j^\phi\underline{p}_j^\phi + \overline{\eta}_j^\phi\overline{q}_j^\phi - \underline{\eta}_j^\phi\underline{q}_j^\phi) \\ & + \sum_{j\sim k,\phi} \nu_{jk}^\phi\overline{\ell}_{jk}^\phi + \text{tr}(\kappa\boldsymbol{\nu}_{\text{ref}}) && \text{(3.8a)} \\ \text{subject to} \quad & \overline{\lambda}_j^\phi, \underline{\lambda}_j^\phi, \overline{\mu}_j^\phi, \underline{\mu}_j^\phi, \overline{\eta}_j^\phi, \underline{\eta}_j^\phi, \nu_{jk}^\phi \geq 0 && \text{(3.8b)} \\ & \boldsymbol{A} \geq 0. && \text{(3.8c)} \end{aligned}$$

Dual variables $(\overline{\lambda}_j^\phi, \underline{\lambda}_j^\phi)$, $(\overline{\mu}_j^\phi, \underline{\mu}_j^\phi)$, $(\overline{\eta}_j^\phi, \underline{\eta}_j^\phi)$, ν_{jk}^ϕ and κ correspond to (3.3b)-(3.3f) in (3.4b), respectively. Specifically $\kappa \in \mathbb{C}^{m\times m}$ is Hermitian but not necessarily

semidefinite positive. Here

$$\boldsymbol{A} := \sum_{j,\phi} (\overline{\lambda}_j^\phi - \underline{\lambda}_j^\phi)\boldsymbol{E}_j^\phi + (\overline{\mu}_j^\phi - \underline{\mu}_j^\phi)\boldsymbol{\Phi}_j^\phi + (\overline{\eta}_j^\phi - \underline{\eta}_j^\phi)\boldsymbol{\Psi}_j^\phi + \sum_{j\sim k,\phi} \nu_{jk}^\phi \boldsymbol{\Theta}_{jk}^\phi + \boldsymbol{C}_0 + \Pi(\kappa). \tag{3.9}$$

Matrix $\Pi(\kappa)$ is an $mn \times mn$ matrix whose upper left $m \times m$ block is κ and other elements are 0. Note that the upper and lower bounds in (3.3c)-(3.3e) could take values of $\pm\infty$. In this case however, since the feasible set prescribed by (3.5b) is compact, the actual values of $\boldsymbol{\Phi}_j^\phi \boldsymbol{W}$ and $\boldsymbol{\Psi}_j^\phi \boldsymbol{W}$ are always finite and hence the dual variables associated with such constraints will be 0. As a result, these constraints can be removed from (3.5) and (3.8).

We can also define the dual problem of (3.5) as

$$\begin{aligned}
\underset{\substack{\overline{\lambda}_j^\phi, \underline{\lambda}_j^\phi, \overline{\mu}_j^\phi, \underline{\mu}_j^\phi, \\ \overline{\eta}_j^\phi, \underline{\eta}_j^\phi, \nu_{jk}^\phi, \kappa}}{\text{maximize}} \quad & -\sum_{j,\phi} (\overline{\lambda}_j^\phi \overline{v}_j^\phi - \underline{\lambda}_j^\phi \underline{v}_j^\phi + \overline{\mu}_j^\phi \overline{p}_j^\phi - \underline{\mu}_j^\phi \underline{p}_j^\phi + \overline{\eta}_j^\phi \overline{q}_j^\phi - \underline{\eta}_j^\phi \underline{q}_j^\phi) \\
& - \sum_{j\sim k,\phi} \nu_{jk}^\phi \overline{\ell}_{jk}^\phi + \operatorname{tr}(\kappa \boldsymbol{v}_{\text{ref}}) && \text{(3.10a)} \\
\text{subject to} \quad & \overline{\lambda}_j^\phi, \underline{\lambda}_j^\phi, \overline{\mu}_j^\phi, \underline{\mu}_j^\phi, \overline{\eta}_j^\phi, \underline{\eta}_j^\phi, \nu_{jk}^\phi \geq 0 && \text{(3.10b)} \\
& \boldsymbol{A}(\varepsilon) \geq 0, && \text{(3.10c)}
\end{aligned}$$

where

$$\boldsymbol{A}(\varepsilon) := \boldsymbol{A} + \varepsilon \boldsymbol{C}_1. \tag{3.11}$$

We will use $\overline{\lambda}_j^\phi(\varepsilon), \underline{\lambda}_j^\phi(\varepsilon)$ and so on to denote the Lagrange multipliers for ε. Clearly, when $\varepsilon = 0$, (3.10) is the same as (3.8) with $\overline{\lambda}_j^\phi(0), \underline{\lambda}_j^\phi(0)$ and so on as the Lagrange multipliers. If the value of ε is clear in the context, we might denote them simply as $\overline{\lambda}_j^\phi, \underline{\lambda}_j^\phi$ and so on for convenience. Let $\boldsymbol{A}^*$ and $\boldsymbol{A}^*(\varepsilon)$ be the dual matrices when dual variables are evaluated at a KKT point for (3.8) and (3.10), respectively.

Note that (3.8) and (3.10) are always strictly feasible, as we can assign sufficiently large values to the diagonal entries of $\boldsymbol{A}$ and $\boldsymbol{A}(\varepsilon)$. Therefore, strong duality always holds for both (3.4), (3.8) and (3.5), (3.10). KKT conditions hold as well.

Definition 15. *An $mn \times mn$ positive semidefinite matrix $\boldsymbol{X}$ is $\mathcal{G}$-invertible for some graph $\mathcal{G}$ if the following two conditions hold:*

1. $\forall (a, b) \in \mathcal{E}$, $[\boldsymbol{X}]_{ab}$ *is invertible.*

2. $\forall a, b \in \mathcal{V}$ *such that* $a \neq b$ *and* $(a, b) \notin \mathcal{E}$, $[\boldsymbol{X}]_{ab}$ *is all zero.*

The next theorem is a generalization of Theorem 3.3 in [37]. While [37] studies the matrices whose non-zero off-diagonal entries correspond to an edge in $\mathcal{G}$, we extend the results to $\mathcal{G}$-invertible matrices.

Theorem 7. *Let* $\boldsymbol{y} \in \mathbb{C}^{mn}$ *be a non-zero vector with the smallest* $|\Omega(\boldsymbol{y})|$ *satisfying* $\boldsymbol{X}\boldsymbol{y} = \boldsymbol{0}$*, where* $\boldsymbol{X}$ *is* $\mathcal{G}$*-invertible. Then* $\Omega(\boldsymbol{y})$ *is connected in* $\mathcal{G}$*.*

Proof. If not, then assume $\Omega(\boldsymbol{y}) = \Omega_1 \cup \Omega_2$ where non-empty sets Ω_1 and Ω_2 are not connected in $\mathcal{G}$. Construct $\tilde{\boldsymbol{y}}$ in the following manner:

$$[\tilde{\boldsymbol{y}}]_k = \begin{cases} [\boldsymbol{y}]_k & , k \notin \Omega_2 \\ \boldsymbol{0} & , k \in \Omega_2 \end{cases}.$$

Then for each $j \in \Omega_1$,

$$\begin{aligned}[\boldsymbol{X}\tilde{\boldsymbol{y}}]_j = \sum_{k \in \mathcal{V}} [\boldsymbol{X}]_{jk}[\tilde{\boldsymbol{y}}]_k = [\boldsymbol{X}]_{jj}[\tilde{\boldsymbol{y}}]_j + \sum_{k:k\sim j} [\boldsymbol{X}]_{jk}[\tilde{\boldsymbol{y}}]_k \\ = [\boldsymbol{X}]_{jj}[\boldsymbol{y}]_j + \sum_{k:k\sim j} [\boldsymbol{X}]_{jk}[\boldsymbol{y}]_k = [\boldsymbol{X}\boldsymbol{y}]_j = \boldsymbol{0}.\end{aligned}$$

The third equality above is due to the fact that $j \in \Omega_1$ is not connected to any nodes in Ω_2. Therefore,

$$\begin{aligned}\tilde{\boldsymbol{y}}^{\mathsf{H}}\boldsymbol{X}\tilde{\boldsymbol{y}} = \sum_{j \in \mathcal{V}} [\tilde{\boldsymbol{y}}]_j^{\mathsf{H}}[\boldsymbol{X}\tilde{\boldsymbol{y}}]_j = \sum_{j \in \Omega_1} [\tilde{\boldsymbol{y}}]_j^{\mathsf{H}}[\boldsymbol{X}\tilde{\boldsymbol{y}}]_j + \sum_{j \notin \Omega_1} [\tilde{\boldsymbol{y}}]_j^{\mathsf{H}}[\boldsymbol{X}\tilde{\boldsymbol{y}}]_j \\ = \sum_{j \in \Omega_1} [\tilde{\boldsymbol{y}}]_j^{\mathsf{H}}\boldsymbol{0} + \sum_{j \notin \Omega_1} \boldsymbol{0}^{\mathsf{H}}[\boldsymbol{X}\tilde{\boldsymbol{y}}]_j = \boldsymbol{0}.\end{aligned}$$

Since $\mathcal{G}$-invertibility implies $\boldsymbol{X} \geq 0$, there must be $\boldsymbol{X}\tilde{\boldsymbol{y}} = \boldsymbol{0}$ as well. As $|\Omega(\tilde{\boldsymbol{y}})| = |\Omega_1| < |\Omega(\boldsymbol{y})|$ and $\tilde{\boldsymbol{y}}$ is non-zero by construction, it contradicts the minimality of $|\Omega(\boldsymbol{y})|$. □

Theorem 8. *Let* $(\boldsymbol{W}^*, \boldsymbol{A}^*)$ *be a pair of primal/dual solutions to* (3.4) *and* (3.8)*. If* $\boldsymbol{A}^*$ *is* $\mathcal{G}$*-invertible,* $\boldsymbol{W}^*$ *must be rank* 1*.*

Proof. Otherwise, we should have $\operatorname{rank}(\boldsymbol{W}^*) \geq 2$. [1] Suppose the eigen-decomposition of $\boldsymbol{W}^*$ is

$$\boldsymbol{W}^* = \sum_{l=1}^{mn} \varrho_l \boldsymbol{u}_l \boldsymbol{u}_l^{\mathsf{H}},$$

[1] Note that $\operatorname{rank}(\boldsymbol{W}^*)$ cannot be 0 as the constraint $[\boldsymbol{W}^*]_{00} = \boldsymbol{v}_{\text{ref}}$ requires $\boldsymbol{W}^*$ to be a non-zero matrix.

where $\varrho_1 \geq \varrho_2 \geq \dots \varrho_{mn} \geq 0$ are $\boldsymbol{W}^*$'s eigenvalues in decreasing order and $\boldsymbol{u}_l$ is the eigenvector associated with ϱ_l. All the $\boldsymbol{u}_l$ are non-zero and orthogonal. As $\operatorname{rank}(\boldsymbol{W}^*) \geq 2$, we have $\varrho_2 > 0$. Now let $2 \leq L \leq mn$ be the largest number such that $\varrho_L > 0$, then we have

$$\boldsymbol{V}_{\text{ref}}\boldsymbol{V}_{\text{ref}}^{\mathsf{H}} = [\boldsymbol{W}^*]_{00} = \sum_{l=1}^{L} \varrho_l [\boldsymbol{u}_l]_0 [\boldsymbol{u}_l]_0^{\mathsf{H}} =: \boldsymbol{U}\boldsymbol{U}^{\mathsf{H}},$$
$$\boldsymbol{U} := \left[\sqrt{\varrho_1}[\boldsymbol{u}_1]_0, \sqrt{\varrho_2}[\boldsymbol{u}_2]_0, \dots, \sqrt{\varrho_L}[\boldsymbol{u}_L]_0\right].$$

If the rank of $\boldsymbol{U}$ is strictly greater than 1, then we can find $\boldsymbol{z} \in \operatorname{span}(\boldsymbol{U})$ such that $\boldsymbol{z}^{\mathsf{H}}\boldsymbol{V}_{\text{ref}} = 0$. Then $\boldsymbol{U}^{\mathsf{H}}\boldsymbol{z} \neq 0$ implies

$$0 = \boldsymbol{z}^{\mathsf{H}}\boldsymbol{V}_{\text{ref}}\boldsymbol{V}_{\text{ref}}^{\mathsf{H}}\boldsymbol{z} = \boldsymbol{z}^{\mathsf{H}}\boldsymbol{U}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{z} > 0.$$

The contradiction means $\operatorname{rank}(\boldsymbol{U}) \leq 1$, and therefore $[\boldsymbol{u}_1]_0$ and $[\boldsymbol{u}_2]_0$ are linearly dependent. If $[\boldsymbol{u}_1]_0 = r[\boldsymbol{u}_2]_0$ for some $r \in \mathbb{C}$, then we construct $\tilde{\boldsymbol{u}} = \boldsymbol{u}_1 - r\boldsymbol{u}_2$. Otherwise $[\boldsymbol{u}_2]_0$ must be zero and we construct $\tilde{\boldsymbol{u}} = \boldsymbol{u}_2$. Clearly we have

$$\tilde{\boldsymbol{u}} \neq 0, [\tilde{\boldsymbol{u}}]_0 = \boldsymbol{0}. \tag{3.12}$$

On the other hand, KKT conditions give $\operatorname{tr}(\boldsymbol{A}^*\boldsymbol{W}^*) = 0$. As both $\boldsymbol{A}^*$ and $\boldsymbol{W}^*$ are positive semidefinite, we have

$$\begin{aligned} 0 =& \operatorname{tr}(\boldsymbol{A}^*\boldsymbol{W}^*) = \operatorname{tr}\Big(\boldsymbol{A}^* \sum_{l=1}^{L} \varrho_l \boldsymbol{u}_l \boldsymbol{u}_l^{\mathsf{H}}\Big) \\ =& \sum_{l=1}^{L} \operatorname{tr}\big(\varrho_l \boldsymbol{A}^* \boldsymbol{u}_l \boldsymbol{u}_l^{\mathsf{H}}\big) = \sum_{l=1}^{L} \operatorname{tr}\big(\varrho_l \boldsymbol{u}_l^{\mathsf{H}} \boldsymbol{A}^* \boldsymbol{u}_l\big) \geq 0. \end{aligned}$$

The equality holds only when $\boldsymbol{A}^*\boldsymbol{u}_l = \boldsymbol{0}$ for all $l \leq L$. Hence

$$\boldsymbol{A}^*\tilde{\boldsymbol{u}} = \boldsymbol{0}. \tag{3.13}$$

As (3.12) has shown $1 \leq |\Omega(\tilde{\boldsymbol{u}})| \leq n-1$, Theorem 7 and (3.13) imply that there exists $\hat{\boldsymbol{u}}$ such that $\Omega(\hat{\boldsymbol{u}})$ is non-empty, connected in $\mathcal{G}$, $1 \leq |\Omega(\hat{\boldsymbol{u}})| \leq n-1$, and $\boldsymbol{A}^*\hat{\boldsymbol{u}} = \boldsymbol{0}$. Let j be a node not in $\Omega(\hat{\boldsymbol{u}})$ but is connected to some node $k \in \Omega(\hat{\boldsymbol{u}})$. Since A2 requires $\mathcal{G}$ to be a tree and $\Omega(\hat{\boldsymbol{u}})$ is connected in $\mathcal{G}$, k must be the only node in $\Omega(\hat{\boldsymbol{u}})$ which is connected to $j \notin \Omega(\hat{\boldsymbol{u}})$. Otherwise there is a cycle. Then

$$\begin{aligned} [\boldsymbol{A}^*\hat{\boldsymbol{u}}]_j =& \sum_{l \in \mathcal{V}} [\boldsymbol{A}^*]_{jl}[\hat{\boldsymbol{u}}]_l = [\boldsymbol{A}^*]_{jj}[\hat{\boldsymbol{u}}]_j + \sum_{l:l\sim j} [\boldsymbol{A}^*]_{jl}[\hat{\boldsymbol{u}}]_l \\ =& [\boldsymbol{A}^*]_{jj}\boldsymbol{0} + [\boldsymbol{A}^*]_{jk}[\hat{\boldsymbol{u}}]_k + \sum_{l:l\sim j, l\notin\Omega(\hat{\boldsymbol{u}})} [\boldsymbol{A}^*]_{jl}[\hat{\boldsymbol{u}}]_l. \end{aligned}$$

As $[\hat{\boldsymbol{u}}]_l = \mathbf{0}$ for $l \notin \Omega(\hat{\boldsymbol{u}})$, we have $[\boldsymbol{A}^*\hat{\boldsymbol{u}}]_j = [\boldsymbol{A}^*]_{jk}[\hat{\boldsymbol{u}}]_k$. Further, $(j,k) \in \mathcal{E}$ and the $\mathcal{G}$-invertibility of $\boldsymbol{A}^*$ implies $[\boldsymbol{A}^*]_{jk}$ is invertible. Node k is in $\Omega(\hat{\boldsymbol{u}})$ implies $[\hat{\boldsymbol{u}}]_k \neq \mathbf{0}$. As a result, $[\boldsymbol{A}^*\hat{\boldsymbol{u}}]_j = [\boldsymbol{A}^*]_{jk}[\hat{\boldsymbol{u}}]_k$ must be non-zero, contracting $\boldsymbol{A}^*\hat{\boldsymbol{u}} = \mathbf{0}$. Therefore, $\boldsymbol{W}^*$ must have rank 1. □

Similarly, we can prove the following corollary for perturbed problems.

Corollary 9. *Let* $(\boldsymbol{W}^*, \boldsymbol{A}^*(\varepsilon))$ *be a pair of primal/dual solutions to* (3.5) *and* (3.10). *If* $\boldsymbol{A}^*(\varepsilon)$ *is* $\mathcal{G}$*-invertible,* $\boldsymbol{W}^*$ *must be rank* 1.

3.5 First Perspective: Sparse Critical Buses

In this section, we will show that if critical buses (i.e., buses that appear in the cost function or yield binding constraints) are sparse in the network, then the dual matrix $\boldsymbol{A}^*$ tends to be sparse, and the relaxation is therefore exact. We first propose the following assumption.

A1: Problem (3.4) has unique optimal solution and we remove line current constraints (3.3e). [2]

The KKT condition is necessary and sufficient optimality condition for the primal (3.5) and the dual (3.8) problem. In this section, $\boldsymbol{W}^*$ refers to the unique solution of (3.4).

Notations

The following notations and definitions will be used throughout the rest of the chapter.

For each bus-phase pair (j,ϕ), we define

$$f_p(j,\phi) := \begin{cases} 0, & \operatorname{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}^*) \notin \{\overline{p}_j^\phi, \underline{p}_j^\phi\} \\ 1, & \operatorname{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}^*) = \overline{p}_j^\phi \\ -1, & \operatorname{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}^*) = \underline{p}_j^\phi \end{cases}.$$

The strong duality guarantees that $\overline{p}_j^\phi$ and $\underline{p}_j^\phi$ cannot be attained simultaneously, so the definition above is fully specified. Similarly we define

$$f_q(j,\phi) := \begin{cases} 0, & \operatorname{tr}(\boldsymbol{\Psi}_j^\phi \boldsymbol{W}^*) \notin \{\overline{q}_j^\phi, \underline{q}_j^\phi\} \\ 1, & \operatorname{tr}(\boldsymbol{\Psi}_j^\phi \boldsymbol{W}^*) = \overline{q}_j^\phi \\ -1, & \operatorname{tr}(\boldsymbol{\Psi}_j^\phi \boldsymbol{W}^*) = \underline{q}_j^\phi \end{cases}.$$

[2] Alternatively, one can assume $\overline{\ell}_{jk}^\phi = \infty$ for all $j \sim k$.

Definition 16. *The critical objective bus set is*

$$\mathcal{S}_{\mathrm{o}} := \{j \in \mathcal{V} : \exists \phi \text{ s.t. } c^{\phi}_{j,re} \neq 0 \text{ or } c^{\phi}_{j,im} \neq 0\}.$$

Definition 17. *The critical constraint bus set is*

$$\mathcal{S}_{\mathrm{c}} := \{j \in \mathcal{V} : \exists \phi \text{ s.t. } f_p(j,\phi) \neq 0 \text{ or } f_q(j,\phi) \neq 0\}.$$

For any $mn \times mn$ matrix $\boldsymbol{X}$, we use $[\boldsymbol{X}]_{j,k}$ to denote the $m \times m$ block of $\boldsymbol{X}$ from rows $jm+1$ to $jm+m$ and from columns $km+1$ to $km+m$. Further, for $\phi \in \mathcal{M}$, we denote $[\boldsymbol{X}]^{\phi,:}_{j,k}$ and $[\boldsymbol{X}]^{:,\phi}_{j,k}$ as the ϕ^{th} row and column of $[\boldsymbol{X}]_{j,k}$, respectively. Similarly, for an mn dimensional vector $\boldsymbol{x}$, we use $[\boldsymbol{x}]_j$ to denote the subvector of $\boldsymbol{x}$ from the $(jm+1)^{\text{th}}$ to $(jm+m)^{\text{th}}$ entry. Denote

$$\Omega(\boldsymbol{x}) := \{j \in \mathcal{V}, [\boldsymbol{x}]_j \neq \boldsymbol{0}\}$$

and we use $|\Omega|$ to denote its cardinality.

We say $\mathcal{V}_1 \subseteq \mathcal{V}$ is *connected* in $\mathcal{G}$ if $\mathcal{G}$ has a connected subgraph whose vertex set is $\mathcal{V}_1$. For any node $j \in \mathcal{V}$, we denote the set of its neighbors in $\mathcal{G}$ as $\mathcal{N}(j)$. For $\mathcal{K} \subseteq \mathcal{V}$, we reload $\mathcal{N}(\mathcal{K}) := \cup_{j\in\mathcal{K}}\mathcal{N}(j)$.

We say a set of real numbers are *sign-semidefinite* if all the non-zero numbers are of the same sign.

Main Results

Consider the following conditions.

A2: The underlying graph $\mathcal{G}$ is a tree.

A3: $(\mathcal{S}_{\mathrm{o}} \cup \mathcal{S}_{\mathrm{c}}) \cap \mathcal{N}(\mathcal{S}_{\mathrm{o}} \cup \mathcal{S}_{\mathrm{c}}) = \emptyset$.

A4: $\mathcal{S}_{\mathrm{o}} \cap \mathcal{S}_{\mathrm{c}} = \emptyset$.

A5: For any $j \in \mathcal{S}_{\mathrm{o}} \cap \mathcal{S}_{\mathrm{c}}$ and $\phi \in \mathcal{M}$, $c^{\phi}_{j,\mathrm{re}} f_p(j,\phi) \geq 0$ and $c^{\phi}_{j,\mathrm{im}} f_q(j,\phi) \geq 0$.

Informally, A3 means all the critical buses are not adjacent to each other. A5 means if a bus is both critical in objective function and constraints, then for all m phases, $\{c^{\phi}_{j,re}, f_p(j,\phi)\}$ and $\{c^{\phi}_{j,im}, f_q(j,\phi)\}$ are sign-semidefinite, respectively. The following two theorems provide two sets of sufficient conditions for exact SDP relaxation.

Theorem 9. *If conditions A1, A2, A3 and A4 hold, then* (3.4) *is exact.*

Theorem 10. *If conditions A1, A2, A3 and A5 hold, then* (3.4) *is exact.*

Both theorems rely on strict feasibility, tree structure and critical buses not be adjacent. Theorem 9 needs $\mathcal{S}_o$ and $\mathcal{S}_c$ to be also disjoint. On the other hand, Theorem 10 allows them to intersect, but says for each (j, ϕ) in the intersection, the objective and constraints should encourage its injection to move in the same direction. [3] Since A4 implies A5, Theorem 10 is stronger than Theorem 9. In the next section, we will only provide a proof of Theorem 10.

One drawback of Theorems 9 and 10 is that the sufficient conditions are given in terms of the optimal solution $\boldsymbol{W}^*$. The next result provides a sufficient condition that depends only on the primal parameters in (3.2). Let

$$\tilde{\mathcal{S}}_c := \{j \in \mathcal{V} : \exists \phi \text{ s.t. } \{\pm\infty\} \nsubseteq \{\underline{p}_j^\phi, \overline{p}_j^\phi, \underline{q}_j^\phi, \overline{p}_j^\phi\}\}.$$

Corollary 10. *Suppose A1 and A2 hold, If* $(\mathcal{S}_o \cup \tilde{\mathcal{S}}_c) \cap \mathcal{N}(\mathcal{S}_o \cup \tilde{\mathcal{S}}_c) = \emptyset$ *and* $\mathcal{S}_o \cap \tilde{\mathcal{S}}_c = \emptyset$*, then* (3.4) *is exact.*

Proof. As $\mathcal{S}_c \subseteq \tilde{\mathcal{S}}_c$, the conditions in the corollary imply A1–A4 and thus exactness holds. □

Informally, Corollary 10 shows that if all the buses involved in the objective function and constraints are not adjacent to each other, then the SDP relaxation is exact.

3.6 Proof of Sufficient Conditions

Review

The existing works [15, 67] prove that the optimal solution of SDP relaxation is of rank 1 in single phase networks. A crucial step in their proof uses the strong duality to show that the product of the primal optimal solution $\boldsymbol{W}^*$ and the dual matrix $\boldsymbol{A}^*$ is a zero matrix, and hence the rank of $\boldsymbol{W}^*$ cannot exceed the dimension of $\boldsymbol{A}^*$'s null space. Under certain conditions [15, 67] prove that $\boldsymbol{A}^*$'s null space is of dimension at most 1. Hence the optimal primal solution $\boldsymbol{W}^*$ must be of rank at most 1.

This argument however breaks down in a multiphase network for the following two reasons. First, although the underlying graph for m phase network is still a tree, each

[3] For example, if $\mathrm{Re}(s_j^\phi)$ is minimized in the objective function, then the lower bound of $\mathrm{Re}(s_j^\phi)$ should not be active in the constraints.

bus now has m different phases and might have m unbalanced voltages in general. If we extend each phase to a separate vertex in the new graph and connect every phase pair between every two neighboring buses, then the m phase network will be transformed into an (mn)-node meshed network with multiple cycles [7, 22, 44]. Hence the theory for single-phase radial network is not applicable. Second, in an m phase network, it is unknown wether the null space of $\boldsymbol{A}^*$ at the optimal point is still of dimension 1. It is therefore not clear how to prove $\text{rank}(\boldsymbol{W}^*) = 1$ via analyzing the dimension of $\text{null}(\boldsymbol{A}^*)$.

In the following argument, we use a similar proof framework to that in [15], but the proof will be based on the eigenvectors of $\boldsymbol{W}^*$ instead of the dimension of $\text{null}(\boldsymbol{A}^*)$. From now on, we suppose A1, A2, A3, and A5 hold.

Preliminaries

Our strategy is to prove the exactness of the perturbed OPF problem and then use Lemma 12 to show (3.4) is also exact. It is important to make sure that all the non-active constraints will remain non-active in the perturbation neighborhood.

Lemma 13. *For any nonzero $\boldsymbol{C}_1$, there exists a positive sequence $\varepsilon \downarrow 0$ such that for each ε in the sequence, one can collect $(\overline{\mu}_j^{\phi}(\varepsilon), \underline{\mu}_j^{\phi}(\varepsilon), \overline{\eta}_j^{\phi}(\varepsilon), \underline{\eta}_j^{\phi}(\varepsilon))$ from at least one of its KKT multiplier tuples satisfying*

$$f_p(j,\phi) = 0 \implies \overline{\mu}_j^{\phi}(\varepsilon) = \underline{\mu}_j^{\phi}(\varepsilon) = 0 \tag{3.14a}$$

$$f_p(j,\phi) \neq 0 \implies f_p(j,\phi) \cdot (\overline{\mu}_j^{\phi}(\varepsilon) - \underline{\mu}_j^{\phi}(\varepsilon)) \geq 0 \tag{3.14b}$$

$$f_q(j,\phi) = 0 \implies \overline{\eta}_j^{\phi}(\varepsilon) = \underline{\eta}_j^{\phi}(\varepsilon) = 0 \tag{3.14c}$$

$$f_q(j,\phi) \neq 0 \implies f_q(j,\phi) \cdot (\overline{\eta}_j^{\phi}(\varepsilon) - \underline{\eta}_j^{\phi}(\varepsilon)) \geq 0. \tag{3.14d}$$

Proof. First consider any positive sequence $\{\varepsilon_l\}_{l=1}^{\infty}$ such that $\lim_{l\to\infty} \varepsilon_l = 0$. Suppose the optimal solution to (3.5) under ε_l is $\boldsymbol{W}_l$ (if there are multiple solutions then select one of them). As (3.5b) prescribes a compact set, using a similar argument as in the proof of Lemma 12 we know there must be a subsequence of $\{\varepsilon_l\}_{l=1}^{\infty}$, denoted by $\{\varepsilon_{z_t}\}_{t=1}^{\infty}$, such that $\boldsymbol{W}_{z_t}$ converges to $\boldsymbol{W}^*$ in the max norm. The difference $\|\boldsymbol{W}_{z_t} - \boldsymbol{W}^*\|_{\infty}$ can be arbitrarily small for sufficiently large t. When t is large enough, the non-active constraints in (3.5b) under $\boldsymbol{W}^*$ will remain non-active under $\boldsymbol{W}_{z_t}$, and

the corresponding KKT multipliers will remain 0. As a result,

$$\begin{aligned} f_p(j,\phi) = 0 \implies & \underline{p}_j^\phi < \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}^*) < \overline{p}_j^\phi \\ \implies \underline{p}_j^\phi < \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}_{z_t}) < \overline{p}_j^\phi \implies & \overline{\mu}_j^\phi(\varepsilon_{z_t}) = \underline{\mu}_j^\phi(\varepsilon_{z_t}) = 0, \end{aligned}$$

$$\begin{aligned} f_p(j,\phi) = +1 \implies & \underline{p}_j^\phi < \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}^*) \\ \implies \underline{p}_j^\phi < \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}_{z_t}) \implies & \underline{\mu}_j^\phi(\varepsilon_{z_t}) = 0 \\ \implies f_p(j,\phi) \cdot (\overline{\mu}_j^\phi(\varepsilon_{z_t}) - & \underline{\mu}_j^\phi(\varepsilon_{z_t})) \geq 0, \end{aligned}$$

$$\begin{aligned} f_p(j,\phi) = -1 \implies & \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}^*) < \overline{p}_j^\phi \\ \implies \mathrm{tr}(\boldsymbol{\Phi}_j^\phi \boldsymbol{W}_{z_t}) < \overline{p}_j^\phi \implies & \overline{\mu}_j^\phi(\varepsilon_{z_t}) = 0 \\ \implies f_p(j,\phi) \cdot (\overline{\mu}_j^\phi(\varepsilon_{z_t}) - & \underline{\mu}_j^\phi(\varepsilon_{z_t})) \geq 0 \end{aligned}$$

all hold. A similar argument can also be applied to prove (3.14c) and (3.14d). □

Properties of Dual Matrix $\boldsymbol{A}^*(\varepsilon)$

In order to apply Lemma 12, we construct $\boldsymbol{C}_1 \in \mathbb{C}^{mn \times mn}$ in the following manner:

$$\begin{aligned} [\boldsymbol{C}_1]_{jj} &= \boldsymbol{0} \in \mathbb{C}^{m \times m}, \quad \text{for } j \in \mathcal{V} \\ [\boldsymbol{C}_1]_{jk} &= \boldsymbol{0} \in \mathbb{C}^{m \times m}, \quad \text{for } (j,k) \notin \mathcal{E}. \end{aligned}$$

When $(j,k) \in \mathcal{E}$, we assume $j < k$. If neither j nor k is in $\mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, then we construct $[\boldsymbol{C}_1]_{jk} = \boldsymbol{Y}_{jk}$.

If $j \in \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, then A3 guarantees $k \notin \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$. $\forall \phi \in \mathcal{M}$, we set $[\boldsymbol{C}_1]_{jk}^{\phi,:}$ to $\boldsymbol{Y}_{jk}^{\phi,:}$ if $c_{j,\mathrm{re}}^\phi = c_{j,\mathrm{im}}^\phi = f_p(j,\phi) = f_q(j,\phi) = 0$, and to $(f_p(j,\phi) + f_q(j,\phi)\boldsymbol{i})\boldsymbol{Y}_{jk}^{\phi,:}$ otherwise.

If $k \in \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, then A3 guarantees $j \notin \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$. $\forall \phi \in \mathcal{M}$, we similarly set $[\boldsymbol{C}_1]_{jk}^{:,\phi}$ to $(\boldsymbol{Y}_{kj}^{\phi,:})^\mathsf{H}$ if $c_{k,\mathrm{re}}^\phi = c_{k,\mathrm{im}}^\phi = f_p(k,\phi) = f_q(k,\phi) = 0$, and to $(f_p(k,\phi) - f_q(k,\phi)\boldsymbol{i})(\boldsymbol{Y}_{kj}^{\phi,:})^\mathsf{H}$ otherwise.

Finally, we set $[\boldsymbol{C}_1]_{kj} := [\boldsymbol{C}_1]_{jk}^\mathsf{H}$ for all $j < k$ to make $\boldsymbol{C}_1$ Hermitian.

The next theorem provides a key intermediate result to prove Theorem 10. Suppose under such $\boldsymbol{C}_1$, the sequence guaranteed by Lemma 13 is $\{\varepsilon_l\}_{l=1}^\infty$.

Theorem 11. *Under A1, A2, A3, and A5, for each ε_l, the dual matrix $\boldsymbol{A}^*(\varepsilon_l)$ is $\mathcal{G}$-invertible.* [4]

[4] If the KKT multiplier tuple at ε_l is non-unique, then $\boldsymbol{A}^*(\varepsilon_l)$ is evaluated at the multiplier tuple in Lemma 13 satisfying (3.14).

Proof. The value of $\boldsymbol{A}^*(\varepsilon_l)$ is the same as the right hand side of (3.11) when all dual variables take values at their corresponding KKT multipliers (with respect to ε_l). If not otherwise specified, all the $(\overline{\mu}_j^\phi, \underline{\mu}_j^\phi, \overline{\eta}_j^\phi, \underline{\eta}_j^\phi)$ in this proof refer to the tuple in Lemma 13 with respect to ε_l. Since for all $a \neq b$, $[\boldsymbol{E}_j^\phi]_{ab}$ and $[\Pi(\kappa)]_{ab}$ are always zero matrices, it is sufficient to show

$$\boldsymbol{Q} := \sum_{j,\phi} \left((\overline{\mu}_j^\phi - \underline{\mu}_j^\phi)\boldsymbol{\Phi}_j^\phi + (\overline{\eta}_j^\phi - \underline{\eta}_j^\phi)\boldsymbol{\Psi}_j^\phi \right) + \boldsymbol{C}_0 + \varepsilon_l \boldsymbol{C}_1$$

satisfies the two conditions in Definition 15.[5]

For $a \neq b$ and $(a,b) \notin \mathcal{E}$, recall that $\boldsymbol{C}_0$ is the linear combination of $\boldsymbol{\Phi}_j^\phi$ and $\boldsymbol{\Psi}_j^\phi$. When $(a,b) \notin \mathcal{E}$, $\boldsymbol{Y}_{ab}$ is a zero matrix and so are all $[\boldsymbol{\Phi}_j^\phi]_{ab}$ and $[\boldsymbol{\Psi}_j^\phi]_{ab}$. The construction of $\boldsymbol{C}_1$ also guarantees $[\boldsymbol{C}_1]_{ab}$ is all zero. Hence $[\boldsymbol{Q}]_{ab}$ is all zero as well.

Now assume $a < b$. If $(a,b) \in \mathcal{E}$, we have

$$\begin{aligned}
&[\boldsymbol{Q}]_{ab} \\
=&\sum_\phi \left((\overline{\mu}_a^\phi - \underline{\mu}_a^\phi + c_{a,\mathrm{re}}^\phi)[\boldsymbol{\Phi}_a^\phi]_{ab} + (\overline{\eta}_a^\phi - \underline{\eta}_a^\phi + c_{a,\mathrm{im}}^\phi)[\boldsymbol{\Psi}_a^\phi]_{ab} \right) \\
+&\sum_\phi \left((\overline{\mu}_b^\phi - \underline{\mu}_b^\phi + c_{b,\mathrm{re}}^\phi)[\boldsymbol{\Phi}_b^\phi]_{ab} + (\overline{\eta}_b^\phi - \underline{\eta}_b^\phi + c_{b,\mathrm{im}}^\phi)[\boldsymbol{\Psi}_b^\phi]_{ab} \right) \\
+&\varepsilon_l [\boldsymbol{C}_1]_{ab}.
\end{aligned} \tag{3.15}$$

If neither a nor b is in $\mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, then by definition, for all $\phi \in \mathcal{M}$ there must be

$$c_{a,\mathrm{re}}^\phi = c_{a,\mathrm{im}}^\phi = f_p(a,\phi) = f_q(a,\phi) = 0, \tag{3.16a}$$

$$c_{b,\mathrm{re}}^\phi = c_{b,\mathrm{im}}^\phi = f_p(b,\phi) = f_q(b,\phi) = 0. \tag{3.16b}$$

Equation (3.15) and Lemma 13 imply $[\boldsymbol{Q}]_{ab} = \varepsilon_l [\boldsymbol{C}_1]_{ab}$. By construction, $[\boldsymbol{C}_1]_{ab} = \boldsymbol{Y}_{ab}$ is invertible, and so is $[\boldsymbol{Q}]_{ab}$.

If $a \in \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, then A3 guarantees $b \notin \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$. Thus (3.16b) holds for all $\phi \in \mathcal{M}$. For a given $\phi \in \mathcal{M}$, if (3.16a) holds, then by construction, we have $[\boldsymbol{Q}]_{ab}^{\phi,:} = \varepsilon_l [\boldsymbol{C}_1]_{ab}^{\phi,:} = \varepsilon_l \boldsymbol{Y}_{ab}^{\phi,:}$. If (3.16a) does not hold for the given ϕ, then we have

$$\begin{aligned}
[\boldsymbol{Q}]_{ab}^{\phi,:} = &(\overline{\mu}_a^\phi - \underline{\mu}_a^\phi + c_{a,\mathrm{re}}^\phi + 2\varepsilon_l f_p(a,\phi)) \frac{\boldsymbol{Y}_{ab}^{\phi,:}}{2} \\
&+ (\overline{\eta}_a^\phi - \underline{\eta}_a^\phi + c_{a,\mathrm{im}}^\phi + 2\varepsilon_l f_q(a,\phi)) \frac{\boldsymbol{Y}_{ab}^{\phi,:}}{2} \boldsymbol{i}.
\end{aligned}$$

[5] The matrix $\boldsymbol{Q}$ itself might not be $\mathcal{G}$-invertible as $\boldsymbol{Q}$ might not be positive semidefinite, but $\boldsymbol{A}^* \geq 0$ always holds.

Note that Condition A5 and Lemma 13 imply both $\{\overline{\mu}_a^\phi - \underline{\mu}_a^\phi, f_p(a,\phi), c_{a,re}^\phi\}$ and $\{\overline{\eta}_a^\phi - \underline{\eta}_a^\phi, f_q(a,\phi), c_{a,\mathrm{im}}^\phi\}$ are sign-semidefinite sets, respectively. When (3.16a) does not hold, at least one of $\{c_{a,\mathrm{re}}^\phi, c_{a,\mathrm{im}}^\phi, f_p(a,\phi), f_q(a,\phi)\}$ is non-zero. As a result, there exists some non-zero $\sigma_{ab}^{\phi,:} \in \mathbb{C}$ such that $[\boldsymbol{Q}]_{ab}^{\phi,:} = \sigma_{ab}^{\phi,:} \boldsymbol{Y}_{ab}^{\phi,:}$. In short, in the case $a \in \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, $[\boldsymbol{Q}]_{ab}^{\phi,:}$ is always a non-zero multiple of $\boldsymbol{Y}_{ab}^{\phi,:}$. The invertibility of $\boldsymbol{Y}_{ab}$ indicates all the $\boldsymbol{Y}_{ab}^{\phi,:}$ are independent for $\phi \in \mathcal{M}$, so $[\boldsymbol{Q}]_{ab}$ is also invertible.

If $b \in \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$, then A3 guarantees $a \notin \mathcal{S}_\mathrm{o} \cup \mathcal{S}_\mathrm{c}$. Then (3.16a) holds for all $\phi \in \mathcal{M}$. For a given $\phi \in \mathcal{M}$, if (3.16b) holds, then by construction, we have $[\boldsymbol{Q}]_{ab}^{:,\phi} = \varepsilon_l [\boldsymbol{C}_1]_{ab}^{:,\phi} = \varepsilon_l (\boldsymbol{Y}_{ba}^{\phi,:})^\mathsf{H}$. If (3.16b) does not hold, then similar to the previous case, there exists some non-zero $\sigma_{ab}^{:,\phi} \in \mathbb{C}$ such that $[\boldsymbol{Q}]_{ab}^{:,\phi} = \sigma_{ab}^{:,\phi} (\boldsymbol{Y}_{ba}^{\phi,:})^\mathsf{H}$. Hence $[\boldsymbol{Q}]_{ab}^{:,\phi}$ is always a non-zero multiple of $(\boldsymbol{Y}_{ba}^{\phi,:})^\mathsf{H}$. The invertibility of $\boldsymbol{Y}_{ba}$ indicates all the $\boldsymbol{Y}_{ba}^{\phi,:}$ are independent for $\phi \in \mathcal{M}$, so $[\boldsymbol{Q}]_{ab}$ is also invertible. □

Proof of Theorem 10

Theorem 11 and Corollary 9 imply that (3.5) is exact under conditions A1, A2, A3, and A5 for any ε. By Lemma 12, Theorem 10 is proved. □

3.7 Discussion and Example

Discussion

Theorem 9, Theorem 10, and Corollary 10 provide us with the first perspective that partially explain why the SDP relaxation for unbalanced multiphase network tends to be exact. Conceptually, those results require the critical buses to be sparse over the network. Particularly, those results give three sets of sufficient conditions which may have slightly different interpretations and implications.

Sufficient conditions in Corollary 10 do not rely on the optimal solution of SDP relaxation, and can be checked *a priori*. Though these conditions are still restrictive in practice, we hope this result can stimulate more work on unbalanced multiphase networks.

Conditions in Theorems 9 and 10 rely on knowing the active constraints at the optimal point, which cannot be checked a priori. Nevertheless, the actual value of the optimal point is not involved as long as one knows where the bottlenecks are. These conditions also suggest that relaxation is more likely to be exact if critical buses turn out to be spread over the network rather than concentrated in some neighborhood.

In A1 we have assumed that (3.4) has a unique optimal solution so that inactive

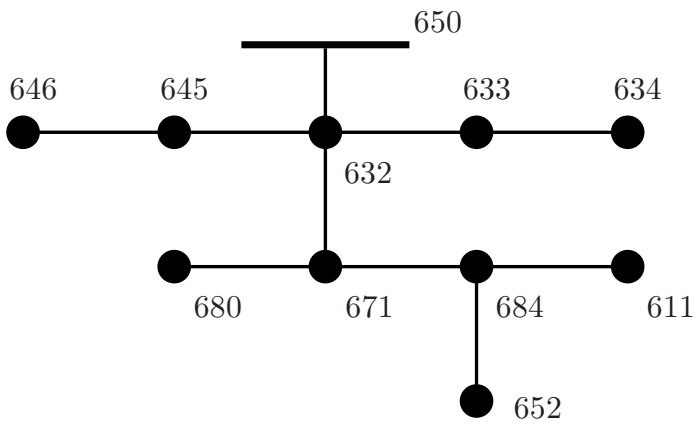

Figure 3.1: An 11 bus network revised from IEEE 13 node test feeder. The switch in the original system is assumed to be open so 2 buses are removed.

constraints at the optimal solution of (3.4) remain inactive under a small perturbation. If (3.4) has multiple solutions, A4 and A5 in Theorems 9 and 10 and condition $\mathcal{S}_{\rm o} \cap \tilde{\mathcal{S}}_{\rm c} = \emptyset$ in Corollary 10 need to be replaced by the linear separability condition proposed in [15]. The proof will be similar.

To generalize the result here to nonlinear cost functions, note that the proposed conditions involving the cost function only rely on the signs of $c^{\phi}_{j,re}$ and $c^{\phi}_{j,\mathrm{i}m}$. The same argument in this chapter can be extended to the nonlinear case when the cost function is convex, monotonic, and additively separable in injections.

Illustrative Example

We use an 11 bus radial network shown in Fig. 3.1, adapted from IEEE 13 node test feeder, to illustrate our theoretical result. The line configuration is reassigned and noise is added to the admittance matrix, so all the buses have three complete phases and each y_{jk} is invertible. For illustrative purpose, all the real/reactive injections are bounded from at most one direction. Table 3.1 summarizes our setup. The '+' and '-' refer to the sign of $c^{\phi}_{j,re}$ or $c^{\phi}_{j,\mathrm{i}m}$ in the cost function. For constraints, 'u' (or 'l') means the upper (or lower) bound for the corresponding injection is finite. It is easy to check that no matter which constraints are active at the optimal point, conditions A1, A2, A3, and A5 must hold, so Theorem 10 implies the optimal solution is of rank 1.

After solving the problem, there are actually nine active constraints, highlighted in light red in Table 3.1. The largest two eigenvalues of the resulting optimal solution

Bus	Phase	Objective (real)	Objective (reactive)	Constraints (real)	Constraints (reactive)
650	a	+	+	u	u
	b	+	+	u	u
	c	+	+	u	u
632	a				
	b				
	c				
633	a	-	-	l	l
	b	-	-	l	l
	c	-	-	l	l
634	a				
	b				
	c				
645	a	+	-	u	l
	b	+	-	u	l
	c	+	-	u	l
646	a				
	b				
	c				
671	a	-	+	l	u
	b	-	+	l	u
	c	-	+	l	u
684	a				
	b				
	c				
611	a	+	+	u	u
	b	+	+	u	u
	c	+	+	u	u
652	a	+			
	b	+			
	c	+			
680	a				
	b				
	c				

Table 3.1: Illustrative example summary.

$\boldsymbol{W}^*$ are 36.90 and 1.44×10^{-10}, respectively. It confirms that $\boldsymbol{W}^*$ is indeed rank 1 up to numerical precision.

Finally, we refer to [30] for more simulation results, which show that semidefinite relaxation is also exact for IEEE 13, 37, 123-bus networks and a real-world 2065-bus network. In the simulation of [30], our sufficient conditions are actually violated

since the cost function is set as

$$\sum_{j\in\mathcal{V}}\sum_{\phi\in\mathcal{M}}\mathrm{Re}(s_j^{\phi}).$$

It means even when all the buses are critical, the semidefinite relaxation can still be exact.

3.8 Second Perspective: Narrow Marginal Price Range

In this section, we provide a new perspective which may also partially explain why $\boldsymbol{A}^*$, the optimal solution to (3.8), tends to be $\mathcal{G}$-invertible. Recall that in [45], one key finding says that having nonnegative nodal prices yields exact solutions. For multi-phase networks, however, we find that such condition is not sufficient. Instead, we need the nodal prices for any two adjacent buses to be close. This condition cannot be checked *a priori*, but for systems maintained at a normal operating point, such assumption is reasonable, and in Chapter 5 we will estimate the approximate values of nodal prices to further support this argument. From now on, we will focus on 3-phase networks, so m is set as 3 throughout this section.

In the first assumption, we let y_{jk} to be a complex matrices that is symmetric across all phases.

B1: For each edge $j \sim k$, y_{jk} has the form

$$y_{jk} = \begin{bmatrix} a_{jk} & b_{jk} & b_{jk} \\ b_{jk} & a_{jk} & b_{jk} \\ b_{jk} & b_{jk} & a_{jk} \end{bmatrix},$$

where both $a_{jk}, b_{jk} \in \mathbb{C}$ and Further, assume

$$-\mathrm{Im}(a_{jk}) > \mathrm{Re}(a_{jk}) > 0, \tag{3.17a}$$

$$-\mathrm{Im}(a_{jk} - b_{jk}) > \mathrm{Re}(a_{jk} - b_{jk}) > 0. \tag{3.17b}$$

Now we can also decompose y_{jk} as $A_{jk} + B_{jk}$ where $A_{jk} = (a_{jk} - b_{jk})\mathbf{I}_3$ and $B_{jk} = b_{jk}\mathbf{1}\mathbf{1}^{\mathsf{T}}$.

We let $\mu_j^{\phi} = \overline{\mu}_j^{\phi} - \underline{\mu}_j^{\phi}$, $\eta_j^{\phi} = \overline{\eta}_j^{\phi} - \underline{\eta}_j^{\phi}$, where $\overline{\mu}_j^{\phi}, \underline{\mu}_j^{\phi}, \overline{\eta}_j^{\phi}, \underline{\eta}_j^{\phi}$ are the optimal dual variables of (3.8). Clearly, for any $j \nsim k$, $[\boldsymbol{A}^*]_{jk} = 0$, so to show $\boldsymbol{A}^*$ is $\mathcal{G}$-invertible, it is sufficient to show that $[\boldsymbol{A}^*]_{jk}$ is invertible for all $j \sim k$. For $j \sim k$, the block

$[\boldsymbol{A}^*]_{jk}$ is

$$\begin{aligned}[\boldsymbol{A}^*]_{jk} = &-\frac{1}{2}\mathrm{diag}(c_{j,\mathrm{re}}+\mu_j)y_{jk}+\frac{1}{2\boldsymbol{i}}\mathrm{diag}(c_{j,\mathrm{im}}+\eta_j)y_{jk}\\ &-\frac{1}{2}y_{jk}^{\mathsf{H}}\mathrm{diag}(c_{k,\mathrm{re}}+\mu_k)-\frac{1}{2\boldsymbol{i}}y_{jk}^{\mathsf{H}}\mathrm{diag}(c_{k,\mathrm{im}}+\eta_k)\\ &-|a_{jk}|^2\mathrm{diag}(\nu_{jk}).\end{aligned}$$

We define

$$\begin{aligned}U &= -\frac{1}{2}\mathrm{diag}(c_{j,\mathrm{re}}+\mu_j)+\frac{1}{2\boldsymbol{i}}\mathrm{diag}(c_{j,\mathrm{im}}+\eta_j)\\ V &= -\frac{1}{2}\mathrm{diag}(c_{k,\mathrm{re}}+\mu_k)-\frac{1}{2\boldsymbol{i}}\mathrm{diag}(c_{k,\mathrm{im}}+\eta_k)\\ D &= -|a_{jk}|^2\mathrm{diag}(\nu_{jk})\end{aligned}$$

to be three diagonal matrices. Then we can evaluate $[\boldsymbol{A}^*]_{jk}$ as

$$\begin{aligned}[\boldsymbol{A}^*]_{jk} &= Uy_{jk}+y_{jk}^{\mathsf{H}}V+D\\ &= U(A_{jk}+B_{jk})+(A_{jk}+B_{jk})^{\mathsf{H}}V+D\\ &= UA_{jk}+A_{jk}^{\mathsf{H}}V+D+[U\mathbf{1}\quad b_{jk}^{\mathsf{H}}\mathbf{1}]\begin{bmatrix}b_{jk}\mathbf{1}^{\mathsf{T}}\\ \mathbf{1}^{\mathsf{T}}V\end{bmatrix},\end{aligned}\tag{3.18}$$

where $UA_{jk}+A_{jk}^{\mathsf{H}}V+D$ is diagonal. By Weinstein-Aronszajn formula, $[\boldsymbol{A}^*]_{jk}$ being invertible can be implied by

$$UA_{jk}+A_{jk}^{\mathsf{H}}V+D \text{ being invertible,}\tag{3.19a}$$

$$\mathbf{I}_2+[b_{jk}\mathbf{1}\quad V^{\mathsf{T}}\mathbf{1}]^{\mathsf{T}}(UA_{jk}+A_{jk}^{\mathsf{H}}V+D)^{-1}[U\mathbf{1}\quad b_{jk}^{\mathsf{H}}\mathbf{1}] \text{ being invertible.}\tag{3.19b}$$

For each $j\sim k$, we define

$$\begin{aligned}u_{jk}^{\phi} &= -\frac{1}{2}(c_{j,\mathrm{re}}^{\phi}+\mu_j^{\phi})+\frac{1}{2\boldsymbol{i}}(c_{j,\mathrm{im}}^{\phi}+\eta_j^{\phi});\\ v_{jk}^{\phi} &= -\frac{1}{2}(c_{k,\mathrm{re}}^{\phi}+\mu_k^{\phi})-\frac{1}{2\boldsymbol{i}}(c_{k,\mathrm{im}}^{\phi}+\eta_k^{\phi});\\ d_{jk}^{\phi} &= -|a_{jk}|^2\nu_{jk}^{\phi}\\ h_{jk}^{\phi} &= (a_{jk}-b_{jk})u_{jk}^{\phi}+(a_{jk}-b_{jk})^{\mathsf{H}}v_{jk}^{\phi}+d_{jk}^{\phi}.\end{aligned}$$

We now complete our new set of sufficient conditions for exactness (in terms of dual variables)

B2: For all j, ϕ, $c^{\phi}_{j,\mathrm{re}} + \mu^{\phi}_{j} > 0$ and $c^{\phi}_{j,\mathrm{im}} + \eta^{\phi}_{j} > 0$.

B3: For each $j \sim k$, let $r_{jk} := \max_{\phi}\{|u^{\phi}_{jk}|, |v^{\phi}_{jk}|\}/\min_{\phi}\{|u^{\phi}_{jk}|, |v^{\phi}_{jk}|\}$, then r_{jk} satisfies

$$1 - \frac{3|b_{jk}|}{\mathrm{Re}(a_{jk} - b_{jk})} - \frac{9}{8}(r_{jk} + \frac{1}{r_{jk}} + 2)\frac{|b_{jk}|^2}{\mathrm{Re}(a_{jk} - b_{jk})^2} > 0.$$

Theorem 12. *Conditions B1-B3 are sufficient for $\boldsymbol{A}^*$ to be $\mathcal{G}$-invertible, and also suffcient for Problem* (3.4) *to be exact respect to* (3.3).

We first present the following lemma.

Lemma 14. *Assume $\alpha_i, \beta_i \geq 0$ for $i = 1, 2, \cdots, n$ and $m \leq \alpha_i/\beta_i \leq M$, then we have*

$$\Big(\sum_{i=1}^{n} \alpha_i\Big)\Big(\sum_{i=1}^{n} \beta_i\Big) \leq \frac{1}{4}\Big(\sqrt{\frac{m}{M}} + \sqrt{\frac{M}{m}} + 2\Big)\Big(\sum_{i=1}^{n} \sqrt{\alpha_i\beta_i}\Big)^2.$$

Proof. Consider the following function

$$\begin{aligned} f(x) &= \Big(\sum_{i=1}^{n} \alpha_i\Big)x^2 - (\sqrt{m} + \sqrt{M})\Big(\sum_{i=1}^{n} \sqrt{\alpha_i\beta_i}\Big)x + \sqrt{mM}\sum_{i=1}^{n} \beta_i \\ &= \sum_{i=1}^{n} (\sqrt{\alpha_i}x - \sqrt{m\beta_i})(\sqrt{\alpha_i}x - \sqrt{M\beta_i}). \end{aligned}$$

As $f(1) \leq 0$ and $f(x) \to \infty$ as $x \to \infty$, we have the discriminant of f is nonnegative. That is

$$(\sqrt{m} + \sqrt{M})^2 \cdot \Big(\sum_{i=1}^{n} \sqrt{\alpha_i\beta_i}\Big)^2 - 4\Big(\sum_{i=1}^{n} \alpha_i\Big)\Big(\sqrt{mM}\sum_{i=1}^{n} \beta_i\Big) \geq 0,$$

which implies the desired result. □

Proof of Theorem 12

It is sufficient to show (3.19a) and (3.19b) are satisfied. For the $j \sim k$, $\mathrm{Re}(h^{\phi}_{jk}) < 0$ for all ϕ (due to B2 and (3.17b)), thus $UA_{jk} + A^{\mathsf{H}}_{jk}V + D = \mathrm{diag}(h)$ is invertible.

Then, notice that

$$\begin{aligned}
&\det\left(\mathbf{I}_2+\begin{bmatrix} b_{jk}\mathbf{1}^\mathsf{T} \\ \mathbf{1}^\mathsf{T}V \end{bmatrix}(UA_{jk}+A_{jk}^\mathsf{H}V+D)^{-1}[U\mathbf{1} \quad b_{jk}^\mathsf{H}\mathbf{1}]\right)\\
=&\det\left(\begin{bmatrix} 1+b_{jk}\sum\limits_\phi \frac{u_{jk}^\phi}{h_{jk}^\phi} & |b_{jk}|^2\sum\limits_\phi \frac{1}{h_{jk}^\phi} \\ \sum\limits_\phi \frac{u_{jk}^\phi v_{jk}^\phi}{h_{jk}^\phi} & 1+b_{jk}\sum\limits_\phi \frac{v_{jk}^\phi}{h_{jk}^\phi} \end{bmatrix}\right)\\
=&\Big(1+b_{jk}\sum_\phi \frac{u_{jk}^\phi}{h_{jk}^\phi}\Big)\Big(1+b_{jk}\sum_\phi \frac{v_{jk}^\phi}{h_{jk}^\phi}\Big)-|b_{jk}|^2\sum_\phi\frac{1}{h_{jk}^\phi}\sum_\phi\frac{u_{jk}^\phi v_{jk}^\phi}{h_{jk}^\phi}.
\end{aligned}$$

To show (3.19b) also holds, it is sufficient to show

$$\begin{aligned}
&1-|b_{jk}|\sum_\phi\Big|\frac{u_{jk}^\phi+v_{jk}^\phi}{h_{jk}^\phi}\Big|-|b_{jk}|^2|\sum_\phi\Big|\frac{u_{jk}^\phi}{h_{jk}^\phi}\Big|\sum_\phi\Big|\frac{v_{jk}^\phi}{h_{jk}^\phi}\Big|-|b_{jk}|^2\sum_\phi\Big|\frac{1}{h_{jk}^\phi}\Big|\sum_\phi\Big|\frac{u_{jk}^\phi v_{jk}^\phi}{h_{jk}^\phi}\Big|\\
>&0. \qquad (3.20)
\end{aligned}$$

As $|h_{jk}^\phi|\geq|\mathrm{Re}(h_{jk}^\phi)|\geq \mathrm{Re}(a_{jk}-b_{jk})(|u_{jk}^\phi|+|v_{jk}^\phi|)$, the left hand side of (3.20) is greater than

$$\begin{aligned}
&1-\frac{3|b_{jk}|}{\mathrm{Re}(a_{jk}-b_{jk})}-\Big(\frac{|b_{jk}|}{\mathrm{Re}(a_{jk}-b_{jk})}\Big)^2\Bigg(\sum_\phi\frac{|u_{jk}^\phi|}{|u_{jk}^\phi|+|v_{jk}^\phi|}\sum_\phi\frac{|v_{jk}^\phi|}{|u_{jk}^\phi|+|v_{jk}^\phi|}\\
&+\sum_\phi\frac{1}{|u_{jk}^\phi|+|v_{jk}^\phi|}\sum_\phi\frac{|u_{jk}^\phi||v_{jk}^\phi|}{|u_{jk}^\phi|+|v_{jk}^\phi|}\Bigg)\\
\geq&1-\frac{3|b_{jk}|}{\mathrm{Re}(a_{jk}-b_{jk})}-\frac{2|b_{jk}|^2}{\mathrm{Re}(a_{jk}-b_{jk})^2}\cdot\frac{r_{jk}+\frac{1}{r_{jk}}+2}{4}\cdot\left(\sum_\phi\frac{\sqrt{|u_{jk}^\phi||v_{jk}^\phi|}}{|u_{jk}^\phi|+|v_{jk}^\phi|}\right)^2\\
\geq&1-\frac{3|b_{jk}|}{\mathrm{Re}(a_{jk}-b_{jk})}-\frac{9}{8}(r_{jk}+\frac{1}{r_{jk}}+2)\frac{|b_{jk}|^2}{\mathrm{Re}(a_{jk}-b_{jk})^2}>0.
\end{aligned}$$

□

Discussion

Conditions B1-B3 can be interpreted as 1) admittance is symmetric across all phases, 2) the coupling between different phases should be very small, and 3) for adjacent buses $j\sim k$, the values for $|u_{jk}^\phi|,|v_{jk}^\phi|$ on all phases should be close and within a narrow range. While B1 can be easily checked from problem data, B2 and B3 actually depend on dual variables which are unavailable until the dual

problem is solved. We notice that for most practical problems, $c^{\phi}_{j,\mathrm{re}}$ usually takes close values for all j, ϕ since we may penalize injections at different locations with similar coefficients. For example, when the cost is the total power loss of the entire network, we would have $c^{\phi}_{j,\mathrm{re}} = 1$ for all j, ϕ. The same argument can also be applied to $c^{\phi}_{j,\mathrm{im}}$. On the other hand, the values of μ^{ϕ}_{j} and η^{ϕ}_{j} reflect the sensitivity of the optimal cost with respect to the changes in $(\overline{p}^{\phi}_{j}, \underline{p}^{\phi}_{j})$ and $(\overline{q}^{\phi}_{j}, \underline{q}^{\phi}_{j})$, respectively. They can be viewed as the nodal prices of real/reactive power. For power systems that are maintained at normal operating points, the nodal prices for neighboring buses tend to be close, and this may qualitatively explain why r_{jk} is typically close to 1 for practical systems.

In Chapter 5, we will use single-phase DC model, a linearized version of power flow equations, to simplify and approximate multi-phase AC model. For DC model we can easily see the approximate values of $c^{\phi}_{j,\mathrm{re}} + \mu^{\phi}_{j}$ are between $\min_{j,\phi} c^{\phi}_{j,\mathrm{re}}$ and $\max_{j,\phi} c^{\phi}_{j,\mathrm{re}}$.

Chapter 4

RELAXATION EXACTNESS FOR MULTI-PHASE NETWORKS: WITH DELTA CONNECTIONS

4.1 Background

In this chapter, we are going to further extend our results in Chapter 3. Recall that the focus of Chapter 3 is to study the conditions under which convex relaxation of OPF problems is exact for multi-phase networks, and the underlying model assumes that all injections are connected in wye configurations. As we can see, the semi-definite relaxation we studied in Chapter 3 has a very similar formulation compared to the single-phase scenario. However, in practical systems, injections could be connected in either wye or delta configurations.

Semi-definite relaxation is recently extended in [77] to networks with both wye and delta connected devices by introducing a new positive semi-definite matrix that represents the outer product of voltages and phase-to-phase currents in the delta connections (matrix $\mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}(j)$ in (4.8b) below). Simulation results in [77] showed that, surprisingly, this matrix was never rank-1 at an optimal solution of the relaxation. This seems to suggest that the SDP relaxation was inexact in these simulations. In this chapter, we show that even though the matrix $\mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}(j)$ fails to attain rank 1, an exact solution can still be recovered under certain conditions; see Theorem 14 and Remark 5. The inexactness in previous works is due to two issues. First, optimal solutions to the SDP relaxation in these simulations are generally not unique, and the exact solution is only one of them which is not returned by the solver. Second, such non-uniqueness could significantly amplify the numerical error and make it computationally challenging to recover the exact solution. We propose two variants of the standard SDP relaxation that address both issues. The first algorithm post-processes the relaxation solution and tends to provide lower cost but larger constraint violation, while the second algorithm adds a penalty term to the cost and tends to provide higher cost but smaller constraint violation. Simulations of both algorithms corroborate the theoretical results and show that they can recover exact solutions for three IEEE distribution feeders.

To summarize the main contribution, this chapter first explains why conventional semi-definite relaxation is often inexact when delta connections are present. Then

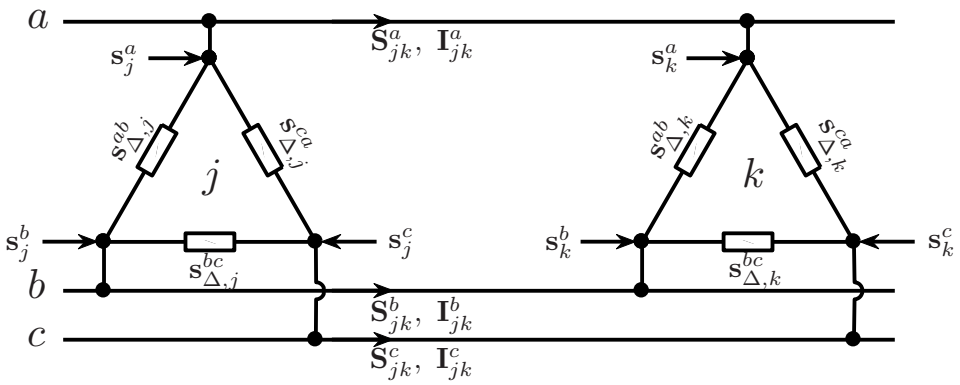

Figure 4.1: Illustration of line (j, k) with both wye and delta connections.

we propose two algorithms and prove that they can recover exact solutions under certain conditions. As a byproduct, we also prove that two models, bus injection model and branch flow model, are equivalent when we relax the problem.

The remainder of the chapter is organized as follows. In Section 4.2, we define the network structure and formulate the three-phase OPF problem in both the bus injection model (BIM) and the branch flow model (BFM). Section 4.3 proves that the global optimal solution to the nonconvex OPF problem can be recovered from its relaxation under certain conditions, and two algorithms are presented. Section 4.4 shows the equivalence between BIM and BFM. Finally, in Section 4.5, we apply our algorithms to IEEE 13-, 37-, and 123-bus systems.

4.2 System Model

Network Structure

We study the model proposed in [30, 77]. Let the directed graph representing the electrical network be $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{0, 1, \ldots, n\}$ denotes the set of buses, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges; let $N := |\mathcal{V}| = n + 1$. In this chapter, we focus on the case where $\mathcal{G}$ represents a radial network (i.e., a tree) because most distribution networks have a tree topology. Throughout the chapter, we will use (graph, vertex, edge) and (power network, bus, line) interchangeably. Without loss of generality, we let bus 0 be the substation bus where the distribution feeder is connected to a transmission network. Suppose the substation also serves as the slack bus, so the voltages at the substation bus are fixed and specified. We use $j \to k$ to denote a directed edge from bus j to k. In many situations, when we do not care about the direction of the edge, we simply use (j, k) and $j \sim k$ interchangeably to

denote an edge connecting bus j and k. That means either $j \to k$ or $k \to j$ is in $\mathcal{E}$. Consider a three-phase line (j,k) characterized by the series impedance matrix $z_{jk} \in \mathbb{C}^{3\times 3}$. When line (j,k) has three phases, the inverse of z_{jk}, denoted as y_{jk}, is the admittance of line (j,k). If branch (j,k) has less than three phases, then we fill the rows and columns of z_{jk} corresponding to the missing phases with zeros, and we let the admittance matrix y_{jk} be the pseudo-inverse of z_{jk}. Last, let $y_j \in \mathbb{C}^{3\times 3}$ denote the admittance of a shunt device connected to bus j.[1]

For each bus j, let the voltages of all three phases at bus j be collected in the vector $\boldsymbol{V}_j \in \mathbb{C}^3$. We use $\boldsymbol{V}_j^\phi$ for $\phi \in \{a,b,c\}$ to indicate the voltage of phase ϕ. The voltage $\boldsymbol{V}_0$ at slack bus 0 is known and denoted by $\mathbf{V}_{\text{ref}}$. Let $\boldsymbol{V} = [\boldsymbol{V}_0^\mathsf{T}, \boldsymbol{V}_1^\mathsf{T}, \ldots, \boldsymbol{V}_n^\mathsf{T}]^\mathsf{T}$ collect the voltages for the entire network. Similarly, we use $\mathbf{s}_j^\phi$ to denote the bus injection for phase ϕ at bus j, and we denote $\mathbf{s}_j$ and $\mathbf{s}$ as the injections at bus j and in the entire network, respectively.

For delta connected components, we use $\boldsymbol{I}_{\Delta,j} \in \mathbb{C}^3$ to collect the delta line currents for phases in $\{ab, bc, ca\}$. Define

$$\Gamma := \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}.$$

Therefore, the complex power injections of the delta connected components at bus j can be expressed as $\mathbf{s}_{\Delta,j} = \text{diag}(\Gamma \boldsymbol{V}_j \boldsymbol{I}_{\Delta,j}^\mathsf{H})$. The net nodal injections contributed by delta connections at bus j are given by $-\text{diag}(\boldsymbol{V}_j \boldsymbol{I}_{\Delta,j}^\mathsf{H} \Gamma)$ (see the illustration in Fig. 4.1). Assume that the operation regions for $\mathbf{s}_j$ and $\mathbf{s}_{\Delta,j}$ at bus j are convex compact sets $\mathcal{S}_j$ and $\mathcal{S}_{\Delta,j}$, respectively.

The AC power flow equations are

$$\begin{aligned} &\mathbf{s}_j - \text{diag}(\boldsymbol{V}_j \boldsymbol{I}_{\Delta,j}^\mathsf{H} \Gamma) - \text{diag}(\boldsymbol{V}_j \boldsymbol{V}_j^\mathsf{H} y_j^\mathsf{H}) \\ &= \sum_{k:j\sim k} \text{diag}((\boldsymbol{V}_j \boldsymbol{V}_j^\mathsf{H} - \boldsymbol{V}_j \boldsymbol{V}_k^\mathsf{H}) y_{jk}^\mathsf{H}) \qquad (4.1a) \\ &\mathbf{s}_{\Delta,j} = \text{diag}(\Gamma \boldsymbol{V}_j \boldsymbol{I}_{\Delta,j}^\mathsf{H}), \qquad (4.1b) \end{aligned}$$

where (4.1a) is the power balance equation at bus j and (4.1b) defines the power through delta connected components.

[1] The shunt here refers to a capacitive device at bus j and not the line charging in the Π circuit model.

Similar to [77], we will adopt $\boldsymbol{X}_j, \rho_j \in \mathbb{C}^{3\times 3}$ as auxiliary matrices to model the outer products of voltages and currents.

$$\boldsymbol{X}_j = \boldsymbol{V}_j \boldsymbol{I}_{\Delta,j}^{\mathsf{H}} \tag{4.2a}$$

$$\rho_j = \boldsymbol{I}_{\Delta,j} \boldsymbol{I}_{\Delta,j}^{\mathsf{H}}. \tag{4.2b}$$

We consider an OPF problem that minimizes a continuous convex cost function $f(\mathbf{s}, \mathbf{s}_\Delta)$ over variables $(\mathbf{s}, \mathbf{s}_\Delta, \boldsymbol{V}, \boldsymbol{I}_\Delta)$ subject to power flow equations (4.1) as well as voltage and injection limits:

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{V},\boldsymbol{I}_\Delta}{\text{minimize}} \quad f(\mathbf{s}, \mathbf{s}_\Delta) \tag{4.3a}$$

$$\text{subject to} \quad (4.1) \tag{4.3b}$$

$$\boldsymbol{V}_0 = \mathbf{V}_{\text{ref}} \tag{4.3c}$$

$$\mathbf{s}_j \in \mathcal{S}_j,\ \mathbf{s}_{\Delta,j} \in \mathcal{S}_{\Delta,j},\ \text{for } j \in \mathcal{V} \tag{4.3d}$$

$$\underline{\boldsymbol{V}} \le |\boldsymbol{V}| \le \overline{\boldsymbol{V}}. \tag{4.3e}$$

In (4.3e), $|\boldsymbol{V}|$ stands for the modulus of $\boldsymbol{V}$ elementwise, and $\underline{\boldsymbol{V}}, \overline{\boldsymbol{V}} \in \mathbb{R}^{3N}$ are the lower and upper limits for voltage magnitudes, respectively. If the limits are homogeneous across all buses and phases, we can denote them as $\underline{V}\mathbf{1}, \overline{V}\mathbf{1}$, where $\underline{V}, \overline{V}$ are scalars and $\mathbf{1}$ is the all-one vector. Next, we will present power flow equations for three-phase radial networks in both bus injection model (4.5) and branch flow model (4.9).

Bus Injection Model

The *bus injection model* (BIM) is defined in terms of $(\mathbf{s}, \mathbf{s}_\Delta, \boldsymbol{W}, \boldsymbol{X}, \rho)$, where we use $\boldsymbol{W} \in \mathbb{C}^{3N\times 3N}$ to replace $\boldsymbol{V}\boldsymbol{V}^{\mathsf{H}}$ in (4.1). The matrix $\boldsymbol{W}_{jk} \in \mathbb{C}^{3\times 3}$ is the (j,k) submatrix of $\boldsymbol{W}$. For notational simplicity, we let

$$\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}(j) := \begin{bmatrix} \boldsymbol{W}_{jj} & \boldsymbol{X}_j \\ \boldsymbol{X}_j^{\mathsf{H}} & \rho_j \end{bmatrix} \qquad \text{for } j \in \mathcal{V}. \tag{4.4}$$

The power flow model is represented as

$$\begin{aligned} &\mathbf{s}_j - \operatorname{diag}(\boldsymbol{X}_j \Gamma) - \operatorname{diag}(\boldsymbol{W}_{jj} y_j^{\mathsf{H}}) \\ &= \sum_{k:j\sim k} \operatorname{diag}((\boldsymbol{W}_{jj} - \boldsymbol{W}_{jk}) y_{jk}^{\mathsf{H}}) \end{aligned} \tag{4.5a}$$

$$\mathbf{s}_{\Delta,j} = \operatorname{diag}(\Gamma \boldsymbol{X}_j) \tag{4.5b}$$

$$\boldsymbol{W}_{00} = \mathbf{V}_{\text{ref}} \mathbf{V}_{\text{ref}}^{\mathsf{H}} \tag{4.5c}$$

$$\boldsymbol{W} \succeq 0 \tag{4.5d}$$

$$\operatorname{rank}(\boldsymbol{W}) = 1 \tag{4.5e}$$

$$\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}(j) \succeq 0 \tag{4.5f}$$

$$\operatorname{rank}\left(\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}(j)\right) = 1. \tag{4.5g}$$

Equations (4.5f), (4.5g) are derived from (4.2) to model the current and power flow of delta connections.

Hence, the AC-OPF problem in BIM formulation is

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{W},\boldsymbol{X},\rho}{\text{minimize}} \quad f(\mathbf{s}, \mathbf{s}_\Delta) \tag{4.6a}$$

$$\text{subject to} \quad (4.5), (4.3d) \tag{4.6b}$$

$$\operatorname{diag}(\underline{\boldsymbol{V}\boldsymbol{V}}^{\mathsf{H}}) \le \operatorname{diag}(\boldsymbol{W}) \le \operatorname{diag}(\overline{\boldsymbol{V}\boldsymbol{V}}^{\mathsf{H}}). \tag{4.6c}$$

Branch Flow Model

In a *branch flow model* (BFM), we introduce $\boldsymbol{S}$, $\boldsymbol{v}$, and $\boldsymbol{\ell}$ to model the branch power flow, squared voltages, and squared currents, respectively. We let $\boldsymbol{I}_{jk} := y_{jk}(\boldsymbol{V}_j - \boldsymbol{V}_k)$ be the sending-end current from bus j to bus k, and $\boldsymbol{S}_{jk}$ be the sending-end branch power from j to k. The matrices $\boldsymbol{S}$, $\boldsymbol{v}$, and $\boldsymbol{\ell}$ can be written as

$$\boldsymbol{S} = (\boldsymbol{S}_{jk} \in \mathbb{C}^{3\times 3})_{(j\to k)\in\mathcal{E}}, \qquad \boldsymbol{S}_{jk} = \boldsymbol{V}_j \boldsymbol{I}_{jk}^{\mathsf{H}} \tag{4.7a}$$

$$\boldsymbol{v} = (\boldsymbol{v}_j \in \mathbb{C}^{3\times 3})_{j\in\mathcal{V}}, \qquad \boldsymbol{v}_j = \boldsymbol{V}_j \boldsymbol{V}_j^{\mathsf{H}} \tag{4.7b}$$

$$\boldsymbol{\ell} = (\boldsymbol{\ell}_{jk} \in \mathbb{C}^{3\times 3})_{(j\to k)\in\mathcal{E}}, \qquad \boldsymbol{\ell}_{jk} = \boldsymbol{I}_{jk} \boldsymbol{I}_{jk}^{\mathsf{H}}. \tag{4.7c}$$

Let

$$\mathbf{M}^{\boldsymbol{v},\boldsymbol{S},\boldsymbol{\ell}}(j,k) := \begin{bmatrix} \boldsymbol{v}_j & \boldsymbol{S}_{jk} \\ \boldsymbol{S}_{jk}^{\mathsf{H}} & \boldsymbol{\ell}_{jk} \end{bmatrix} \qquad \text{for } j \to k \tag{4.8a}$$

$$\mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}(j) := \begin{bmatrix} \boldsymbol{v}_j & \boldsymbol{X}_j \\ \boldsymbol{X}_j^{\mathsf{H}} & \rho_j \end{bmatrix} \qquad \text{for } j \in \mathcal{V}. \tag{4.8b}$$

The branch flow model is defined in terms of variables $(\mathbf{s}, \mathbf{s}_\Delta, \boldsymbol{S}, \boldsymbol{v}, \boldsymbol{\ell}, \boldsymbol{X}, \rho)$, and it is expressed as

$$\boldsymbol{v}_k = \boldsymbol{v}_j - (\boldsymbol{S}_{jk} z_{jk}^{\mathsf{H}} + z_{jk} \boldsymbol{S}_{jk}^{\mathsf{H}}) + z_{jk} \boldsymbol{\ell}_{jk} z_{jk}^{\mathsf{H}} \tag{4.9a}$$

$$\begin{aligned} &\sum_{k:j\to k} \operatorname{diag}(\boldsymbol{S}_{jk}) - \sum_{l:l\to j} \operatorname{diag}(\boldsymbol{S}_{lj} - z_{lj}\boldsymbol{\ell}_{lj}) \\ &= -\operatorname{diag}(\boldsymbol{v}_j y_j^{\mathsf{H}} + \boldsymbol{X}_j \Gamma) + \mathbf{s}_j \end{aligned} \tag{4.9b}$$

$$\mathbf{s}_{\Delta,j} = \operatorname{diag}(\Gamma \boldsymbol{X}_j) \tag{4.9c}$$

$$\boldsymbol{v}_0 = \mathbf{V}_{\text{ref}} \mathbf{V}_{\text{ref}}^{\mathsf{H}} \tag{4.9d}$$

$$\mathbf{M}^{\boldsymbol{v},\boldsymbol{S},\boldsymbol{\ell}}(j,k) \succeq 0 \tag{4.9e}$$

$$\operatorname{rank}\left(\mathbf{M}^{\boldsymbol{v},\boldsymbol{S},\boldsymbol{\ell}}(j,k)\right) = 1 \tag{4.9f}$$

$$\mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}(j) \succeq 0 \tag{4.9g}$$

$$\operatorname{rank}\left(\mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}(j)\right) = 1. \tag{4.9h}$$

Similar to BIM, (4.9g), (4.9h) are also derived from (4.2).

The AC-OPF problem in the BFM form can be formulated as

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{S},\boldsymbol{v},\boldsymbol{\ell},\boldsymbol{X},\rho}{\text{minimize}} \quad f(\mathbf{s}, \mathbf{s}_\Delta) \tag{4.10a}$$

$$\text{subject to} \quad (4.9), (4.3d) \tag{4.10b}$$

$$\operatorname{diag}(\underline{\boldsymbol{V}}_j \underline{\boldsymbol{V}}_j^{\mathsf{H}}) \le \operatorname{diag}(\boldsymbol{v}_j) \le \operatorname{diag}(\overline{\boldsymbol{V}}_j \overline{\boldsymbol{V}}_j^{\mathsf{H}}). \tag{4.10c}$$

4.3 Analytical Results

The main challenge to solving OPF problems (4.6) and (4.10) is the nonconvex rank constraints in (4.5e), (4.5g), (4.9f), and (4.9h). If we drop all the rank-1 constraints, then we obtain

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{W},\boldsymbol{X},\rho}{\text{minimize}} \quad f(\mathbf{s}, \mathbf{s}_\Delta) \tag{4.11a}$$

$$\text{subject to} \quad (4.5a) - (4.5d), (4.5f), (4.3d), (4.6c) \tag{4.11b}$$

as the relaxation for the BIM and

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{S},\boldsymbol{v},\boldsymbol{\ell},\boldsymbol{X},\rho}{\text{minimize}} \quad f(\mathbf{s}, \mathbf{s}_\Delta) \tag{4.12a}$$

$$\text{subject to} \quad (4.3d), (4.10c), (4.9a) - (4.9e), (4.9g) \tag{4.12b}$$

as the relaxation for the BFM. Solving the relaxed problems (4.11) and (4.12) could lead to solutions that are infeasible for the original nonconvex problems (4.6) and (4.10) respectively when the solutions do not satisfy the rank-1 constraints. In what follows, we will explore conditions under which optimal solutions of (4.6) and (4.10) can be recovered from their respective relaxations. First, the following lemma is presented, which is the main ingredient for subsequent results.

Lemma 15. *Consider a block Hermitian matrix*

$$\mathbf{M} := \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\mathsf{H}} & \mathbf{C} \end{bmatrix} \tag{4.13}$$

where $\mathbf{A}$ *and* $\mathbf{C}$ *are both square matrices. If* $\mathbf{M} \succeq 0$ *and* $\mathbf{A} = \mathbf{x}\mathbf{x}^{\mathsf{H}}$ *for some vector* $\mathbf{x}$*, then there must exist some vector* $\mathbf{y}$ *such that* $\mathbf{B} = \mathbf{x}\mathbf{y}^{\mathsf{H}}$.

Proof. As $\mathbf{M} \succeq 0$, it can be decomposed as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \begin{bmatrix} \mathbf{M}_1^{\mathsf{H}} & \mathbf{M}_2^{\mathsf{H}} \end{bmatrix} \tag{4.14}$$

and $\mathbf{A} = \mathbf{M}_1\mathbf{M}_1^{\mathsf{H}}$, $\mathbf{B} = \mathbf{M}_1\mathbf{M}_2^{\mathsf{H}}$, $\mathbf{C} = \mathbf{M}_2\mathbf{M}_2^{\mathsf{H}}$. Because $\mathbf{A} = \mathbf{x}\mathbf{x}^{\mathsf{H}}$ has rank-1, matrix $\mathbf{M}_1$ is in the column space of $\mathbf{x}$ and has rank-1 as well. There must exist vector $\mathbf{z}$ such that $\mathbf{M}_1 = \mathbf{x}\mathbf{z}^{\mathsf{H}}$. As a result, $\mathbf{B} = \mathbf{M}_1\mathbf{M}_2^{\mathsf{H}} = \mathbf{x}\mathbf{z}^{\mathsf{H}}\mathbf{M}_2^{\mathsf{H}} = \mathbf{x}(\mathbf{M}_2\mathbf{z})^{\mathsf{H}}$. □

One observation in Lemma 15 is when submatrices $\mathbf{A}$ and $\mathbf{B}$ are fixed and specified as $\mathbf{x}\mathbf{x}^{\mathsf{H}}$ and $\mathbf{x}\mathbf{y}^{\mathsf{H}}$, there are non-unique $\mathbf{C}$ to make $\mathbf{M}$ positive semi-definite. Similarly, in the relaxations (4.11) and (4.12), the optimal solutions are always non-unique. Taking (4.11) as an example, for any optimal solution $(\mathbf{s}^*, \mathbf{s}_\Delta^*, \boldsymbol{W}^*, \boldsymbol{X}^*, \rho^*)$, one could add to ρ^* an arbitrary positive semi-definite matrix to obtain a different optimal solution $(\mathbf{s}^*, \mathbf{s}_\Delta^*, \boldsymbol{W}^*, \boldsymbol{X}^*, \rho^* + \mathbf{K}\mathbf{K}^{\mathsf{H}})$. This non-uniqueness in the optimal ρ explains why in existing literature such as [77], the relaxation (4.11) could compute rank-1 $\boldsymbol{W}$ (within numerical tolerance) but the resulting $\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}$ is always not rank-1. In fact, the next result shows in theory, if the optimal $\boldsymbol{W}$ is perfectly of rank 1 without any numerical error, then a feasible and optimal solution of (4.6) is recoverable.

Theorem 13. *If* $\boldsymbol{u}^* = (\mathbf{s}^*, \mathbf{s}_\Delta^*, \boldsymbol{W}^*, \boldsymbol{X}^*, \rho^*)$ *is an optimal solution to* (4.11) *that satisfies* $\operatorname{rank}(\boldsymbol{W}^*) = 1$*, then a feasible and optimal solution of* (4.6) *can be recovered from* $\boldsymbol{u}^*$.

Proof. We decompose $\boldsymbol{W}^*_{jj}$ as $\boldsymbol{V}_j\boldsymbol{V}_j^{\mathsf{H}}$ for each j, where $\boldsymbol{V}_j$ is a vector. By Lemma 15, there exists vector $\boldsymbol{I}_{\Delta,j}$ such that $\boldsymbol{X}^*_j = \boldsymbol{V}_j\boldsymbol{I}^{\mathsf{H}}_{\Delta,j}$. One could construct $\tilde{\rho}$ such that $\tilde{\rho}_j = \boldsymbol{I}_{\Delta,j}\boldsymbol{I}^{\mathsf{H}}_{\Delta,j}$.

Since (4.11) is a relaxation of (4.6), for $(\mathbf{s}^*, \mathbf{s}^*_\Delta, \boldsymbol{W}^*, \boldsymbol{X}^*, \tilde{\rho})$ to be optimal for (4.6), it is sufficient that it is feasible for (4.6). Clearly, constraints (4.3d), (4.6c), (4.5a)–(4.5d) are satisfied because they are also the constraints in (4.11) and they do not involve the decision variable ρ. Constraint (4.5e) also holds as $\operatorname{rank}(\boldsymbol{W}^*) = 1$. Further, by Lemma 15, we have

$$\begin{bmatrix} \boldsymbol{W}^*_{jj} & \boldsymbol{X}^*_j \\ (\boldsymbol{X}^*_j)^{\mathsf{H}} & \tilde{\rho}_j \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_j \\ \boldsymbol{I}_{\Delta,j} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_j \\ \boldsymbol{I}_{\Delta,j} \end{bmatrix}^{\mathsf{H}}$$

is both positive semi-definite and of rank-1. Hence, (4.5f) and (4.5g) are also satisfied. Hence, $(\mathbf{s}^*, \mathbf{s}^*_\Delta, \boldsymbol{W}^*, \boldsymbol{X}^*, \tilde{\rho})$ is feasible for (4.6), and this completes the proof. □

Theorem 14. *If* $\boldsymbol{u}^* = (\mathbf{s}^*, \mathbf{s}^*_\Delta, \boldsymbol{S}^*, \boldsymbol{v}^*, \boldsymbol{\ell}^*, \boldsymbol{X}^*, \rho^*)$ *is an optimal solution to* (4.12) *and satisfies* $\operatorname{rank}(\mathbf{M}^{\boldsymbol{v}^*,\boldsymbol{S}^*,\boldsymbol{\ell}^*}(j,k)) = 1$ *for* $j \sim k$ *and* $\operatorname{rank}(\boldsymbol{v}^*_j) = 1$ *for* $j \in \mathcal{V}$*, then an optimal solution of* (4.10) *can be recovered from* $\boldsymbol{u}^*$.

The proof of Theorem 14 is omitted because it is similar to the proof of Theorem 13.

Theorem 13 asserts that in theory, the only critical non-convex constraint of (4.6) is (4.5e), in the sense that a solution satisfying (4.5g) could always be recovered whenever (4.5e) holds. However in practice, $\boldsymbol{W}^*$ is typically not exactly rank-1 due to numerical precision and therefore Theorem 14 could not be directly applied to recover the optimal solution as long as numerical error exists. This is because the recovery method in Theorem 13 relies on the rank-1 decomposition of $\boldsymbol{X}^*$. In practice even if $\boldsymbol{W}^*$ is close to being rank-1, the optimal $\boldsymbol{X}^*$ could still be very different from being rank-1, as we will explain below in Remark 5.

Remark 5 (Spectrum Error)**.** *The matrix* $\mathbf{A}$ *in* (4.14) *being* approximately *rank-1 does not necessarily mean that* $\mathbf{B}$ *is also approximately rank-1.*[2] *For example, consider the case where* $\boldsymbol{x}$*,* $\boldsymbol{e}_1$*, and* $\boldsymbol{e}_2$ *are orthogonal vectors with norms* 1, 10^{-4}*, and* 10^{-5}*, respectively. Similarly, let* $\boldsymbol{y}$*,* $\boldsymbol{z}_1$*, and* $\boldsymbol{z}_2$ *be orthogonal vectors with norms* 1, 10^4*, and* 10^5*, respectively. Then, construct the matrix* $\boldsymbol{M}$ *as in* (4.14) *with*

[2]Here, being approximately rank-1 means that the second largest eigenvalue of the matrix is nonzero but smaller than the largest eigenvalue by several orders of magnitude.

$\mathbf{M}_1 = [\boldsymbol{x}\ \ \boldsymbol{e}_1\ \ \boldsymbol{e}_2]$ *and* $\mathbf{M}_2 = [\boldsymbol{y}\ \ z_1\ \ z_2]$. *Clearly,* $\mathbf{M}$ *has the upper left diagonal block that is approximately rank-1. On the other hand, the upper right block is of rank 3 with three singular values of* 1. *Consequently, even when* $\boldsymbol{W}^*$ *is close to rank-1 within a certain numerical tolerance,* $\boldsymbol{X}^*$ *could be far from being a rank-1 matrix, especially if* ρ^* *already contains a large redundant positive semi-definite matrix* $\mathbf{K}\mathbf{K}^{\mathsf{H}}$. *Decomposing* $\boldsymbol{X}^*$ *as the product of two vectors, as in the proof of Theorem 13, could result in a large numerical error. In other words, the non-uniqueness of* ρ *could significantly amplify the numerical error. In fact, as long as the second largest eigenvalue of* $\boldsymbol{W}^*$ *is not exactly* 0*, then no matter how small it is, such spectrum error could potentially be significant, especially when the trace of* $\mathbf{K}\mathbf{K}^{\mathsf{H}}$ *is large.*

To summarize, there are two factors that prevent the relaxation output from being exact. The first is the non-uniqueness in the relaxation solution, and the second is that such non-uniqueness further greatly amplify the numerical error in computation. This finding motivates two algorithms for practical implementation.

Relaxation with Post-Processing

Remark 5 shows recovering the vector $\boldsymbol{I}_{\Delta,j}$ from $\boldsymbol{X}_j^*$ can lead to poor numerical performance. In the first algorithm, we instead recover $\boldsymbol{I}_{\Delta,j}$ as $\big(\mathrm{diag}(\Gamma \boldsymbol{V}_j)\big)^{-1}\mathbf{s}^*_{\Delta,j}$ from (4.1b), and then we reconstruct $\tilde{\boldsymbol{X}}_j$ as $\boldsymbol{V}_j\boldsymbol{I}^{\mathsf{H}}_{\Delta,j}$. If there is no numerical error, $\boldsymbol{X}^*$ and $\tilde{\boldsymbol{X}}$ should be equal; however, in the presence of spectrum error, they could be different, as discussed in Remark 5. The pseudo code is provided in Algorithm 1.

Algorithm 1 Relaxation Algorithm with Post-Processing.

Input: $y, \mathcal{S}, \mathcal{S}_\Delta$

Output: Optimal solution $(\mathbf{s}, \mathbf{s}_\Delta, \boldsymbol{W}, \boldsymbol{X}, \rho)$ to (4.6).

1: Solve (4.11) to obtain $(\mathbf{s}^*, \mathbf{s}^*_\Delta, \boldsymbol{W}^*, \boldsymbol{X}^*, \rho^*)$.
2: **if** $(\mathrm{rank}(\boldsymbol{W}^*) > 1)$ **then**
3: Output 'Failed!'
4: Exit
5: **else**
6: Decompose $\boldsymbol{W}^*_{jj} = \boldsymbol{V}_j\boldsymbol{V}_j^{\mathsf{H}}$
7: $\boldsymbol{I}_{\Delta,j} \leftarrow \big(\mathrm{diag}(\Gamma \boldsymbol{V}_j)\big)^{-1}\mathbf{s}^*_{\Delta,j}$
8: $\tilde{\boldsymbol{X}} \leftarrow \boldsymbol{V}_j\boldsymbol{I}^{\mathsf{H}}_{\Delta,j}$, $\tilde{\rho}_j \leftarrow \boldsymbol{I}_{\Delta,j}\boldsymbol{I}^{\mathsf{H}}_{\Delta,j}$
9: **return** $(\mathbf{s}^*, \mathbf{s}^*_\Delta, \boldsymbol{W}^*, \tilde{\boldsymbol{X}}, \tilde{\rho})$
10: **end if**

Theorem 15. *If Algorithm 1 does not fail, then its output is an optimal solution of* (4.6).

Theorem 15 is the direct consequence of Theorem 13. Similarly for BFM, one could also apply post-processing to recover the solution of (4.10) from an optimal solution of (4.12). In the BFM, instead of checking the rank of $\boldsymbol{W}^*$, we check the rank of $\mathbf{M}^{\boldsymbol{v}^*,\boldsymbol{S}^*,\boldsymbol{\ell}^*}(j,k)$ for each $j \sim k$ and $\boldsymbol{v}_j^*$ for $j \in \mathcal{V}$.

Relaxation with Penalized Cost Function

Since the inexactness of relaxations (4.11) and (4.12) originates from two issues: the non-uniqueness in ρ^* and the spectrum error, where the latter is essentially amplified by the former. The second algorithm we propose is to penalize and suppress the trace of ρ_j in the cost function. With such penalty term, the value of ρ^* will be unique for fixed $\boldsymbol{W}^*$ and $\boldsymbol{X}^*$ in the solution of (4.11) and the spectrum error can also be restricted. Similar penalization approaches were also previously proposed in [55, 58] to promote low-rank solutions. The penalized relaxed formulation under the BIM becomes

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{W},\boldsymbol{X},\rho}{\text{minimize}} \qquad f(\mathbf{s},\mathbf{s}_\Delta) + \lambda \sum_{j\in\mathcal{V}} \operatorname{tr}(\rho_j) \tag{4.15a}$$

$$\text{subject to} \qquad (4.5a) - (4.5d), (4.5f), (4.3d), (4.6c). \tag{4.15b}$$

Similarly, the penalized relaxed program under the BFM becomes

$$\underset{\mathbf{s},\mathbf{s}_\Delta,\boldsymbol{S},\boldsymbol{v},\boldsymbol{\ell},\boldsymbol{X},\rho}{\text{minimize}} \qquad f(\mathbf{s},\mathbf{s}_\Delta) + \lambda \sum_{j\in\mathcal{V}} \operatorname{tr}(\rho_j) \tag{4.16a}$$

$$\text{subject to} \qquad (4.9a) - (4.9e), (4.9g), (4.3d), (4.10c). \tag{4.16b}$$

Because $\operatorname{tr}(\rho_j)$ is linear and all constraints in (4.15b) and (4.16b) are convex, both (4.15) and (4.16) are convex optimization problems and can be efficiently solved in polynomial time. Here, $\lambda > 0$ controls the weight of $\sum \operatorname{tr}(\rho_j)$ in the cost function. The pseudo code (based on BIM) is summarized in Algorithm 2. The algorithm for BFM is similar.

Because the cost function in the penalized program has been changed, the output of Algorithm 2 might not be the global optimal solution of (4.6). We next show that the output of Algorithm 2 serves as an approximation of the true optimal solution. We make the following assumption.

Assumption 2. *The problem* (4.11) *has at least one finite optimal solution.*

Algorithm 2 Relaxation Algorithm with Penalized Cost Function.

Input: $y, \mathcal{S}, \mathcal{S}_\Delta$
Output: Optimal solution $(\mathbf{s}, \mathbf{s}_\Delta, \boldsymbol{W}, \boldsymbol{X}, \rho)$ to (4.6).

```
Pick a sufficiently small λ > 0
Solve (4.15) and obtain u* := (s*, s*_Δ, W*, X*, ρ*)
if (rank(W*) > 1) then
    Output 'Failed!'
    Exit
else
    return u*
end if
```

Now consider a sequence of positive and decreasing λ_i for $i = 1, 2, \cdots$ such that $\lambda_i \to 0$ as $i \to \infty$. Taking BIM as an example, let the optimal solution of (4.15) with respect to λ_i be $\boldsymbol{u}^{(i)}$.[3] Then the following lemma implies the sequence $\boldsymbol{u}^{(i)}$ has a limit point.

Lemma 16. *The sequence* $(\boldsymbol{u}^{(i)})_{i=1}^{\infty}$ *resides in a compact set, and hence has a limit point.*

Proof. Because all the constraints in (4.15) are closed, we only need to prove boundedness. By assumption, $\mathbf{s}_j^{(i)}$ and $\mathbf{s}_{\Delta,j}^{(i)}$ at bus j are in compact sets $\mathcal{S}_j$ and $\mathcal{S}_{\Delta,j}$ respectively. The positive semi-definite matrix $\boldsymbol{W}^{(i)}$ has upper bounds on its diagonal elements and is therefore bounded. We only need to show that $\sum_j \mathrm{tr}(\rho_j^{(i)})$ is also bounded because the boundedness of $\boldsymbol{X}^{(i)}$ is implied by the constraint (4.5f) as long as $\sum_j \mathrm{tr}(\rho_j^{(i)})$ is bounded.

To show $\sum_j \mathrm{tr}(\rho_j^{(i)})$ is bounded, let $\hat{\boldsymbol{u}} = (\hat{\mathbf{s}}, \hat{\mathbf{s}}_\Delta, \hat{\boldsymbol{W}}, \hat{\boldsymbol{X}}, \hat{\rho})$ be an optimal solution of (4.11). Then $\hat{\boldsymbol{u}}$ is feasible for (4.15) regardless of the value of λ. For any i, we must have $\sum_j \mathrm{tr}(\rho_j^{(i)}) \le \sum_j \mathrm{tr}(\hat{\rho}_j)$; otherwise, $\hat{\boldsymbol{u}}$ will always give a strictly smaller cost value in (4.15) for $\lambda = \lambda_i$ and it would contradict the optimality of $\boldsymbol{u}^{(i)}$. □

Suppose $\tilde{\boldsymbol{u}} := (\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_\Delta, \tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \tilde{\rho})$ is an arbitrary limit point of the sequence $\boldsymbol{u}^{(i)}$. We present sufficient conditions for $\tilde{\boldsymbol{u}}$ to be an optimal solution of (4.6).

Lemma 17. *Consider the positive semi-definite matrix* $\mathbf{M}$ *as in* (4.13) *where* $\mathbf{A} = \mathbf{x}\mathbf{x}^{\mathsf{H}}$ *for some vector* $\mathbf{x}$ *such that* $\mathbf{x} \neq \mathbf{0}$*, and* $\mathbf{B} = \mathbf{x}\mathbf{y}^{\mathsf{H}}$*. Then,*

$$\mathrm{tr}(\mathbf{M}) \ge \mathbf{x}^{\mathsf{H}}\mathbf{x} + \mathbf{y}^{\mathsf{H}}\mathbf{y}$$

[3] If the program has multiple solutions, then pick any one of them.

and equality holds if and only if $\mathbf{C} = \mathbf{y}\mathbf{y}^{\mathsf{H}}$.

Proof. It is sufficient to prove that $\mathbf{C} - \mathbf{y}\mathbf{y}^{\mathsf{H}} \succeq 0$. If not, then suppose there exists $\mathbf{z}$ such that $\mathbf{z}^{\mathsf{H}}(\mathbf{C} - \mathbf{y}\mathbf{y}^{\mathsf{H}})\mathbf{z} < 0$. Because $\mathbf{x} \neq \mathbf{0}$, we can always find $\mathbf{w}$ such that $\mathbf{w}^{\mathsf{H}}\mathbf{x} = -\mathbf{z}^{\mathsf{H}}\mathbf{y}$. Consider

$$\begin{aligned}
&\begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix}^{\mathsf{H}} \mathbf{M} \begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{x}\mathbf{x}^{\mathsf{H}} & \mathbf{x}\mathbf{y}^{\mathsf{H}} \\ \mathbf{y}\mathbf{x}^{\mathsf{H}} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix} \\
&= \mathbf{w}^{\mathsf{H}}\mathbf{x}\mathbf{x}^{\mathsf{H}}\mathbf{w} + \mathbf{z}^{\mathsf{H}}\mathbf{y}\mathbf{x}^{\mathsf{H}}\mathbf{w} + \mathbf{w}^{\mathsf{H}}\mathbf{x}\mathbf{y}^{\mathsf{H}}\mathbf{z} + \mathbf{z}^{\mathsf{H}}\mathbf{C}\mathbf{z} \\
&= \mathbf{z}^{\mathsf{H}}\mathbf{y}\mathbf{y}^{\mathsf{H}}\mathbf{z} - \mathbf{z}^{\mathsf{H}}\mathbf{y}\mathbf{y}^{\mathsf{H}}\mathbf{z} - \mathbf{z}^{\mathsf{H}}\mathbf{y}\mathbf{y}^{\mathsf{H}}\mathbf{z} + \mathbf{z}^{\mathsf{H}}\mathbf{C}\mathbf{z} \\
&= \mathbf{z}^{\mathsf{H}}(\mathbf{C} - \mathbf{y}\mathbf{y}^{\mathsf{H}})\mathbf{z} < 0.
\end{aligned}$$

This contradicts the positive semi-definiteness of $\mathbf{M}$. □

Theorem 16. *If* $\operatorname{rank}(\tilde{\boldsymbol{W}}) = 1$*, then* $\tilde{\boldsymbol{u}}$ *is a globally optimal solution of* (4.6)*.*

Proof. We first show that $\tilde{\boldsymbol{u}}$ is the optimal solution of (4.11). Since (4.11) and (4.15) have the same feasible set, which is closed, $\tilde{\boldsymbol{u}}$ is also feasible for (4.11) and (4.15) for any λ. If $\tilde{\boldsymbol{u}}$ is not optimal for (4.11), then there must exist another point $\bar{\boldsymbol{u}}$ such that $f(\bar{\mathbf{s}}, \bar{\mathbf{s}}_\Delta) + \alpha = f(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_\Delta)$ and $\alpha > 0$. Then for some sufficiently large i_0, we have

$$\begin{aligned}
\lambda_{i_0} \sum_{j \in \mathcal{V}} \operatorname{tr}(\bar{\rho}_j) &< \frac{\alpha}{2} \\
|f(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_\Delta) - f(\mathbf{s}^{(i_0)}, \mathbf{s}_\Delta^{(i_0)})| &< \frac{\alpha}{2}.
\end{aligned}$$

Therefore

$$f(\bar{\mathbf{s}}, \bar{\mathbf{s}}_\Delta) + \lambda_{i_0} \sum_{j \in \mathcal{V}} \operatorname{tr}(\bar{\rho}_j) < f(\mathbf{s}^{(i_0)}, \mathbf{s}_\Delta^{(i_0)}) + \lambda_{i_0} \sum_{j \in \mathcal{V}} \operatorname{tr}(\rho_j^{(i_0)}),$$

which contradicts the optimality of $\boldsymbol{u}^{(i_0)}$.

Then for each j, we decompose $\tilde{\boldsymbol{W}}_{jj} = \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j^{\mathsf{H}}$ and $\tilde{\boldsymbol{X}} = \tilde{\mathbf{x}}_j \tilde{\mathbf{y}}_j^{\mathsf{H}}$, and construct $\rho_j^\dagger$ as $\tilde{\mathbf{y}}_j \tilde{\mathbf{y}}_j^{\mathsf{H}}$. Under the same argument as in the proof of Theorem 13, the solution $\boldsymbol{u}^\dagger := (\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_\Delta, \tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \rho^\dagger)$ is an optimal solution for both (4.6) and (4.11). We want to prove $\tilde{\boldsymbol{u}} = \boldsymbol{u}^\dagger$ and conclude that $\tilde{\boldsymbol{u}}$ is also an optimal solution for (4.6). The proof is by contradiction.

Since $\boldsymbol{u}^\dagger$ is optimal for (4.6), $\mathbf{M}^{\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \rho^\dagger}(j)$ must be of rank-1 for all j. By Lemma 17, we have

$$\operatorname{tr}(\mathbf{M}^{\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \tilde{\rho}}(j)) \geq \operatorname{tr}(\mathbf{M}^{\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}, \rho^\dagger}(j))$$

and therefore $\text{tr}(\tilde{\rho}_j) \geq \text{tr}(\rho_j^\dagger)$ for all j. If $\tilde{\boldsymbol{u}} \neq \boldsymbol{u}^\dagger$, then some equalities cannot be achieved, and as a result, $\sum_j \text{tr}(\tilde{\rho}_j) - \sum_j \text{tr}(\rho_j^\dagger) = \beta$ for some $\beta > 0$.

As $\tilde{\boldsymbol{u}}$ is a limit point of $(\boldsymbol{u}^{(i)})_{i=1}^{\infty}$, there must be some sufficiently large i_1 such that

$$\Big| \sum_{j\in\mathcal{V}} \text{tr}(\tilde{\rho}_j) - \sum_{j\in\mathcal{V}} \text{tr}(\rho_j^{(i_1)}) \Big| < \frac{\beta}{2}.$$

Hence

$$\sum_{j\in\mathcal{V}} \text{tr}(\rho_j^\dagger) < \sum_{j\in\mathcal{V}} \text{tr}(\rho_j^{(i_1)}).$$

On the other hand, (4.11) and (4.15) have the same feasible set, so the optimality of $\boldsymbol{u}^\dagger$ for (4.11) implies $f(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_\Delta) \leq f\left(\mathbf{s}^{(i_1)}, \mathbf{s}_\Delta^{(i_1)}\right)$. Therefore

$$f(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}_\Delta) + \lambda_{i_1} \sum_{j\in\mathcal{V}} \text{tr}(\rho_j^\dagger) < f\left(\mathbf{s}^{(i_1)}, \mathbf{s}_\Delta^{(i_1)}\right) + \lambda_{i_1} \sum_{j\in\mathcal{V}} \text{tr}\left(\rho_j^{(i_1)}\right)$$

which contradicts the fact that $\boldsymbol{u}^{(i_1)}$ is the optimal solution for (4.15) with respect to λ_{i_1}. □

Theorem 16 shows that when we solve the penalized program with a sequence of decreasing λ_i that converge to 0, any limit point would be a global optimal for (4.6) as long as the $\boldsymbol{W}$ matrix associated with the limit point is of rank-1. In our simulations, we apply Algorithm 2 to solve (4.15) with a fixed but sufficiently small λ, which usually results in rank-1 solutions.

Remark 6. *Further, if all optimal solutions of* (4.11) *have the same value for* $\mathbf{s}, \mathbf{s}_\Delta, \boldsymbol{W}, \boldsymbol{X}$*, then Algorithm 1 succeeds if and only if* $\text{rank}(\tilde{\boldsymbol{W}}) = 1$ *holds in Theorem 16. If Algorithm 1 succeeds, its output will also be the same as* $\tilde{\boldsymbol{u}}$*.*

4.4 Model Equivalence

In the previous sections, our results for the BIM and BFM always come in pairs and are analogous. A natural question is whether there exist instances where one model produces an exact solution while the other does not. In single-phase networks and multi-phase systems with only wye connections, [14] and [30] have shown that the two models are equivalent in the sense that one will produce an exact solution if and only if the other will. We show in this subsection that a similar result holds in the presence of delta connections. We first define the equivalence between two optimization problems as follows.

Definition 18. *Consider two optimization problems*

$$\underset{x}{\text{minimize}} \;\; f_A(x) \quad \text{subject to} \;\; x \in \mathcal{X} \tag{4.17}$$

$$\underset{y}{\text{minimize}} \;\; f_B(y) \quad \text{subject to} \;\; y \in \mathcal{Y}. \tag{4.18}$$

We say (4.17) *and* (4.18) *are equivalent if there exist mappings* $g_1 : \mathcal{X} \to \mathcal{Y}$ *and* $g_2 : \mathcal{Y} \to \mathcal{X}$ *such that*

$$x \in \mathcal{X} \Rightarrow g_1(x) \in \mathcal{Y}, \; f_A(x) = f_B(g_1(x)),$$
$$y \in \mathcal{Y} \Rightarrow g_2(y) \in \mathcal{X}, \; f_B(y) = f_A(g_2(y)).$$

We do not require g_1 and g_2 to be bijections, but if one of the mappings is a bijection, then we can always set the other as its inverse. We denote the decision variables for the BIM as

$$\boldsymbol{u}^{\mathrm{BIM}} = (\mathbf{s}^{\mathrm{BIM}}, \mathbf{s}_\Delta^{\mathrm{BIM}}, \boldsymbol{W}^{\mathrm{BIM}}, \boldsymbol{X}^{\mathrm{BIM}}, \rho^{\mathrm{BIM}})$$

and the decision variables for the BFM as

$$\boldsymbol{u}^{\mathrm{BFM}} = (\mathbf{s}^{\mathrm{BFM}}, \mathbf{s}_\Delta^{\mathrm{BFM}}, \boldsymbol{S}^{\mathrm{BFM}}, \boldsymbol{v}^{\mathrm{BFM}}, \boldsymbol{\ell}^{\mathrm{BFM}}, \boldsymbol{X}^{\mathrm{BFM}}, \rho^{\mathrm{BFM}}).$$

The superscripts here are to distinguish the same variable for different models.

Proposition 1. *Problems* (4.11) *and* (4.12) *are equivalent. Moreover, for the pairs* g_1 *and* g_2 *in Definition 18, if* $\boldsymbol{u}^{\mathrm{BIM}}$ *satisfies* (4.5e)*, then* $g_1(\boldsymbol{u}^{\mathrm{BIM}})$ *satisfies* (4.9f)*. If* $\boldsymbol{u}^{\mathrm{BFM}}$ *satisfies* (4.9f)*, then* $g_2(\boldsymbol{u}^{\mathrm{BFM}})$ *satisfies* (4.5e)*.*

Note that (4.11) and (4.12) are the relaxed BIM and BFM models. The proposition above implies that besides the equivalence between nonconvex BIM and BFM models as we derived in Section 4.2, their relaxations are also equivalent. We only sketch a proof here by providing the mappings g_1 and g_2, where g_1 can be written as

$$\mathbf{s}^{\mathrm{BFM}} = \mathbf{s}^{\mathrm{BIM}}, \; \mathbf{s}_\Delta^{\mathrm{BFM}} = \mathbf{s}_\Delta^{\mathrm{BIM}} \tag{4.19a}$$
$$\boldsymbol{S}_{jk}^{\mathrm{BFM}} = (\boldsymbol{W}_{jj}^{\mathrm{BIM}} - \boldsymbol{W}_{jk}^{\mathrm{BIM}}) y_{jk}^{\mathsf{H}} \tag{4.19b}$$
$$\boldsymbol{v}_j^{\mathrm{BFM}} = \boldsymbol{W}_{jj}^{\mathrm{BIM}} \tag{4.19c}$$
$$\boldsymbol{\ell}_{jk}^{\mathrm{BFM}} = y_{jk} (\boldsymbol{W}_{jj}^{\mathrm{BIM}} + \boldsymbol{W}_{kk}^{\mathrm{BIM}} - \boldsymbol{W}_{jk}^{\mathrm{BIM}} - \boldsymbol{W}_{kj}^{\mathrm{BIM}}) y_{jk}^{\mathsf{H}} \tag{4.19d}$$
$$\boldsymbol{X}_j^{\mathrm{BFM}} = \boldsymbol{X}_j^{\mathrm{BIM}}, \; \rho_j^{\mathrm{BFM}} = \rho_j^{\mathrm{BIM}} \tag{4.19e}$$

and g_2 as

$$\mathbf{s}^{\mathrm{BIM}} = \mathbf{s}^{\mathrm{BFM}},\ \mathbf{s}_{\Delta}^{\mathrm{BIM}} = \mathbf{s}_{\Delta}^{\mathrm{BFM}} \tag{4.20a}$$

$$\boldsymbol{W}_{jj}^{\mathrm{BIM}} = \boldsymbol{v}_{j}^{\mathrm{BFM}} \tag{4.20b}$$

$$\boldsymbol{W}_{jk}^{\mathrm{BIM}} = \begin{cases} \boldsymbol{v}_{j}^{\mathrm{BFM}} - \boldsymbol{S}_{jk}^{\mathrm{BFM}} z_{jk}^{\mathsf{H}}, & \text{if } j \to k \\ (\boldsymbol{W}_{kj}^{\mathrm{BIM}})^{\mathsf{H}}, & \text{if } k \to j \end{cases} \tag{4.20c}$$

$$\boldsymbol{X}_{j}^{\mathrm{BIM}} = \boldsymbol{X}_{j}^{\mathrm{BFM}},\ \rho_{j}^{\mathrm{BIM}} = \rho_{j}^{\mathrm{BFM}}. \tag{4.20d}$$

For g_2, the value of $\boldsymbol{W}_{jk}^{\mathrm{BIM}}$ where $j \neq k$ and $(j,k) \notin \mathcal{E}$ can be determined arbitrarily as long as $\boldsymbol{W} \succeq 0$. As $\mathcal{G}$ is a tree, we can always complete the matrix $\boldsymbol{W}^{\mathrm{BIM}} \succeq 0$, but not necessarily in a unique way.

Proposition 1 shows that to apply Algorithm 1, if an optimal solution of (4.11) can produce an exact solution of (4.6), then there must also be an optimal solution of (4.12) that can produce an exact solution of (4.10), even though both (4.6) and (4.10) may have multiple solutions. The converse is also true. Informally, for Algorithm 1, both the BIM and BFM have the same capability of producing exact solutions.

The same holds for the penalized program. The next proposition can be easily proved using the same mappings g_1 and g_2 in (4.19) and (4.20), respectively.

Proposition 2. *Problems* (4.15) *in the BIM and* (4.16) *in the BFM are equivalent when λ takes the same value for both problems.*

Beyond OPF problems, mappings g_1 and g_2 also provide the correspondence between feasible points under the two models. Thereby, a solution of power flow equations under one model can also be translated into a solution with the same physical meaning under the other model by applying g_1 or g_2. Note that power flow equations may have multiple solutions.

Though BIM and BFM are mathematically equivalent, the two models may behave differently in practice and shed lights on different properties. Some analysis may rely on the structure of one model but not the other, which is indeed the case for single-phase networks [52]. The equivalence implies that one could freely choose a model that is more convenient for a specific problem. For instance, our result in Chapter 3 was derived based on BIM.

4.5 Numerical Results

In this section, we show the ability of the proposed relaxation algorithms to recover the optimal solution to (4.6) and (4.10). We use the IEEE 13-, 37-, and 123-node

distribution feeders [65] to assess the exactness of both algorithms for both the BIM and BFM models. Note that the IEEE 123-bus feeder does not include delta connected components. Hence, we artificially added 4 delta connected loads to the feeder to assess the performance of the proposed approaches. Therefore, all feeders in our simulation will include delta connections. In our experiments, we check how close the output matrices $\boldsymbol{W}, \mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}, \mathbf{M}^{\boldsymbol{v},\boldsymbol{S},\boldsymbol{\ell}}, \mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}$ are to being rank-1, and we evaluate the maximum violation of the constraints when the decision variables are produced from the two proposed algorithms. For all the experiments in this section, we show that both algorithms succeed up to numerical precision, and each has its own advantages and disadvantages.

In previous sections, when we refer to Algorithm 1 and 2 as being exact, the claim is in the sense that Algorithm 1 in theory would produce the globally optimal solution of (4.6) *if there were no numerical error* (i.e., Theorem 15), and the output of Algorithm 2 would converge to the globally optimal solution of (4.6) *as λ goes to* 0 (i.e., Theorem 16). In practice, the machine always has finite precision and we always set λ as a fixed small number in the program. Therefore, the output cost of Algorithm 1 in our simulation should be regarded as a lower bound of the globally optimal cost and the cost of Algorithm 2 should be regarded as an approximation. With higher precision and smaller λ (depending on the precision), the output cost of both algorithms will be closer to the actual globally optimal cost.

Experimental Setup

The load transformer in the IEEE test feeders are modeled as lines with equivalent impedance, whereas the substation transformers and regulators are removed. The switches are assumed to be open or short according to their default status. The capacitor banks are modeled as controllable reactive power sources with continuous control space. The same modification is also commonly applied in the literature [24, 77].

The voltage at the substation is assumed to be $\mathbf{V}_{\text{ref}} = \overline{V}[1, e^{\frac{-\boldsymbol{i}2\pi}{3}}, e^{\frac{\boldsymbol{i}2\pi}{3}}]^{\mathsf{T}}$, where $\overline{V}$ is the maximum allowed voltage magnitude. The operational constraints for controllable loads are set as in [77]. The AC-OPF problem is solved with the cost function $f(\mathbf{s}, \mathbf{s}_\Delta)$ comprising three parts. [4] The first part minimizes the total power

[4] We will use $\mathbf{p}, \mathbf{q}$ to denote the real and imaginary parts of $\mathbf{s}$, and $\mathbf{p}_\Delta, \mathbf{q}_\Delta$ to denote the real and imaginary parts of $\mathbf{s}_\Delta$.

losses in the network, and it can be written as

$$p_{\text{loss}} = \sum_{j \in \mathcal{V}} \sum_{\phi \in \breve{}^{j}_{\text{Y}}} \mathbf{p}_j^{\phi} + \sum_{j \in \mathcal{V}} \sum_{\phi \in \breve{}^{j}_{\Delta}} \mathbf{p}_{\Delta,j}^{\phi}.$$

The second part penalizes deviations of the active and reactive injection profile from nominal profiles, and it is given by $d_p(\mathbf{p}, \mathbf{p}_\Delta)$ and $d_q(\mathbf{q}, \mathbf{q}_\Delta)$ as follows:

$$d_p(\mathbf{p}, \mathbf{p}_\Delta) = \sum_{\substack{j \in \mathcal{V} \\ \phi \in \breve{}^{j}_{\text{Y}}}} \frac{(\mathbf{p}_j^{\phi} - \overline{\mathbf{p}}_j^{\phi})^2}{2\overline{\mathbf{p}}_j^{\phi}} + \sum_{\substack{j \in \mathcal{V} \\ \phi \in \breve{}^{j}_{\Delta}}} \frac{(\mathbf{p}_{\Delta,j}^{\phi} - \overline{\mathbf{p}}_{\Delta,j}^{\phi})^2}{2\overline{\mathbf{p}}_{\Delta,j}^{\phi}},$$

$$d_q(\mathbf{q}, \mathbf{q}_\Delta) = \sum_{\substack{j \in \mathcal{V} \\ \phi \in \breve{}^{j}_{\text{Y}}}} \frac{(\mathbf{q}_j^{\phi} - \overline{\mathbf{q}}_j^{\phi})^2}{2\overline{\mathbf{q}}_j^{\phi}} + \sum_{\substack{j \in \mathcal{V} \\ \phi \in \breve{}^{j}_{\Delta}}} \frac{(\mathbf{q}_{\Delta,j}^{\phi} - \overline{\mathbf{q}}_{\Delta,j}^{\phi})^2}{2\overline{\mathbf{q}}_{\Delta,j}^{\phi}}.$$

The values $\overline{\mathbf{p}}_j^{\phi}, \overline{\mathbf{q}}_j^{\phi}, \overline{\mathbf{p}}_{\Delta,j}^{\phi}, \overline{\mathbf{q}}_{\Delta,j}^{\phi}$ represent the nominal active and reactive injection values for phase ϕ at bus j. All the tracking errors are normalized by their nominal values to have the same order of magnitude for all quantities. In addition, $\breve{}^{j}_{\text{Y}} \subseteq \{a, b, c\}$ and $\breve{}^{j}_{\Delta} \subseteq \{ab, bc, ca\}$ denote the available wye and delta connections at bus $j \in \mathcal{V}$, respectively. Penalizing the deviation of power injection can characterize either the operational cost of controllable loads, the curtailment of photovoltaic systems, or the charging cost of batteries. The same cost expression was also used in [25, 77].

The last part minimizes the deviation of the power injections at the substation from the reference injections $\overline{p}_0, \overline{q}_0 \in \mathbb{R}$ provided by the transmission system operator. Therefore, the system operational cost function can be written as

$$\begin{aligned} f(\mathbf{s}, \mathbf{s}_\Delta) =& \mu_\ell \; p_{\text{loss}} + w_p \; d_p(\mathbf{p}, \mathbf{p}_\Delta) + w_q \; d_q(\mathbf{q}, \mathbf{q}_\Delta) \\ &+ \mu_p \; \frac{(\mathbf{1}^\mathsf{T}\mathbf{p}_0 - \overline{p}_0)^2}{\overline{p}_0} + \mu_q \; \frac{(\mathbf{1}^\mathsf{T}\mathbf{q}_0 - \overline{q}_0)^2}{\overline{q}_0}. \end{aligned}$$

The nonnegative weights w_p, w_q, μ_ℓ, μ_p, and μ_q are used to reflect the relative importance of the components of the cost function and are set as follows:

$$w_p = w_q = \mu_\ell = 1, \qquad \mu_p = \mu_q = 4. \tag{4.21}$$

Table 4.1: Rank and infeasibility for the outputs of Algorithm 1 (with post-processing).

Network	# of Δ-loads	Voltage	BIM		BFM	
			$\boldsymbol{W}$-ratio	Infeas. (kW)	$\mathbf{M}^{v,S,\ell}$-ratio	Infeas. (kW)
IEEE-13	2	3%	1.91×10^{-8}	3.20×10^{-1}	1.74×10^{-5}	4.29×10^{-2}
		5%	2.73×10^{-9}	3.20×10^{-1}	1.58×10^{-5}	4.10×10^{-2}
IEEE-37	25	3%	4.81×10^{-10}	9.84×10^{-2}	7.81×10^{-5}	9.67×10^{-2}
		5%	2.77×10^{-8}	9.83×10^{-2}	2.70×10^{-5}	9.72×10^{-2}
IEEE-123	4	3%	7.75×10^{-8}	1.54×10^{-3}	1.07×10^{-4}	1.03×10^{-2}
		5%	7.67×10^{-8}	1.54×10^{-3}	9.64×10^{-5}	1.02×10^{-2}

Table 4.2: Rank and infeasibility for the outputs of Algorithm 2 (with penalized cost function).

Network	Voltage	BIM			BFM		
		$\boldsymbol{W}$-ratio	$\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}$-ratio	Infeas. (kW)	$\mathbf{M}^{\boldsymbol{v},\boldsymbol{S},\boldsymbol{\ell}}$-ratio	$\mathbf{M}^{\boldsymbol{v},\boldsymbol{X},\rho}$-ratio	Infeas. (kW)
IEEE-13	3%	1.34×10^{-10}	2.36×10^{-9}	8.85×10^{-2}	1.44×10^{-10}	1.97×10^{-10}	4.43×10^{-5}
	5%	1.31×10^{-10}	1.96×10^{-9}	8.84×10^{-2}	1.36×10^{-10}	1.57×10^{-10}	1.46×10^{-5}
IEEE-37	3%	3.04×10^{-8}	6.22×10^{-8}	5.75×10^{-6}	8.85×10^{-8}	3.38×10^{-5}	1.45×10^{-6}
	5%	2.94×10^{-8}	1.05×10^{-8}	1.06×10^{-6}	2.12×10^{-8}	3.18×10^{-5}	1.00×10^{-6}
IEEE-123	3%	1.45×10^{-9}	7.03×10^{-9}	9.13×10^{-7}	1.05×10^{-8}	8.99×10^{-9}	1.34×10^{-6}
	5%	1.94×10^{-8}	9.31×10^{-8}	7.84×10^{-6}	7.98×10^{-9}	6.59×10^{-9}	1.40×10^{-6}

Table 4.3: Effect of the penalty parameter on the cost and infeasibility.

λ	BIM		BFM	
	Cost	Infeas. (kW)	Cost	Infeas. (kW)
0	100.0036	9.84×10^{-2}	100.0194	9.67×10^{-2}
0.1	103.9504	1.15×10^{-2}	104.7141	5.99×10^{-4}
1	104.7846	6.00×10^{-5}	104.7840	1.80×10^{-5}
10	104.7886	5.75×10^{-6}	104.7982	1.45×10^{-6}
100	104.9431	3.17×10^{-6}	105.0332	1.23×10^{-7}

Exactness Results for Algorithm 1

In this subsection, we assess the quality of the solutions recovered using Algorithm 1. We solve (4.11) for the BIM as well as (4.12) for the BFM with different values of voltage limits for the three considered feeders. We invoke the Mosek 8.0 conic solver using CVX, a MATLAB-based convex optimization toolbox.

The left-hand side of Table 4.1 provides the result of Algorithm 1 based on the BIM. The voltage column represents the maximum and minimum voltage deviation allowed, i.e., 3% means that the value of $\overline{V}$ and $\underline{V}$ are set to 1.03 pu and 0.97 pu, respectively. We assess the rank of matrices $\boldsymbol{W}_{jj}$, for all $j \in \mathcal{V}$, in terms of the ratio between the top two largest eigenvalues of these matrices. The maximum ratio among all $j \in \mathcal{V}$ is listed in the table. In the solution of (4.11) (before post-processing), the ratio between the two maximum eigenvalues of the matrices $\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}$ is on the order of 10^{-1}, and after the post-processing in Algorithm 1, the final $\mathbf{M}^{\boldsymbol{W},\boldsymbol{X},\rho}$-ratio will be dominated by $\boldsymbol{W}$-ratio and is thus not informative to be displayed in the table. Because of the spectrum error, the output $\tilde{\boldsymbol{X}}$ could be different from $\boldsymbol{X}^*$, and thus having a very small $\boldsymbol{W}$-ratio is not enough to guarantee the feasibility of the final output of Algorithm 1. Therefore, we also assess the infeasibilty of the power flow equations by measuring the maximum violation in (4.5a) for the solutions returned by Algorithm 1. Here, the violation is defined as the difference between the left- and right-hand sides of (4.5a) (in kW) when $\mathbf{s}$, $\boldsymbol{W}$, $\boldsymbol{X}$ are evaluated as the output of Algorithm 1. In our simulations, the infeasibility is on the order of 10^{-3} to 10^{-1} kW, which reflects the effect of the spectrum error after the post-processing. As a benchmark, the load injections for those feeders are on the order of 10^1 to 10^2 kW, and are typically two orders of magnitude higher than the infeasibility.

On the right-hand side of Table 4.1, the rank of $\mathbf{M}^{\boldsymbol{v},\boldsymbol{S},\boldsymbol{\ell}}(j,k)$ for all lines $(j,k) \in \mathcal{E}$

is examined for the same algorithm under the BFM. Again, we present the maximum ratio between the two largest eigenvalues. Similar to the BIM, the infeasibility, i.e., the violation of (4.1a), is shown in the table.

Exactness Results for Algorithm 2

In our setting, the penalized formulations (4.15) and (4.16) are solved with the parameter $\lambda = 10$ in all experiments. We will later show how the value of λ affects the solution quality.

Table 4.2 presents the maximum ratio between the top two largest eigenvalues of $\mathbf{M}^{W,X,\rho}(j)$ for the BIM and $\mathbf{M}^{v,X,\rho}(j)$ for the BFM returned by the solvers. Comparing the infeasibility of the solutions obtained using Algorithm 1, shown in Table 4.1, and Algorithm 2, shown in Table 4.2, it is clear that adding a penalty helps reduce the effect of the spectrum error and leads to globally optimal solutions with much lower infeasibility.

To assess the effect of the penalization approach on the quality of the solutions in terms of cost and feasibility, Table 4.3 shows the effect of increasing the penalty parameter in the cost function as well as the maximum infeasibility of the power equations (in kVA) for the IEEE 37-bus network with 3% voltage limits. The cost in Table 4.3 is evaluated without the penalty term. Note that the case $\lambda = 0$ corresponds to the output of Algorithm 1. Although the solution feasibility is enhanced by increasing the penalty parameter, the cost associated with the solution obtained also increases. Note that the cost obtained with $\lambda = 0$, i.e., from Algorithm 1, represents a lower bound for the optimal cost of the original AC-OPF problem. [5] In addition, increasing the penalty parameter beyond the values considered in Table 4.3 leads to uninteresting solutions because the cost function becomes dominated by the penalty term. For real applicaitons, we suggest to use binary search to find the smallest λ such that the infeasibility of the solution is within the user-specified tolerance range.

Results with Distributed Energy Resources

We now assess the performance of the proposed approach in a more general setting where distributed energy resources (DERs), such as photovoltaics (PV), are installed. In this simulation, we utilize the IEEE 37-bus distribution feeder where we assume

[5] Here is the reason why the cost of Algorithm 1 is regarded as a lower bound. One consequence of having the spectrum error is that numerical error in BIM formulation could lead to larger constraint violation (as indicated by the infeasibility). Therefore, the output cost (with slight constraint violation) may be lower than the actual optimal cost within the feasible set. We would expect the actual optimal cost to be exact if there were no numerical error.

Table 4.4: Results on IEEE 37-bus network with DERs installed when minimizing the electrical losses.

Model	Algorithm 1		Algorithm 2 ($\lambda = 1$)	
	Cost	Infeas. (kW)	Cost	Infeas. (kW)
BIM	2.5571	6.68×10^{-2}	4.0415	4.42×10^{-6}
BFM	2.5631	6.57×10^{-2}	4.0442	7.00×10^{-6}

that five PV systems are installed in delta connections at the buses 725, 729, 731, 732, and 740. The available power at these units is set at 120, 75, 90, 105, and 180 KW, respectively. We also assume that all the PV inverters can provide reactive power support such that the resultant power factor is at least 0.8. Using this modified feeder, we evaluate the performance of the proposed algorithms when the cost function is p_{loss}, i.e., $\mu_p = \mu_q = w_p = w_q = 0$ and $\mu_\ell = 1$. In addition, we set the upper and lower bounds on voltage magnitudes to be 1.03 pu and 0.97 pu, respectively, in this simulation. In Table 4.4, the results of both Algorithm 1 and Algorithm 2 are presented. It is consistent with previous sections that Algorithm 2 often has lower infeasibility compared to Algorithm 1. To assess the voltage magnitudes resulting from the proposed algorithms, Fig. 4.2 depicts the voltage magnitude at all phases for the solution produced by Algorithm 2 ($\lambda = 1$). It is worth noting that the same voltage profile is obtained by both the BFM and BIM formulations. The results confirms that the voltages in the solution are within the operational limits.

In additon, we evaluate the performance of both the BIM and BFM under a different cost function which also includes the substation power deviation, i.e., $w_p = w_q = 0$, $\mu_p = \mu_q = 4$, and $\mu_\ell = 1$. Furthermore, we set the reference substation injection such that it is achievable only if the available PV power is curtailed. Therefore, the reference power tracking term in the cost function becomes not increasing in power injections. This is known to lead to inexactness of the relaxation. We test this cost function using Algorithm 1 and Algorithm 2 ($\lambda = 1$). Table 4.5 shows the infeasibility and the cost function of the solutions obtained using both BIM and BFM. We can see that the infeasibility of the solution obtained using Algorithm 1 is aggravated due to the use of the cost function that is not increasing with power injections. However, adding the penalty term in Algorithm 2 is enough to significantly reduce the infeasibility.

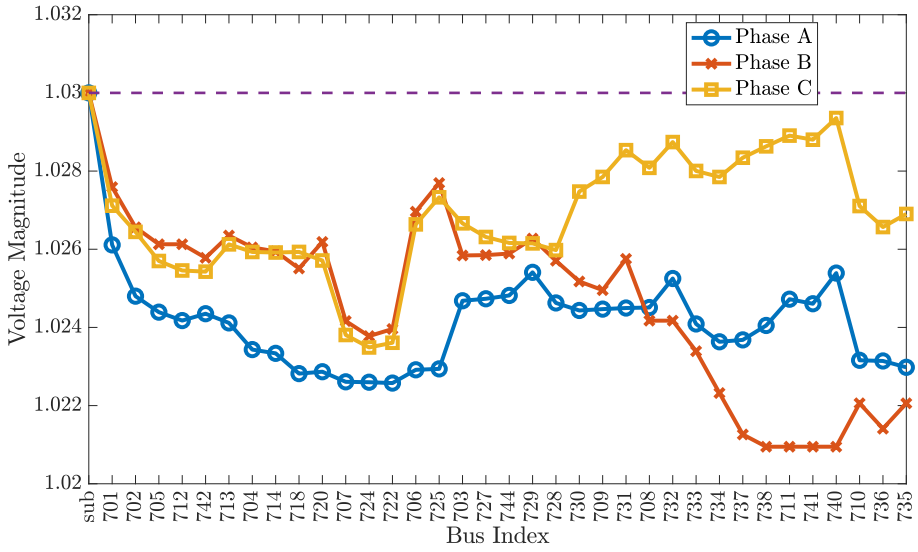

Figure 4.2: The voltage magnitude at all buses for the solution obtained through Algorithm 2 ($\lambda = 1$) for the modified IEEE 37-bus feeder.

Table 4.5: Results on IEEE 37-bus network with DERs installed when minimizing reference tracking cost and electrical loss.

Model	Algorithm 1		Algorithm 2 ($\lambda = 1$)	
	Cost	Infeas. (kW)	Cost	Infeas. (kW)
BIM	18.2124	1.42×10^{-1}	19.6029	6.00×10^{-7}
BFM	18.2281	1.40×10^{-1}	19.6026	6.35×10^{-6}

Algorithm Summary and Comparison

Algorithms 1 and 2 can be useful for different applications. Algorithm 1 solves the un-penalized problem and therefore prioritizes cost minimization at the cost of larger constraint violation. The simulation shows that the infeasibility is typically two orders of magnitude smaller than the load injections and should be acceptable. Algorithm 2, on the other hand, can recover a solution with much smaller constraint violation, but the optimal cost is higher because of the penalty term.

The simulation results also show that the methods under BFM are more numerically stable than BIM, in terms of the infeasibility in Table 4.1 and 4.2. This observation is consistent with the performance of two models for single-phase feeders and feeders without delta connections, as shown in [30, 57].

Table 4.6: Computational Time for both BIM and BFM (in seconds).

Model	IEEE-13	IEEE-37	IEEE-123
BIM	2.46	5.56	9.03
BFM	2.77	4.93	9.87

We also benchmark the computational time of the proposed algorithms in our case studies. Since both algorithms require solving similar optimization problems with slightly different cost functions, the computational time of the two algorithms is similar. Hence, we only present the computational time of Algorithm 1 for all networks using both the BIM and BFM formulations. The algorithm was implemented using Mosek 8.0 as a conic solver on a laptop with Intel Core i9 CPU (2.40 GHz), 16 GB RAM, macOS Catalina OS, and MATLAB R2019b. The results show that the proposed algorithms take less than 10 seconds to solve the AC-OPF problem for the IEEE 123-bus network on a standard laptop, which demonstrates the computational efficiency of the proposed algorithms. More advanced methods such as sparse semi-definite programming solvers, e.g., [75], can further scale the implementation to thousands of buses.

4.6 Conclusion

This chapter studied the SDP relaxation of the AC-OPF problem for an unbalanced three-phase radial network with delta connections, formulated under both the BIM and BFM. We showed the equivalence between the BIM and BFM formulations and presented sufficient conditions for recovering exact solutions of the nonconvex AC-OPF formulations from their respective relaxations. The chapter also showed why conventional relaxation (by directly dropping rank-1 constraints) always fails when the sufficient conditions are approximately satisfied. It is due to the non-uniqueness in the relaxation solution and the spectrum error in computation. Inspired by this finding, we then proposed two algorithms which are guaranteed to produce exact solutions whenever our sufficient conditions are satisfied. One applies post-processing and produces lower cost but larger constraint violation. The other adds a penalty term and produces higher cost but smaller constraint violation. In simulations, we demonstrated that for three IEEE standard test cases, both algorithms are able to recover near globally optimal solutions with tolerable constraint violation and cost suboptimality.

Chapter 5

PRICE AND SENSITIVITY

In this chapter, we are concerned with determining how the optimal cost and the optimal solution of an given OPF problem vary as the demand changes. As the penetration of distributed energy resources increases (for example, causing generation limits to fluctuate depending on the weather), it will become more important to understand such questions. Similar questions have been addressed recently in [10, 35, 63]. In contrast to other chapters, this chapter is based on a more tractable DC model, since it offers the advantage of admitting a linear programming formulation as opposed to a non-convex quadratic program, or in the relaxed case a semidefinite or second-order cone program. Work in [26, 48, 62] explores how good an approximation the DC power flow provides.

We divide the injections into two classes: power load and power generation, while power load (denoted by vector $\boldsymbol{s}^l$) refers to noncontrollable power demand and power generation (denoted by vector $\boldsymbol{s}^g$) refers to controllable power supply. Conceptually the problem can be formulated as a linear program as follows:

$$\begin{aligned} \underset{\boldsymbol{s}^g}{\text{minimize}} \quad & \boldsymbol{f}^\mathsf{T}\boldsymbol{s}^g \\ \text{subject to} \quad & \boldsymbol{A}_{\text{eq}}\boldsymbol{s}^g = \boldsymbol{b}_{\text{eq}}(\boldsymbol{s}^l, \boldsymbol{b}) \\ & \boldsymbol{A}_{\text{in}}\boldsymbol{s}^g \leq \boldsymbol{b}_{\text{in}}. \end{aligned} \tag{5.1}$$

where $\boldsymbol{f}$ is a vector of generation costs (per unit time). The function $\boldsymbol{b}_{\text{eq}}$ is linear in both $\boldsymbol{s}^l$ and $\mathbf{b}$.[1] We are concerned with how an optimal $(\boldsymbol{s}^g)^*$ changes as a function of $\boldsymbol{s}^l$, and how the optima cost $\boldsymbol{f}^\mathsf{T}(\boldsymbol{s}^g)^*$ changes.

In the first half of this chapter, we studied how the optimal cost varies when the load changes, and this is also referred to as the nodal price of the system. Such price is an approximation of $2|u^{\phi}_{jk}|, 2|v^{\phi}_{jk}|$ in Chapter 3. For radial networks, we prove that the nodal price is always between $\min_j \boldsymbol{f}_j$ and $\max_j \boldsymbol{f}_j$, and it helps explain why the values of $|u^{\phi}_{jk}|, |v^{\phi}_{jk}|$ are close for $j \sim k$. In the second half, we go a step further,

[1] The vector $\boldsymbol{b}$ is reserved as a placeholder for any constants which may affect the feasible domain where $\boldsymbol{s}^g$ resides.

and study the sensitivity of the optimal power generation with respect to the load. It is found that for radial networks, such sensitivity is always bounded by 1, but for meshed networks, the value of sensitivity could rapidly grow up in the worst case if the network contains multiple loops.

Notation

Vectors and matrices are typically written in bold while scalars are not. Given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{a} \geq \mathbf{b}$ denotes the element-wise partial order $\mathbf{a}_i \geq \mathbf{b}_i$ for $i = 1, \ldots, n$. For a scalar k, we define $[k]^- := \min\{0, k\}$. We define $\|\mathbf{x}\|_0$ as the number of non-zero elements of the vector $\mathbf{x}$. Identity and zero matrices are denoted by $\mathbf{I}^n$ and $\mathbf{0}^{n\times m}$ while vectors of all ones are denoted by $\mathbf{1}_n$ where superscripts and subscripts indicate their dimensions. To streamline notation, we omit the dimensions when the context makes it clear. The notation $\mathbb{R}_+$ denotes the nonnegative real set $[0, +\infty)$. For $\mathbf{X} \in \mathbb{R}^{n\times m}$, the restriction $\mathbf{X}_{\{1,3,5\}}$ denotes the $3 \times m$ matrix composed of stacking rows $1, 3$, and 5 on top of each other. We will frequently use a set to describe the rows we wish to form the restriction from; in this case we assume the elements of the set are arranged in increasing order. We will use $\boldsymbol{e}_m$ to denote the standard base for the m^{th} coordinate, and its dimension will be clear from the context. Let $(\cdot)^\dagger$ be the Moore-Penrose inverse. Finally, let $[m] := \{1, 2, \ldots, m\}$ and $[n, m] := \{n, n+1, \ldots, m\}$. The indicator function is denoted as $\mathbb{I}$.

5.1 System Model

System model

Consider a power network modeled by an undirected connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} := \mathcal{V}_\mathrm{G} \cup \mathcal{V}_\mathrm{L}$ denotes the set of buses which can be further classified into subsets of generators $\mathcal{V}_\mathrm{G}$ and loads $\mathcal{V}_\mathrm{L}$, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of all branches linking those buses. Suppose $\mathcal{V}_\mathrm{G} \cap \mathcal{V}_\mathrm{L} = \emptyset$ and there are $|\mathcal{V}_\mathrm{G}| =: N_\mathrm{G}$ generator and $|\mathcal{V}_\mathrm{L}| =: N_\mathrm{L}$ loads, respectively. For simplicity, let $\mathcal{V}_\mathrm{G} = [N_\mathrm{G}]$, $\mathcal{V}_\mathrm{L} = [N_\mathrm{G} + 1, N_\mathrm{G} + N_\mathrm{L}]$. Let $N = N_\mathrm{G} + N_\mathrm{L}$. Without loss of generality, $\mathcal{G}$ is a connected graph with $|\mathcal{E}| =: E$ edges labelled as $1, 2, \ldots, E$. Let $\boldsymbol{C} \in \mathbb{R}^{N\times E}$ be the incidence matrix. We will use e, (j, k) or (k, j) interchangeably to denote the same edge. Let $\boldsymbol{B} = \mathrm{diag}(b_1, b_2, \ldots, b_E)$, where $b_e > 0$ is the susceptance of branch e and the value of b_e is $-\mathrm{Im}(y_e)$ where y_e is the line admittance defined in previous chapters. The Laplacian matrix is defined as $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^\mathsf{T}$. As we adopt a DC power flow model, all branches are assumed lossless. Further, we denote the generation and load as $\boldsymbol{s}^g \in \mathbb{R}^{N_\mathrm{G}}$, $\boldsymbol{s}^l \in \mathbb{R}^{N_\mathrm{L}}$, respectively. Thus $\boldsymbol{s}^g_i$ refers to the generation on

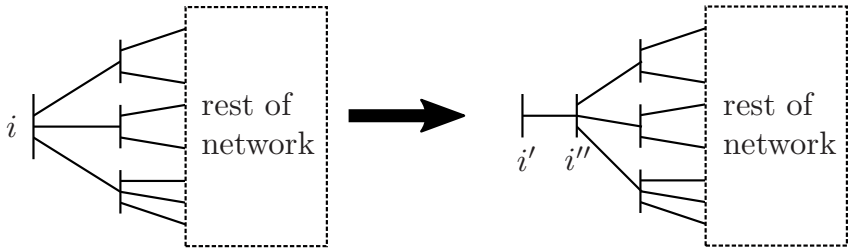

Figure 5.1: Illustration on splitting a bus with both a generator and a load into a pure generation bus and a pure load bus.

bus i while $\boldsymbol{s}_i^l$ refers to the load on bus $N_{\mathrm{G}} + i$. We will refer to bus $N_{\mathrm{G}} + i$ simply as load i for simplicity. The power flow on branch $e \in \mathcal{E}$ is denoted as $\boldsymbol{p}_e$, and $\boldsymbol{p} := [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_E]^\mathsf{T} \in \mathbb{R}^E$ is the vector of all branch power flows. To simplify analysis, we assume that there are no buses in the network that are both loads and generators. This setting is different from previous chapters, where injections are regarded as the net injection of both generation and load. Also, we assume that all generators have degree 1. In fact, those are not restricted assumptions, as we can always split a bus with both a generator and a load into a bus with only the generator connected to another bus with only the load, and connect all the neighbors of the original bus to that load bus (see Fig. 5.1 for a more detailed illustration). Therefore, any network in previous chapters can be equivalently transformed into the model discussed in this chapter.

The DC power flow model assumes that the voltage magnitudes are fixed and known, and the lines are lossless. The DC-OPF problem is a linear program:

$$\begin{aligned}
\underset{\boldsymbol{s}^g, \boldsymbol{\theta}}{\text{minimize}} \quad & \boldsymbol{f}^\mathsf{T} \boldsymbol{s}^g && \text{(5.2a)}\\
\text{subject to} \quad & \boldsymbol{\theta}_1 = 0 && \text{(5.2b)}\\
& \boldsymbol{L}\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{s}^g \\ -\boldsymbol{s}^l \end{bmatrix} && \text{(5.2c)}\\
& \underline{\boldsymbol{s}}^g \le \boldsymbol{s}^g \le \overline{\boldsymbol{s}}^g && \text{(5.2d)}\\
& \underline{\boldsymbol{p}} \le \boldsymbol{B}\boldsymbol{C}^\mathsf{T}\boldsymbol{\theta} \le \overline{\boldsymbol{p}}. && \text{(5.2e)}
\end{aligned}$$

The decision variables are the power generations $\boldsymbol{s}^g$ and voltage angles $\boldsymbol{\theta} \in \mathbb{R}^N$. The cost vector $\boldsymbol{f} \in \mathbb{R}_+^{N_{\mathrm{G}}}$ is the unit cost for each generator and constraint (5.2b) indicates that bus 1 has been set as the slack-bus. All voltage magnitudes are fixed at 1. In constraint (5.2c), we let the injections for generators be positive while the injections for loads be $-\boldsymbol{s}^l$. The upper and lower limits on the generations are set

as $\overline{\boldsymbol{s}}^g$ and $\underline{\boldsymbol{s}}^g$, respectively, and $\overline{\boldsymbol{p}}$ and $\underline{\boldsymbol{p}}$ are the limits on branch power flows. We assume that (5.2) has a non-empty feasible set.

One of our main purposes here is to use the single-phase, DC model to provide an approximation for μ_j^ϕ in the three-phase, AC model. Consider that our definitions for injections are slightly differently in two models, we first want to bridge the gap between them. Assume that the model in Chapter 3 contains N_{L} buses, and each bus has its own load, which is noncontrollable. A subset of buses have controllable generations, and can be split into a pure generator and a pure load. When we apply the splitting process as shown in Fig. 5.1, we also construct a matrix $\boldsymbol{P}$ that if bus i is split into i' (a generator) and i'' (a load), then the i^{th} row of $\boldsymbol{P}$ is constructed as $\boldsymbol{e}_{i'}^\mathsf{T} + \boldsymbol{e}_{i''}^\mathsf{T}$, where $\boldsymbol{e}_{i'}, \boldsymbol{e}_{i''}$ are the basis vectors. Consider the following problem:

$$\begin{aligned}
\underset{\boldsymbol{s}^g, \boldsymbol{\theta}}{\text{minimize}} \quad & \boldsymbol{f}^\mathsf{T} \boldsymbol{P} \boldsymbol{L} \boldsymbol{\theta} && \text{(5.3a)}\\
\text{subject to} \quad & \boldsymbol{\theta}_1 = 0 && \text{(5.3b)}\\
& \boldsymbol{L}\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{s}^g \\ -\boldsymbol{s}^l \end{bmatrix} && \text{(5.3c)}\\
& \boldsymbol{P} \begin{bmatrix} \underline{\boldsymbol{s}}^g \\ -\boldsymbol{s}^l \end{bmatrix} \le \boldsymbol{P}\boldsymbol{L}\boldsymbol{\theta} \le \boldsymbol{P} \begin{bmatrix} \overline{\boldsymbol{s}}^g \\ -\boldsymbol{s}^l \end{bmatrix} && \text{(5.3d)}\\
& \underline{\boldsymbol{p}} \le \boldsymbol{B}\boldsymbol{C}^\mathsf{T}\boldsymbol{\theta} \le \overline{\boldsymbol{p}}. && \text{(5.3e)}
\end{aligned}$$

Problem 5.3 has linear cost in real injection, real power limits, and line constraints. Compared to (3.3), it drops all the reactive power and voltage constraints. By setting $(\boldsymbol{f}, \overline{\boldsymbol{s}}^g, \underline{\boldsymbol{s}}^g, \underline{\boldsymbol{p}}, \overline{\boldsymbol{p}})$ at suitable values, (5.3) can serve as an approximation of problem (3.3). [2] Conceptually, the value of μ_j^ϕ can also be approximated by the dual variables associated with constraint (5.3d). [3] Let $g(\boldsymbol{s}^l)$ be the optimal solution of (5.3) parameterized by $\boldsymbol{s}^l$. By perturbation analysis, the dual variables for (5.3d) are equal to $\partial g(\boldsymbol{s}^l)/\partial \boldsymbol{s}^l$. On the other hand, problems (5.2) and (5.3) are equivalent in the sense that they share the same optimal solution $\boldsymbol{\theta}^*$. However, the cost function of (5.3) penalizes the power load while (5.2) does not. Therefore, we have

$$\frac{\partial \boldsymbol{f}^\mathsf{T}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l} = \frac{\partial g(\boldsymbol{s}^l)}{\partial \boldsymbol{s}^l} + \boldsymbol{f} \approx \mu + \boldsymbol{f}.$$

[2] Here, the approximation includes two steps: 1) approximate a three-phase network by three decoupled single-phase networks; 2) approximate each single-phase network by the linearized model.

[3] μ_j^ϕ is the dual variable for both (3.2) and (3.3) since they share the same dual problem.

Here, $(\boldsymbol{s}^g)^*$ is the optimal solution of (5.2). Compared with Chapter 3, $\boldsymbol{f}$ and $c^{\phi}_{j,\text{re}}$ play the same role as the linear coefficients in the cost function. Therefore, the values of $u^{\phi}_{jk}, v^{\phi}_{jk}$ in Chapter 3 can be approximated by $\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$,the nodal prices of DC OPF problem (5.2), as long as DC model is accurate and the cost on reactive power is negligible. In the later sections, we will prove that the values in $\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$ can be bounded by $\min \boldsymbol{f}_j$ and $\max \boldsymbol{f}_j$. This result partially explains that if the cost function penalizes similarly on all the buses, then the values of $u^{\phi}_{jk}, v^{\phi}_{jk}$ should take values close to each other.

Set Definitions

In general, deriving the closed-form expression for derivatives $\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$ and $\frac{\partial (\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$ is both analytically and computationally challenging. Alternatively, we would like to express those derivatives in terms of the indices of binding constraints. To do so, we first introduce the $\mathcal{OPF}$ operator and provide the following definitions.

Let $\boldsymbol{\xi}$ be a vector of $2N_{\text{G}} + 2E$ network limits arranged as

$$\boldsymbol{\xi} := [(\overline{\boldsymbol{s}}^g)^{\mathsf{T}}, (\underline{\boldsymbol{s}}^g)^{\mathsf{T}}, \overline{\boldsymbol{p}}^{\mathsf{T}}, \underline{\boldsymbol{p}}^{\mathsf{T}}]^{\mathsf{T}}.$$

Define the sets

$$\begin{aligned}
\Omega_{\boldsymbol{\xi}} &:= \{\boldsymbol{\xi} | \underline{\boldsymbol{s}}^g \geq 0, \text{(5.2b)} - \text{(5.2e)} \text{ are feasible for some } \boldsymbol{s}^l > 0\}, \\
\Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi}) &:= \{\boldsymbol{s}^l | \boldsymbol{s}^l > 0, \text{(5.2b)} - \text{(5.2e)} \text{ are feasible}\} \text{ for } \boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}}.
\end{aligned}$$

The set $\Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi})$ is convex and non-empty. When $\boldsymbol{\xi}$ is clear we will simply refer to this set as $\Omega_{\boldsymbol{s}^l}$. These two sets collect the parameters we care about. The next set ensures (5.2) has a unique solution:

$$\begin{aligned}
\Omega_{\boldsymbol{f}} := \{\boldsymbol{f} \geq 0 \mid \forall \boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}}, \boldsymbol{s}^l \in \Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi}),\ &\text{(5.2) has a unique} \\
&\text{solution, and} \geq N_{\text{G}} - 1 \text{ nonzero dual} \\
&\text{variables at the optimal point.}\}
\end{aligned}$$

The definition above is in fact more restrictive than what is needed for uniqueness. We impose the additional constraint on the number of non-zero dual variables as it paves the way for further desirable properties, where we show, that up to perturbation all the binding constraints are independent and there are exactly $N_{\text{G}} - 1$ of them.

With these definitions in hand, we are ready to define the $\mathcal{OPF}$ operator abstraction of (5.2).

5.2 The OPF Operator

Existence and Smoothness

Instead of dealing with the convex program (5.2) directly, we instead treat it as an operator that maps loads to optimal generations.

Definition 19. *Assume $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$. Let $\mathcal{OPF}$ be the operator $\mathcal{OPF} : \Omega_{\boldsymbol{s}^l} \to \mathbb{R}^{N_G}$ such that $\mathcal{OPF}(\boldsymbol{s}^l)$ returns an optimal solution to* (5.2)*, i.e. $(\boldsymbol{s}^g)^* = \mathcal{OPF}(\boldsymbol{s}^l)$.*

We collect a few observations pertaining to $\mathcal{OPF}$:

- The assumption that $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$ ensures that $\mathcal{OPF}$ is a singleton, i.e. it returns a unique element.
- $\mathcal{OPF}$ defines a parametric linear program. Solution sets to parametric LPs are both upper and lower hemi-continuous, thus the $\mathcal{OPF}$ solution set inherits hemi-continuity. Furthermore, when $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$, $\mathcal{OPF}$ is continuous.
- $\Omega_{\boldsymbol{f}}$ is dense in $\mathbb{R}^{N_G}_+$ (See Proposition 1 in [79]). Thus, if $\boldsymbol{f} \notin \Omega_{\boldsymbol{f}}$ applying a small perturbation to $\boldsymbol{f}$ will with probability 1 ensure that $\boldsymbol{f}' \in \Omega_{\boldsymbol{f}}$. So the assumption that $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$ is mild.

We require one final set definition before we can state the differentiability properties of the $\mathcal{OPF}$ operator.

$$\widetilde{\Omega}_{\boldsymbol{s}^l}(\boldsymbol{\xi}, \boldsymbol{f}) := \{\boldsymbol{s}^l \in \Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi}) \mid (5.2) \text{ has exactly } N_G - 1 \text{ binding inequalities.}\}$$

The $N_G - 1$ binding inequalities condition above ensures (when combined with the restriction of $\boldsymbol{f}$ to $\Omega_{\boldsymbol{f}}$) that the set of binding inequalities are independent. This technical assumption is required in the proof of Theorem 17.

Theorem 17. *Assume that $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$. Then there exists a dense set $\widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f}) \subseteq \Omega_{\boldsymbol{\xi}}$ such that for all $\boldsymbol{\xi} \in \widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f})$ the following hold:*

1. *$closure(interior(\Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi}))) = closure(\Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi}))$,*
2. *$\widetilde{\Omega}_{\boldsymbol{s}^l}(\boldsymbol{\xi}, \boldsymbol{f})$ is dense in $\Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi})$.*

Then, when $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$ and $\boldsymbol{\xi} \in \widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f})$, the derivative $\partial_{\boldsymbol{s}^l}\mathcal{OPF}(\boldsymbol{s}^l)$ exists for $\boldsymbol{s}^l \in \widetilde{\Omega}_{\boldsymbol{s}^l}$, and the set of binding constraints remain unchanged in some neighborhood of $\boldsymbol{s}^l$.

Proof. The proof of the two topological properties of $\widetilde{\Omega}_{s^l}(\boldsymbol{\xi}, \boldsymbol{f})$ is somewhat involved but can be found in Appendix C of [79]. With these definitions in hand, by construction, the appropriate sets possess the necessary topological properties such that when combined with the OPF problem (3.2), they satisfy all the necessary conditions in Lemma 4.1 of [35] which guarantee that the derivatives always exist and binding constraints do not change locally.

□

Corollary 11. *If $\boldsymbol{s}^l \in \widetilde{\Omega}_{s^l}$, the $N_{\mathrm{G}} - 1$ binding inequalities, along with $N+1$ equality constraints, are independent.*

Now, suppose at point $\boldsymbol{s}^l$, the set of generators corresponding to binding inequalities is $\mathcal{S}_{\mathrm{G}} \subseteq \mathcal{V}_{\mathrm{G}}$, while the set of branches corresponding to binding inequalities is $\mathcal{S}_{\mathrm{B}} \subseteq \mathcal{E}$. As a consequence of 1) and 2) in Theorem 17 we obtain the following:

Corollary 12. *When $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}, \boldsymbol{\xi} \in \widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f}), \boldsymbol{s}^l \in \widetilde{\Omega}_{s^l}(\boldsymbol{\xi}, \boldsymbol{f})$, we have*

$$|\mathcal{S}_{\mathrm{G}}| + |\mathcal{S}_{\mathrm{B}}| = N_{\mathrm{G}} - 1.$$

We use Fig. 5.2 to summarize the relationship among the sets $\Omega_{\boldsymbol{f}}$, Ω_{s^l}, $\widetilde{\Omega}_{s^l}$, $\Omega_{\boldsymbol{\xi}}$, $\widetilde{\Omega}_{\boldsymbol{\xi}}$ defined above. Informally, set $\Omega_{\boldsymbol{\xi}}$ contains all the $\boldsymbol{\xi}$ that make the OPF problem feasible, and $\Omega_{\boldsymbol{f}}$ contains $\boldsymbol{f}$ that guarantee the unique solution for feasible OPF problems and sufficiently many non-zero Lagrangian multipliers. Each $\boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}}$ maps to a set $\Omega_{s^l}(\boldsymbol{\xi})$, while each $(\boldsymbol{\xi}, \boldsymbol{f})$ maps to set $\widetilde{\Omega}_{s^l}(\boldsymbol{\xi}, \boldsymbol{f})$, which is a subset of $\Omega_{s^l}(\boldsymbol{\xi})$. For fixed $\boldsymbol{f}$, by collecting all the $\boldsymbol{\xi}$ such that $\Omega_{s^l}(\boldsymbol{\xi})$ has "good" topological property and $\widetilde{\Omega}_{s^l}(\boldsymbol{\xi}, \boldsymbol{f})$ is dense in $\Omega_{s^l}(\boldsymbol{\xi})$, we obtain a set $\widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f})$ depending on $\boldsymbol{f}$, and Theorem 17 implies $\widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f})$ is dense in $\Omega_{\boldsymbol{\xi}}$.

We have placed a lot of emphasis on sets being dense. The reason for this is that if the parameter of the OPF problem under consideration does not satisfy the necessary assumptions, then there exists another parameter arbitrarily nearby that does. Thus applying a perturbation to the parameter will provide an OPF problem that does satisfy the necessary conditions for the derivative to be well defined. In summary when $\boldsymbol{f}, \boldsymbol{\xi}, \boldsymbol{s}^l$ belong to $(\Omega_{\boldsymbol{f}}, \widetilde{\Omega}_{s^l}, \widetilde{\Omega}_{\boldsymbol{\xi}})$, we have shown that $\mathcal{OPF}$ has a unique solution, and is well defined and differentiable, as well as that at the optimal solution the binding constraints are independent.

Definition 20. *We use $\mathcal{S}_{\mathrm{G}} \perp \mathcal{S}_{\mathrm{B}}$ to denote that in* (3.2)*, all the inequality constraints corresponding to $\mathcal{S}_{\mathrm{G}}$ and $\mathcal{S}_{\mathrm{B}}$, as well as equality constraints, are independent to each other.*

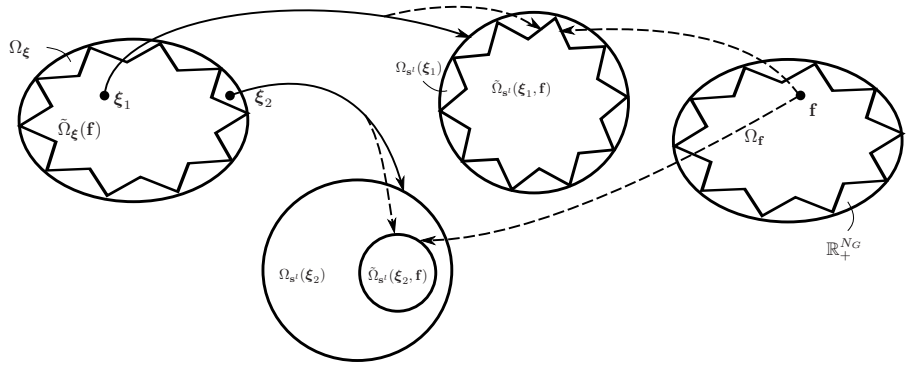

Figure 5.2: Relationship among definitions in Sections 5.1 and 5.2. Solid arrows show the mapping from $\boldsymbol{\xi}$ to $\Omega_{\boldsymbol{s}^l}(\boldsymbol{\xi})$, and dashed arrows show the mapping from $(\boldsymbol{\xi}, \boldsymbol{f})$ to $\widetilde{\Omega}_{\boldsymbol{s}^l}(\boldsymbol{\xi}, \boldsymbol{f})$. A star set being inscribed in a circular set means the former is dense in the latter.

Definition 21. *We say that $\mathcal{S}_{\mathrm{G}} \in \mathcal{V}_{\mathrm{G}}$ and $\mathcal{S}_{\mathrm{B}} \in \mathcal{E}$ form a perfect pair if $\mathcal{S}_{\mathrm{G}} \perp \mathcal{S}_{\mathrm{B}}$ and $|\mathcal{S}_{\mathrm{G}}| + |\mathcal{S}_{\mathrm{B}}| = N_{\mathrm{G}} - 1$. A perfect pair is denoted as $\mathcal{S}_{\mathrm{G}} \sim \mathcal{S}_{\mathrm{B}}$.*

Jacobian Matrix

Notice that the nodal prices $\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$ can also be computed from $\frac{\partial (\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l} = \partial_{\boldsymbol{s}^l} \mathcal{OPF}(\boldsymbol{s}^l)$, we now focus on the Jacobian matrix $\partial_{\boldsymbol{s}^l} \mathcal{OPF}(\boldsymbol{s}^l)$. Let

$$\boldsymbol{J}(\boldsymbol{s}^l; \boldsymbol{f}, \boldsymbol{\xi}) := \partial_{\boldsymbol{s}^l} \mathcal{OPF}(\boldsymbol{s}^l) \in \mathbb{R}^{N_{\mathrm{G}} \times N_{\mathrm{L}}} \tag{5.4}$$

for $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}, \boldsymbol{\xi} \in \widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f}), \boldsymbol{s}^l \in \widetilde{\Omega}_{\boldsymbol{s}^l}(\boldsymbol{f}, \boldsymbol{\xi})$. Suppose at point $\boldsymbol{s}^l$, the set of generators corresponding to binding inequalities is $\mathcal{S}_{\mathrm{G}} \subseteq \mathcal{V}_{\mathrm{G}}$, while the set of branches corresponding to binding inequalities is $\mathcal{S}_{\mathrm{B}} \subseteq \mathcal{E}$. By Corollary 12, we have $|\mathcal{S}_{\mathrm{G}}| + |\mathcal{S}_{\mathrm{B}}| = N_{\mathrm{G}} - 1$. Further, let

$$\boldsymbol{R}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})^{\mathsf{T}} := \begin{bmatrix} \mathbf{I}^N_{\mathcal{V}_{\mathrm{L}}} \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} \\ \mathbf{I}^N_{\mathcal{S}_{\mathrm{G}}} \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} \\ \mathbf{I}^E_{\mathcal{S}_{\mathrm{B}}} \boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} \\ \boldsymbol{e}_1^{\mathsf{T}} \end{bmatrix}, \qquad \boldsymbol{U} = \mathbf{I}^N_{\mathcal{V}_{\mathrm{G}}} \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} \big(\boldsymbol{R}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})^{\mathsf{T}}\big)^{-1}. \tag{5.5}$$

We can derive

$$\boldsymbol{J} = -\boldsymbol{U}(\mathbf{I}^N_{[N_{\mathrm{L}}]})^{\mathsf{T}}. \tag{5.6}$$

It is worth noting that the value of $\boldsymbol{J}$ computed via (5.5)-(5.6) depends on knowing the binding constraints $\mathcal{S}_{\mathrm{G}}$ and $\mathcal{S}_{\mathrm{B}}$ for given $(\boldsymbol{f}, \boldsymbol{\xi}, \boldsymbol{s}^l)$. We abuse notation slightly and let $\boldsymbol{J}(\boldsymbol{s}^l; \boldsymbol{f}, \boldsymbol{\xi})$ be the Jacobian matrix when $(\boldsymbol{f}, \boldsymbol{\xi}, \boldsymbol{s}^l)$ is known and let $\boldsymbol{J}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})$

be the Jacobian matrix when $(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})$ is known. When it is clear from context or not relevant we simply use $\boldsymbol{J}$.

5.3 Topology Analysis

Starting from this section, we study how the network topology, specifically the properties of $\mathcal{S}_{\mathrm{G}}$ and $\mathcal{S}_{\mathrm{B}}$, may affect the value of $\boldsymbol{J}$. The following lemma indicates that under our assumptions, if we view $\mathcal{S}_{\mathrm{B}}$ as a cut of the graph, then each subgraph contains at least one non-binding generator.

Lemma 18. *Suppose $\mathcal{S}_{\mathrm{B}}$ partitions $\mathcal{G}$ into m disjoint subgraphs $\{\mathcal{G}_i(\mathcal{V}_i, \mathcal{E}_i)\}_{i=1}^m$, where $\cup_i \mathcal{V}_i = \mathcal{V}$ and $(\cup_i \mathcal{E}_i) \cup \mathcal{S}_{\mathrm{B}} = \mathcal{E}$. Then for any i, we have $(\mathcal{V}_{\mathrm{G}} \setminus \mathcal{S}_{\mathrm{G}}) \cap \mathcal{V}_i \neq \emptyset$.*

Proof. If the conclusion does not hold for some fixed i, then all the generators in $\mathcal{V}_i$ are binding. Denote

$$\mathbf{T} := \begin{bmatrix} \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} \\ \boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} \end{bmatrix}$$

and let $\mathcal{E}_0$ be the subset of $\mathcal{S}_{\mathrm{B}}$ satisfying $\forall e = (u, v) \in \mathcal{E}_0$, $u \notin \mathcal{V}_i$ and $v \in \mathcal{V}_i$. The condition that $\mathcal{S}_{\mathrm{G}} \perp \mathcal{S}_{\mathrm{B}}$ implies all the rows $\mathbf{T}_{\{j\}}$ for $j \in \mathcal{V}_i$ and $\mathbf{T}_{\{N+e\}}$ for $e \in \mathcal{E}_0$ must be independent of each other. Recall that for each edge, we use its integer index e and (u, v) interchangeably. Note that

$$\begin{aligned}
&\sum_{j \in \mathcal{V}_i} \mathbf{T}_{\{j\}} = \sum_{j \in \mathcal{V}_i} \sum_{e=(j,j') \in \mathcal{E}} (b_e \boldsymbol{e}_j^{\mathsf{T}} - b_e \boldsymbol{e}_{j'}^{\mathsf{T}}) \\
=& \sum_{\substack{j,j' \in \mathcal{V}_i \\ e=(j,j') \in \mathcal{E}_i}} (b_e \boldsymbol{e}_j^{\mathsf{T}} - b_e \boldsymbol{e}_{j'}^{\mathsf{T}}) + \sum_{j \in \mathcal{V}_i} \sum_{\substack{e=(j,j') \\ e \in \mathcal{E}_0}} (b_e \boldsymbol{e}_j^{\mathsf{T}} - b_e \boldsymbol{e}_{j'}^{\mathsf{T}}) \\
=& 0 + \sum_{\substack{e=(u,v) \in \mathcal{E}_0 \\ u \notin \mathcal{V}_i, v \in \mathcal{V}_i}} (b_e \boldsymbol{e}_v^{\mathsf{T}} - b_e \boldsymbol{e}_u^{\mathsf{T}}) = \sum_{e \in \mathcal{E}_0} \boldsymbol{C}_{v,e} \mathbf{T}_{\{N+e\}}.
\end{aligned}$$

Here $\boldsymbol{C}_{v,e}$ is the (v, e) element of matrix $\boldsymbol{C}$. It contradicts to $\mathcal{S}_{\mathrm{G}} \perp \mathcal{S}_{\mathrm{B}}$. Thereby, $(\mathcal{V}_{\mathrm{G}} \setminus \mathcal{S}_{\mathrm{G}}) \cap \mathcal{V}_i \neq \emptyset$. □

Now for a fixed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we can divide the edges into two disjoint sets $\mathcal{E}^{\mathrm{I}}$ and $\mathcal{E}^{\mathrm{II}}$ where

$$\mathcal{E}^{\mathrm{I}} := \{e \in \mathcal{E} \mid \mathcal{G}(\mathcal{V}, \mathcal{E} \setminus \{e\}) \text{ is not connected.}\},\ \mathcal{E}^{\mathrm{II}} := \mathcal{E} \setminus \mathcal{E}^{\mathrm{I}}.$$

For set $\mathcal{S}$, we use $\mathcal{S}(n)$ to denote the n^{th} smallest element in $\mathcal{S}$, and we denote its inverse operation as $\mathcal{S}^{-1}(\cdot)$.

Theorem 18. *Suppose $\mathcal{S}_{\mathrm{B}}$ partitions $\mathcal{G}$ into n disjoint connected subgraphs $\{\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)\}_{l=1}^{n}$. For $i \in \mathcal{V}_{\mathrm{G}}$, if $\exists l$ such that $\mathcal{V}_l \cap (\mathcal{V}_{\mathrm{G}} \setminus \mathcal{S}_{\mathrm{G}}) = \{i\}$, then for $j \in [N_{\mathrm{L}}]$, we have $\boldsymbol{J}_{i,j} = \mathbb{I}_{\{j+N_{\mathrm{G}} \in \mathcal{V}_l\}}$.*

Proof. Let $\mathcal{E}^{\mathrm{bri}} := \{e = (u, v) \in \mathcal{E} \mid u \in \mathcal{V}_l, v \notin \mathcal{V}_l)\}$ be the set of bridges connecting $\mathcal{V}_l$ and $\mathcal{V} \setminus \mathcal{V}_l$. We have $\mathcal{E}^{\mathrm{bri}} \subseteq \mathcal{S}_{\mathrm{B}}$. [4] We construct $\boldsymbol{q} \in \mathbb{R}^N$ as

$$\boldsymbol{q}_t = \begin{cases} 1, & t \le N_{\mathrm{L}} \text{ and } t + N_{\mathrm{G}} \in \mathcal{V}_l \\ 1, & N_{\mathrm{L}} < t < N - n \text{ and } \mathcal{S}_{\mathrm{G}}(t - N_{\mathrm{L}}) \in \mathcal{V}_l \\ \boldsymbol{C}_{v,e}, & e = (u, v) := \mathcal{S}_{\mathrm{B}}(t - N_{\mathrm{L}} - |\mathcal{S}_{\mathrm{G}}|) \in \mathcal{E}^{\mathrm{bri}} \\ 0, & \text{otherwise} \end{cases}.$$

By assumption, generator i has degree 1, so we let $i' \in \mathcal{V}$ be its only neighbor and $e^* := (i, i')$ be the edge linking i and i'. If $e^* \in \mathcal{S}_{\mathrm{B}}$, then $\mathcal{V}_l = \{i\}$, and by construction we have $\boldsymbol{R}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})\boldsymbol{q} = -\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{e}_i$. Otherwise, e^* is in neither $\mathcal{S}_{\mathrm{B}}$ nor $\mathcal{E}^{\mathrm{bri}}$, and by (5.7) we also have $\boldsymbol{R}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})\boldsymbol{q} = -\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{e}_i$.

$$\begin{aligned} \boldsymbol{R}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})\boldsymbol{q} &= \sum_{k \in \mathcal{V}_l \setminus \{i\}} \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{e}_k - \sum_{e=(u,v) \in \mathcal{E}^{\mathrm{bri}}} b_e(\boldsymbol{e}_u - \boldsymbol{e}_v) \\ &= \sum_{k \in \mathcal{V}_l \setminus \{i\}} \sum_{e=(k,k') \in \mathcal{E}} b_e(\boldsymbol{e}_k - \boldsymbol{e}_{k'}) - \sum_{e=(u,v) \in \mathcal{E}^{\mathrm{bri}}} b_e(\boldsymbol{e}_u - \boldsymbol{e}_v) \\ &= \sum_{k \in \mathcal{V}_l} \sum_{e=(k,k') \in \mathcal{E}} b_e(\boldsymbol{e}_k - \boldsymbol{e}_{k'}) - \sum_{e=(u,v) \in \mathcal{E}^{\mathrm{bri}}} b_e(\boldsymbol{e}_u - \boldsymbol{e}_v) - b_{e^*}(\boldsymbol{e}_i - \boldsymbol{e}_{i'}) \\ &= \sum_{\substack{k,k' \in \mathcal{V}_l: \\ e=(k,k') \in \mathcal{E}}} b_e(\boldsymbol{e}_k - \boldsymbol{e}_{k'}) + \sum_{\substack{k \in \mathcal{V}_l, k' \in \mathcal{V} \setminus \mathcal{V}_l: \\ e=(k,k') \in \mathcal{E}}} b_e(\boldsymbol{e}_k - \boldsymbol{e}_{k'}) \\ &\quad - \sum_{e=(u,v) \in \mathcal{E}^{\mathrm{bri}}} b_e(\boldsymbol{e}_u - \boldsymbol{e}_v) - b_{e^*}(\boldsymbol{e}_i - \boldsymbol{e}_{i'}) \\ &= - b_{e^*}(\boldsymbol{e}_i - \boldsymbol{e}_{i'}) = -\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{e}_i. \end{aligned} \tag{5.7}$$

Thereby, (5.5) implies that $\boldsymbol{q}$ is the i^{th} row of $\boldsymbol{U}$. Therefore, we have $\boldsymbol{J}_{i,j} = \boldsymbol{q}_j = \mathbb{I}_{j+N_{\mathrm{G}} \in \mathcal{V}_l}$ for $j \in [N_{\mathrm{L}}]$. □

Informally, Theorem 18 indicates if, after the removal of $\mathcal{S}_{\mathrm{B}}$, generator i is the only non-binding generator in some connected subgraph, then the small change in any load within the same subgraph will directly affects the generation at i by the same

[4] Otherwise, $\mathcal{V}_l$ and $\mathcal{V} \setminus \mathcal{V}_l$ will be connected by a path consisting of edges in $\mathcal{E} \setminus \mathcal{S}_{\mathrm{B}}$. It contradicts to the fact that $\mathcal{V}_l$ is one of the subgraphs partitioned by $\mathcal{S}_{\mathrm{B}}$.

amount, while the change in other loads outside the subgraph will have no effects on i at all.

Radial Networks

When $\mathcal{G}$ is a tree, $\mathcal{E}^{\mathrm{I}} = \mathcal{E}$. An set of $|\mathcal{S}_{\mathrm{B}}|$ edges will always partition the graph into $|\mathcal{S}_{\mathrm{B}}| + 1$ connected subgraphs. Lemma 18, as well as $|\mathcal{S}_{\mathrm{G}}| + |\mathcal{S}_{\mathrm{B}}| = N_{\mathrm{G}} - 1$, indicates that each subgraph should contain exactly one non-binding generator. Using Theorem 18 again and we will obtain

$$\boldsymbol{J}_{i,j}(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}}) = \begin{cases} 1, & \text{if } i \Leftrightarrow j + N_{\mathrm{G}} \text{ in } \mathcal{G}(\mathcal{V}, \mathcal{E} \setminus \mathcal{S}_{\mathrm{B}}), i \notin \mathcal{S}_{\mathrm{G}} \\ 0, & \text{otherwise} \end{cases}.$$

Here $i \Leftrightarrow j + N_{\mathrm{G}}$ means generator i and load j (i.e., bus $j + N_{\mathrm{G}}$) are connected by some path. Informally, it means for tree network, $\mathcal{S}_{\mathrm{B}}$ partitions the network into subgraphs and there will be exactly one non-binding generator in each subgraph. Any change in load will lead to the same amount of change in the generator within the same subgraph as long as $(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})$ does not alter.

Therefore, when the binding sets are $\mathcal{S}_{\mathrm{G}}$ and $\mathcal{S}_{\mathrm{B}}$, nodel prices $\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$ can be computed as

$$\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l} = \boldsymbol{J}^{\mathsf{T}} \boldsymbol{f}.$$

Since each column of $\boldsymbol{J}$ is a basis vector, the values in $\frac{\partial \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{s}^g)^*}{\partial \boldsymbol{s}^l}$ are taken from some entries of $\boldsymbol{f}$. As a result, all nodal prices must be bounded by $\min \boldsymbol{f}_j$ and $\max \boldsymbol{f}_j$. For practical problems, the coefficients in $\boldsymbol{f}$ usually fall within a narrow range, and it partially explains why the values of $u^{\phi}_{jk}, v^{\phi}_{jk}$ in Chapter 3 are typically close to each other for multi-phase systems.

5.4 Worst Case Sensitivity

As we can see in previous sections, if the power network is a tree, the sensitivity is always binary, implying increasing one unit of load requires increasing one unit of generation somewhere and all other generators remain unaffected. Starting from this section, we want to study for networks with general topology, specifically with cycles, how the sensitivity would change. There are indeed many sensitivity problems that can be formulated. Before we do so, it will be helpful to make concrete the link between continuity and the Jacobian matrix. To avoid notational overload we will refer to arbitrary functions and sets and then provide the definition specific to $\mathcal{OPF}$.

Recall that a function $h : \mathcal{D} \to \mathbb{R}^n$ with $\mathcal{D}$ an open subset of $\mathbb{R}^n$ is said to be Lipschitz on $\mathcal{D}$ if there exists some $L \geq 0$ such that

$$\|h(\boldsymbol{x}) - h(\boldsymbol{x}')\| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\| \tag{5.8}$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$. Suppose that the Jacobian $\boldsymbol{J} := \frac{\partial h}{\partial \boldsymbol{x}}$ exists and is continuous on D. Then if for some convex subset $\mathcal{B} \subseteq \mathcal{D}$, there exists a constant $K \geq 0$ such that

$$\left\| \frac{\partial h(\boldsymbol{x})}{\partial \boldsymbol{x}} \right\| \leq K$$

on $\mathcal{B}$, then (5.8) holds for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{B}$ with $L = K$. This establishes a clear link between a bound on the norm of the Jacobian and the Lipschitz constant of a function. We now define the notion of Lipschitz continuity for a generator-load pair, and then formulate three sensitivity definitions.

Definition 22. *The matrices $\boldsymbol{C}, \boldsymbol{B}$ are fixed, as is the cost vector $\boldsymbol{f} \in \Omega_f$ and $\boldsymbol{\xi} \in \widetilde{\Omega}_{\xi}(\boldsymbol{f})$. Select a generator i and load j. The pair (i, j) is said to be C-Lipschitz if for all $\delta > 0$ and $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \Omega_{s^l}(\boldsymbol{\xi})$ such that $|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'_j| \leq \delta$ and $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}'_k$ for all $k \neq j$, we have that*

$$|\mathcal{OPF}_i(\boldsymbol{\alpha}) - \mathcal{OPF}_i(\boldsymbol{\alpha}')| < C\delta,$$

where $\mathcal{OPF}_i(\cdot)$ denotes the i^{th} coordinate of $\mathcal{OPF}(\cdot)$.

This Lipschitz-like definition forms the basis of the sensitivity analysis formulation we are proposing. In the remainder of this section we formulate several sensitivity problems that will be of interest to grid operators.

Remark 7. *Recall that in our notation, when we refer to the (i, j)-generator-load pair, this corresponds to vertices $(i, N_{\mathrm{G}} + j)$.*

Problem 1: SISO Sensitivity

In this formulation we consider the problem of computing the worst-case sensitivity of the generator-load pair (i, j). We use SISO to mean single-input, single-output, i.e. the change in one output when one input is changed. Recall that according to our indexing of vertices, load j corresponds to the vertex $N_G + j$.

Definition 23. *The (SISO) sensitivity of generator i with respect to load j is the minimum value which we denote by $C_{i \leftarrow j}$, such that (i, j) is a $C_{i \leftarrow j}$-Lipschitz pair, i.e., $C_{i \leftarrow j}$ is the minimal C such that $|\mathcal{OPF}_i(\boldsymbol{\alpha}) - \mathcal{OPF}_i(\boldsymbol{\alpha}')| < C\delta$ for every $\boldsymbol{\alpha}, \boldsymbol{\alpha}'$ that differ only in their j^{th} coordinates with $|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'_j| \leq \delta$.*

Problem 2: Worst-Case SISO Sensitivity

In the SISO sensitivity formulation, it was assumed that all the network parameters and the OPF cost function were fixed. In this version of the problem we allow the network parameters to change (apart from those which define the network structure, e.g., $\boldsymbol{C}$, the graph incidence matrix).

Definition 24. *The worst-case (SISO) sensitivity of generator i with respect to load j is*

$$C^{\mathrm{wc}}_{i\leftarrow j} := \max_{\boldsymbol{f}\in\Omega_{\boldsymbol{f}}} \max_{\boldsymbol{\xi}\in\widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f})} C_{i\leftarrow j}. \tag{5.9}$$

The ability to allow parameter variations means $C^{\mathrm{wc}}_{i\leftarrow j}$ provides information about various network scenarios. For example, a generator instantaneously going offline can be modeled by $\overline{\boldsymbol{p}}, \underline{\boldsymbol{p}} \to \epsilon$, where ϵ is a small constant. Taking $\epsilon = 0$ would potentially break the independence conditions we require. In practice a small constant such as $\epsilon = 10^{-5}$ suffices.

Problem 3: MISO Sensitivity

Consider a set of m load buses $\mathcal{V}'_{\mathrm{L}} \subseteq \mathcal{V}_{\mathrm{L}}$ and let $\mathcal{L}$ denote the set of indices corresponding to those loads. The MISO part of the definition refers to the fact that here, we are interested in how a single output (generation) changes when multiple inputs (loads) are allowed to simultaneously change. To make this definition concrete, we must first modify Definition 22.

Definition 25. *Assume that $\boldsymbol{C}, \boldsymbol{B}$ are fixed, as is the cost vector $\boldsymbol{f} \in \Omega_{\boldsymbol{f}}$ and $\boldsymbol{\xi} \in \widetilde{\Omega}_{\boldsymbol{\xi}}(\boldsymbol{f})$. We say that $(i, \mathcal{L})$ is $C^{(m)}$-Lipschitz if for all $\delta > 0$ and $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \Omega_{s^l}(\boldsymbol{\xi})$ such that $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| \leq \delta$ and $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}'_k$ for all $k \notin \mathcal{L}$, there exists a constant $C^{(m)}$ such that*

$$\|\mathcal{OPF}_i(\boldsymbol{\alpha}) - \mathcal{OPF}_i(\boldsymbol{\alpha}')\| < C^{(m)}\delta.$$

Definition 26. *The (MISO) sensitivity of generator i with respect to the set $\mathcal{L}$ of loads, denoted by $C_{i\leftarrow\mathcal{L}}$, is the minimum value of $C^{(m)}$ such that $(i, \mathcal{L})$ is C^m-Lipschitz.*

The worst-case MISO sensitivity problem can then be derived analogously to Definition 24.

Computing Worst Case SISO Sensitivity

It can be shown that the worst-case sensitivity is computed by solving a discrete optimization problem based on the binding constraint formulation of the Jacobian

matrix.

$$C^{\mathrm{wc}}_{i\leftarrow j} = \max_{\substack{\mathcal{S}_{\mathrm{G}}\in\mathcal{V}_{\mathrm{G}},\mathcal{S}_{\mathrm{B}}\in\mathcal{E}\\ |\mathcal{S}_{\mathrm{G}}|+|\mathcal{S}_{\mathrm{B}}|=N_{\mathrm{G}}-1\\ \mathcal{S}_{\mathrm{G}}\perp\mathcal{S}_{\mathrm{B}}}} |\boldsymbol{J}_{i,j}|. \tag{5.10}$$

Unfortunately (5.10) is a non-convex, discrete optimization problem and thus intractable in general. However, we provide Algorithm 3 that produces small subgraphs such that a brute-force search is possible.

Algorithm 3 Decomposition of the computation of $C^{\mathrm{wc}}_{i\leftarrow j}$.

Input: $\boldsymbol{B}, \boldsymbol{C}, i \in [N_{\mathrm{G}}], j \in [N_{\mathrm{L}}], \mathcal{G}(\mathcal{V}_{\mathrm{G}} \cup \mathcal{V}_{\mathrm{L}}, \mathcal{E})$
Output: $C^{\mathrm{wc}}_{i\leftarrow j}$
  for $e = (u, v)$ in $\mathcal{E}^{\mathrm{bri}}$ **do**
    if $u, v \notin \mathcal{V}_{\mathrm{G}}$ and $i \Leftrightarrow j + N_{\mathrm{G}}$ in $\mathcal{G}(\mathcal{V}, \mathcal{E} \setminus \{e\})$ **then**
      e partitions $\mathcal{G}$ into $\mathcal{G}_1$ and $\mathcal{G}_2$ (assume $i, j + N_{\mathrm{G}}$ are both in $\mathcal{G}_1$)
      if $\mathcal{G}_2$ contains any vertex in $\mathcal{V}_{\mathrm{G}}$ **then**
        Replace $\mathcal{G}_2$ by a single generator
      else
        Replace $\mathcal{G}_2$ by a single load
      end if
    end if
  end for
  Find a shortest path connecting i and $j + N_{\mathrm{G}}$
  Get $\{\mathcal{G}_l\}_{l=1}^{m}$ and add p_l, q_l to subgraphs
  for $l = 0$ to $m - 1$ **do**
    call subroutine to compute $C^{\mathrm{wc}}_{p_l\leftarrow q_{l+1}}$
  end for
  $C^{\mathrm{wc}}_{i\leftarrow j} \leftarrow \prod_{l=0}^{m-1} C^{\mathrm{wc}}_{p_l\leftarrow q_{l+1}}$
  return $C^{\mathrm{wc}}_{i\leftarrow j}$

This algorithm aims to reduce the computational complexity by breaking the large-scale computation down into independent smaller tasks, which are usually much easier than the original problem and can be processed in parallel.

Here a bridge is an edge in $\mathcal{E}$ whose deletion disconnects the graph. Define $\mathcal{E}^{\mathrm{bri}}$ as the set of bridges in $\mathcal{E}$. In a not necessarily connected graph $\mathcal{G}'$, we say $i \Leftrightarrow j + N_{\mathrm{G}}$ if there exists a path between nodes i and $j + N_{\mathrm{G}}$.

In the first step, for any bridge $e = (u, v) \in \mathcal{E}^{\mathrm{bri}}$ that partitions $\mathcal{G}$ into $\mathcal{G}_1$ and $\mathcal{G}_2$, if $i \Leftrightarrow j + N_{\mathrm{G}}$ after e is deleted, then without loss of generality we assume both i and $j + N_{\mathrm{G}}$ are in $\mathcal{G}_1$. In this case we can replace the whole of $\mathcal{G}_2$ by a single bus. The rule is if $\mathcal{G}_2$ contains only load buses then it will be replaced by a single load, else it is replaced by a single generator.

In the second step, we find a shortest path (in terms of the number of edges along the path) connecting i and $j + N_{\mathrm{G}}$, and the bridges along the path will partition the graph into subgraphs $\{\mathcal{G}_l(\mathcal{V}_l, \mathcal{E}_l)\}_{l=1}^{m}$. Assume the indices are assigned such that $\mathcal{G}_{l-1}$ is always closer to i than $\mathcal{G}_l$. At the location of each bridge connecting $\mathcal{G}_{l-1}$ and $\mathcal{G}_l$, we add a single load q_{l-1} to $\mathcal{G}_{l-1}$ and a single generator p_{l-1} to $\mathcal{G}_l$, as shown in Figure 5.3. For notational consistency, we refer to i as generator p_0 and j as load q_m. Then the computation of $C^{\mathrm{wc}}_{i\leftarrow j}$ can be composed as $\prod_{l=0}^{m-1} C^{\mathrm{wc}}_{p_l\leftarrow q_{l+1}}$, where each $C^{\mathrm{wc}}_{p_l\leftarrow q_{l+1}}$ only depends on computing the sensitivity for smaller graphs.

5.5 Examples

We now consider two numerical examples that demonstrate Algorithm 3 presented in the previous section. Both examples make use of the IEEE 9-bus test network, full details of the model can be found in the MATPOWER toolbox [81].

9-Bus Example

The IEEE 9-bus test network is shown in Figure 5.4. The network consists of 3 generators, $\{1, 2, 3\}$ and 6 loads, $\{4, \ldots, 9\}$. In Table 5.1 we have computed the worst-case SISO sensitivity for every generator load pair in the network.

Table 5.1: E.g. 1: worst-case SISO sensitivity for the 9-bus network.

$\mathcal{V}_{\mathrm{G}}$ \ $\mathcal{V}_{\mathrm{L}}$	4	5	6	7	8	9
1	1.0000	1.3935	2.0650	2.4748	1.9389	1.3244
2	2.4236	2.9560	1.7024	1.4748	1.0000	2.0081
3	2.5162	1.9838	1.0000	1.3847	1.6595	3.0081

This example shows that the network is most sensitive to perturbations to bus 9 as felt by generator 3. It is interesting to note that the distance (in terms of number of lines between the pair) between this pair of buses is as large as it could be for a network of this topology. The worst-case sensitivities were computed using a brute-force search over the discrete sets $(\mathcal{S}_{\mathrm{G}}, \mathcal{S}_{\mathrm{B}})$. This example is small enough for such an approach to easily be computationally tractable. In the next sub-section we consider an example where this is not the case.

27-Bus Example

This example computes the worst-case SISO sensitivity of a 27-bus network. The network is constructed by chaining together three copies of the 9-bus network

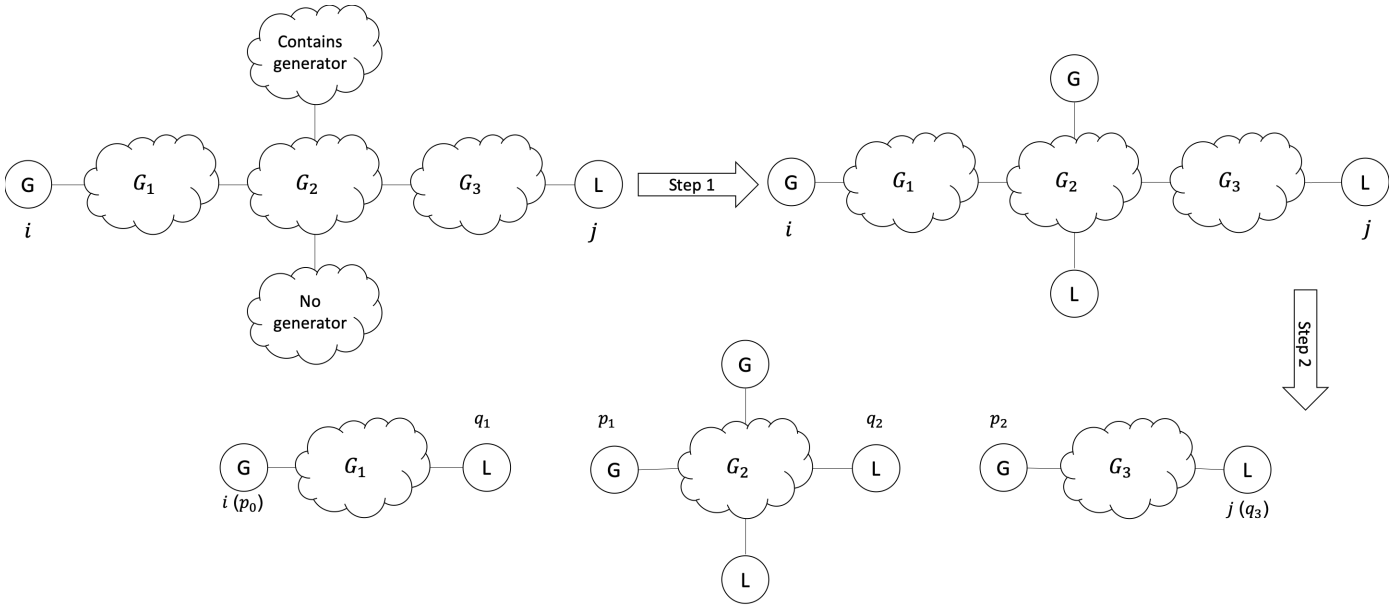

Figure 5.3: Algorithm to decompose the worst-case SISO sensitivity of generator i with respect to load j into computation involving smaller graphs. Step1: Replace the off-path subgraphs by a single node. Step 2: Partition the on-path subgraphs and complement each subgraph by adding a pair of generator/load.

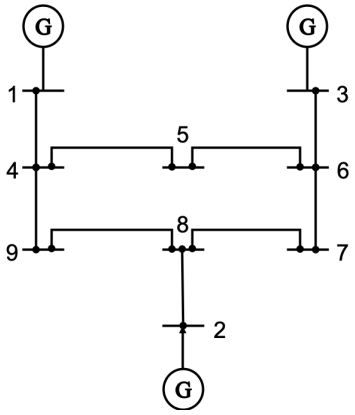

Figure 5.4: IEEE 9-bus network.

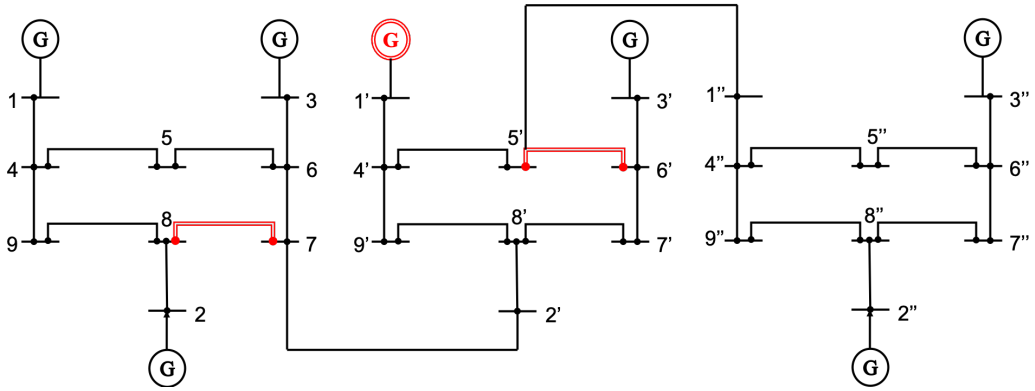

Figure 5.5: A 27-bus auxiliary network constructed by chaining three identical 9-bus networks together. Nodes and edges in red indicate that their corresponding generation and flow constraints are binding for every generator-load worst-case SISO sensitivity pairing in Table 5.3.

described in the previous example, it is illustrated in Figure 5.5. This system was chosen to demonstrate Algorithm 3 as it easily decomposes into three 9-bus subgraphs. The worst-case sensitivity can then be computed (in parallel) for each of the subgraphs, with the global solution then given by multiplying the sensitivities of each of the subproblems together.

In Table 5.2 (on the next page) we show a subset of the SISO worst-case generator-load pairs. We have chosen to show the results of the computation from loads located at the far right of the network to generators at the far left. From the decomposition algorithm, we know that these values are likely to be larger than those of pairings that are closer together because the graph in the middle, i.e. the 9-bus network with nodes labeled with a single prime, e.g. $4'$, acts as a multiplier for generator-load

pairs that have a shortest path passing through it.

In Table 5.3, for every sensitivity pairing we have listed the binding constraints, i.e., the edge flows and generations that hit their limits. Observe that generator $1'$ and lines $(7, 8)$, $(5', 6')$ are active for all pairings and hence omitted from the table (they are, however, marked in red in Figure 5.5).

Table 5.2: E.g. 2: worst-case SISO sensitivity for the 27-bus chained network.

$\mathcal{V}_{\mathrm{G}}$ \ $\mathcal{V}_{\mathrm{L}}$	4”	5”	6”	7”	8”	9”
1	7.3155	10.1942	15.1069	18.1045	14.1843	9.6889
2	4.3595	6.0750	9.0026	10.7889	8.4528	5.7739
3	4.0933	5.7040	8.4528	10.1301	7.9366	5.4213

Table 5.3: Binding generators/branches corresponding to the worst-case SISO sensitivity for the 27-bus chained network. Since generator $1'$ and the branch $(5', 6')$ (marked by red double-lines in Fig. 5.5) are always binding for all 18 pairs, this table only lists other binding constraints besides $1'$ and $(5', 6')$.

$\mathcal{V}_{\mathrm{G}}$ \ $\mathcal{V}_{\mathrm{L}}$	4”	5”	6”
1	3, 2”, (6”, 7”)	3, 3”, (4”, 5”)	3, 3”, (6”, 7”)
2	3, 2”, (6”, 7”)	3, 3”, (4”, 5”)	3, 3”, (6”, 7”)
3	2, 2”, (6”, 7”)	2, 3”, (4”, 5”)	2, 3”, (6”, 7”)

$\mathcal{V}_{\mathrm{G}}$ \ $\mathcal{V}_{\mathrm{L}}$	7”	8”	9”
1	3, 3”, (7”, 8”)	3, 2”, (6”, 7”)	3, 2”, (6”, 7”)
2	3, 3”, (7”, 8”)	3, 2”, (6”, 7”)	3, 2”, (6”, 7”)
3	2, 3”, (7”, 8”)	2, 2”, (6”, 7”)	2, 2”, (6”, 7”)

BIBLIOGRAPHY

[1] Sanjeev Arora et al. "Simple, efficient, and neural algorithms for sparse coding". In: *Proceedings of Machine Learning Research* 40 (Jan. 2015).

[2] Xiaoqing Bai et al. "Semidefinite programming for optimal power flow problems". In: *International Journal of Electrical Power & Energy Systems* 30.6-7 (2008), pp. 383–392.

[3] Mesut Baran and Felix F Wu. "Optimal capacitor placement on radial distribution systems". In: *IEEE Trans. Power Delivery* 4.1 (1989), pp. 725–734.

[4] Mesut Baran and Felix F Wu. "Optimal sizing of capacitors placed on a radial distribution system". In: *IEEE Trans. Power Delivery* 4.1 (1989), pp. 735–743.

[5] Alexander I. Barvinok. "Problems of distance geometry and convex properties of quadratic maps". In: *Discrete & Computational Geometry* 13.2 (1995), pp. 189–202.

[6] Mohammadhafez Bazrafshan and Nikolaos Gatsis. "Convergence of the Z-bus method for three-phase distribution load-flow with ZIP loads". In: *IEEE Transactions on Power Systems* 33.1 (2017), pp. 153–165.

[7] R Berg, ES Hawkins, and WW Pleines. "Mechanized calculation of unbalanced load flow on radial distribution circuits". In: *IEEE Transactions on power apparatus and systems* 4 (1967), pp. 415–421.

[8] Andrey Bernstein and Emiliano Dall'Anese. "Real-time feedback-based optimization of distribution grids: A unified approach". In: *IEEE Transactions on Control of Network Systems* 6.3 (2019), pp. 1197–1209.

[9] Andrey Bernstein et al. "Load flow in multiphase distribution networks: Existence, uniqueness, non-singularity and linear models". In: *IEEE Transactions on Power Systems* 33.6 (2018), pp. 5832–5843.

[10] Daniel Bienstock and Apurv Shukla. "Variance-Aware Optimal Power Flow: addressing the trade-off between cost, security and variability". In: *IEEE Transactions on Control of Network Systems* (2019).

[11] Daniel Bienstock and Abhinav Verma. "Strong NP-hardness of AC power flows feasibility". In: *Operations Research Letters* 47.6 (2019), pp. 494–501.

[12] Edward Bierstone. "Differentiable functions". In: *Boletim da Sociedade Brasileira de Matemática-Bulletin/Brazilian Mathematical Society* 11.2 (1980), pp. 139–189.

[13] Edward Bierstone and Pierre D Milman. "Semianalytic and subanalytic sets". In: *Publications Mathématiques de l'Institut des Hautes Études Scientifiques* 67.1 (1988), pp. 5–42.

[14] Subhonmesh Bose et al. “Equivalent relaxations of optimal power flow”. In: *IEEE Transactions on Automatic Control* 60.3 (2014), pp. 729–742.

[15] Subhonmesh Bose et al. “Quadratically constrained quadratic programs on acyclic graphs with application to power flow”. In: *IEEE Transactions on Control of Network Systems* 2.3 (2015), pp. 278–287.

[16] Nicolas Boumal. “Nonconvex phase synchronization”. In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2355–2377.

[17] Samuel Burer and Renato DC Monteiro. “Local minima and convergence in low-rank semidefinite programming”. In: *Mathematical Programming* 103.3 (2005), pp. 427–444.

[18] Mary B Cain, Richard P O’neill, Anya Castillo, et al. “History of optimal power flow and formulations”. In: *Federal Energy Regulatory Commission* 1 (2012), pp. 1–36.

[19] Emmanuel J Candès and Benjamin Recht. “Exact matrix completion via convex optimization”. In: *Foundations of Computational mathematics* 9.6 (2009), p. 717.

[20] Emmanuel J Candès and Terence Tao. “The power of convex relaxation: Near-optimal matrix completion”. In: *IEEE Transactions on Information Theory* 56.5 (2010), pp. 2053–2080.

[21] J Carpentier. “Contribution to the economic dispatch problem”. In: *Bulletin de la Societe Francoise des Electriciens* 3.8 (1962), pp. 431–447.

[22] Tsai-Hsiang Chen et al. “Distribution system power flow analysis – a rigid approach”. In: *EEE Transactions on Power Delivery* 6.3 (July 1991).

[23] Konstantina Christakou et al. “Efficient computation of sensitivity coefficients of node voltages and line currents in unbalanced radial electrical distribution networks”. In: *IEEE Transactions on Smart Grid* 4.2 (2013), pp. 741–750.

[24] Emiliano Dall’Anese, Hao Zhu, and Georgios B Giannakis. “Distributed optimal power flow for smart microgrids”. In: *IEEE Transactions on Smart Grid* 4.3 (2013), pp. 1464–1475.

[25] Emiliano Dall’Anese et al. “Optimal regulation of virtual power plants”. In: *IEEE Transactions on Power Systems* 33.2 (2017), pp. 1868–1881.

[26] K. Dvijotham and D. K. Molzahn. “Error bounds on the DC power flow approximation: A convex relaxation approach”. In: *2016 IEEE 55th Conference on Decision and Control*. IEEE. 2016, pp. 2411–2418.

[27] Masoud Farivar and Steven H Low. “Branch flow model: Relaxations and convexification–Part I”. In: *IEEE Transactions on Power Systems* 28.3 (2013), pp. 2554–2564.

[28] Masoud Farivar and Steven H Low. “Branch Flow Model: Relaxations and Convexification–Part II”. In: *IEEE Transactions on Power Systems* 3.28 (2013), pp. 2565–2572.

[29] Andrei Mikhailovich Gabrièlov. “Projections of semi-analytic sets”. In: *Functional Analysis and its applications* 2.4 (1968), pp. 282–291.

[30] Lingwen Gan and Steven H Low. “Convex relaxations and linear approximation for optimal power flow in multiphase radial networks”. In: *2014 Power Systems Computation Conference*. IEEE. 2014, pp. 1–9.

[31] Lingwen Gan et al. “Exact convex relaxation of optimal power flow in radial networks”. In: *IEEE Transactions on Automatic Control* 60.1 (2015), pp. 72–87.

[32] Rong Ge, Chi Jin, and Yi Zheng. “No spurious local minima in nonconvex low rank problems: A unified geometric analysis”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1233–1242.

[33] Rong Ge, Jason D Lee, and Tengyu Ma. “Matrix completion has no spurious local minimum”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2973–2981.

[34] S Gopinath et al. “Proving global optimality of ACOPF solutions”. In: *Electric Power Systems Research* 189 (2020), p. 106688.

[35] Paul R Gribik et al. “Optimal power flow sensitivity analysis”. In: *IEEE Transactions on Power Systems* 5.3 (1990), pp. 969–976.

[36] Robert Hardt. “Some analytic bounds for subanalytic sets”. In: *Differential geometric control theory, Progress in Math* 27 (1983), pp. 259–267.

[37] Hein van der Holst. “Graphs whose positive semi-definite matrices have nullity at most two”. In: *Linear Algebra and its Applications* 375 (2003), pp. 1–11.

[38] Roger A Horn, Roger A Horn, and Charles R Johnson. *Matrix analysis*. Cambridge university press, 1990.

[39] IEA. *Renewable Energy Market Update 2021*. 2021. URL: `https://www.iea.org/news/renewables-are-stronger-than-ever-as-they-power-through-the-pandemic`.

[40] Rabih A Jabr. “Radial distribution load flow using conic programming”. In: *IEEE transactions on power systems* 21.3 (2006), pp. 1458–1459.

[41] Rabih A Jabr, Alun H Coonick, and Brian J Cory. “A primal-dual interior point method for optimal power flow dispatching”. In: *IEEE Transactions on Power Systems* 17.3 (2002), pp. 654–662.

[42] Joakim Jaldén, Cristoff Martin, and Björn Ottersten. "Semidefinite programming for detection in linear systems-optimality conditions and space-time decoding". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* Vol. 4. IEEE. 2003, pp. IV–9.

[43] W. H. Kersting. *Distribution systems modeling and analysis*. CRC, 2002.

[44] M. Laughton. "Analysis of unbalanced polyphase networks by the method of phase co-ordinates. Part 1: System representation in phase frame of reference". In: *Proc. Inst. Electr. Eng.* 115.8 (1968).

[45] Javad Lavaei and Steven H Low. "Zero duality gap in optimal power flow problem". In: *IEEE Transactions on Power Systems* 27.1 (2012), pp. 92–107.

[46] Karsten Lehmann, Alban Grastien, and Pascal Van Hentenryck. "AC-feasibility on tree networks is NP-hard". In: *IEEE Transactions on Power Systems* 31.1 (2016), pp. 798–801.

[47] Bernard C Lesieutre et al. "Examining the limits of the application of semidefinite programming to power flow problems". In: *2011 49th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE. 2011, pp. 1492–1499.

[48] Fangxing Li and Rui Bo. "DCOPF-based LMP simulation: algorithm, comparison with ACOPF, and sensitivity". In: *IEEE Transactions on Power Systems* 22.4 (2007), pp. 1475–1485.

[49] Zhengshuo Li et al. "Sufficient conditions for exact relaxation of complementarity constraints for storage-concerned economic dispatch". In: *IEEE Transactions on Power Systems* 31.2 (2015), pp. 1653–1654.

[50] Stanisław Łojasiewicz. "On semi-analytic and subanalytic geometry". In: *Banach Center Publications* 34.1 (1995), pp. 89–104.

[51] Steven H Low. "Convex relaxation of optimal power flow–Part II: Exactness". In: *IEEE Transactions on Control of Network Systems* 1.2 (2014), pp. 177–189.

[52] Steven H Low. "Convex relaxation of optimal power flow-Part I: Formulations and equivalence". In: *IEEE Transactions on Control of Network Systems* 1.1 (2014), pp. 15–27.

[53] Cheng Lu et al. "Tightness of a new and enhanced semidefinite relaxation for MIMO detection". In: *SIAM Journal on Optimization* 29.1 (2019), pp. 719–742.

[54] Jan Machowski, Janusz Bialek, and Jim Bumby. *Power system dynamics: stability and control*. John Wiley & Sons, 2011.

[55] Ramtin Madani, Morteza Ashraphijuo, and Javad Lavaei. “Promises of conic relaxation for contingency-constrained optimal power flow problem”. In: *IEEE Transactions on Power Systems* 31.2 (2015), pp. 1297–1307.

[56] Ramtin Madani, Somayeh Sojoudi, and Javad Lavaei. “Convex relaxation for optimal power flow problem: Mesh networks”. In: *IEEE Transactions on Power Systems* 30.1 (2014), pp. 199–211.

[57] Daniel K Molzahn and Ian A Hiskens. “A Survey of Relaxations and Approximations of the Power Flow Equations”. In: *Foundations and Trends® in Electric Energy Systems* 4.1-2 (2019), pp. 1–221.

[58] Daniel K Molzahn et al. “A Laplacian-based approach for finding near globally optimal solutions to OPF problems”. In: *IEEE Transactions on Power Systems* 32.1 (2017), pp. 305–315.

[59] James A Momoh, Rambabu Adapa, and ME El-Hawary. “A review of selected optimal power flow literature to 1993. I. Nonlinear and quadratic programming approaches”. In: *IEEE transactions on power systems* 14.1 (1999), pp. 96–104.

[60] James A Momoh, ME El-Hawary, and Ramababu Adapa. “A review of selected optimal power flow literature to 1993. II. Newton, linear programming and interior point methods”. In: *IEEE Transactions on Power Systems* 14.1 (1999), pp. 105–111.

[61] Gábor Pataki. “On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues”. In: *Mathematics of operations research* 23.2 (1998), pp. 339–358.

[62] K. Purchala et al. “Usefulness of DC power flow for active power flow analysis”. In: *IEEE Power Engineering Society General Meeting, 2005*. IEEE. 2005, pp. 454–459.

[63] L. Roald and D. K. Molzahn. “Implied Constraint Satisfaction in Power System Optimization: The Impacts of Load Variations”. In: *arXiv preprint arXiv:1904.01757* (2019).

[64] Walter Rudin et al. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1964.

[65] KP Schneider et al. “Analytic considerations and design basis for the IEEE distribution test feeders”. In: *IEEE Transactions on power systems* 33.3 (2017), pp. 3181–3188.

[66] Stephan Singer, Jean-Philippe Denruyter, and Deniz Yener. “The energy report: 100% renewable energy by 2050”. In: *Towards 100% renewable energy*. Springer, 2017, pp. 379–383.

[67] Somayeh Sojoudi and Javad Lavaei. “Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure”. In: *SIAM Journal on Optimization* 24.4 (2014), pp. 1746–1778.

[68] MS Srinivas. "Distribution load flows: a brief review". In: *2000 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 00CH37077)*. Vol. 2. IEEE. 2000, pp. 942–945.

[69] Ju Sun, Qing Qu, and John Wright. "Complete dictionary recovery over the sphere I: Overview and the geometric picture". In: *IEEE Transactions on Information Theory* 63.2 (2016), pp. 853–884.

[70] Ju Sun, Qing Qu, and John Wright. "When are nonconvex problems not scary?" In: *arXiv preprint arXiv:1510.06096* (2015).

[71] Victor A Toponogov. *Differential geometry of curves and surfaces*. Springer, 2006.

[72] Abhinav Verma. "Power Grid Security Analysis : An Optimization Approach". PhD thesis. Columbia University, 2009.

[73] Ahmed S Zamzam, Nicholas D Sidiropoulos, and Emiliano Dall'Anese. "Beyond relaxation and Newton–Raphson: Solving AC OPF for multi-phase systems with renewables". In: *IEEE Transactions on Smart Grid* 9.5 (2018), pp. 3966–3975.

[74] Baosen Zhang and David Tse. "Geometry of injection regions of power networks". In: *IEEE Transactions on Power Systems* 28.2 (2013), pp. 788–797.

[75] Richard Y Zhang and Javad Lavaei. "Sparse semidefinite programs with near-linear time complexity". In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE. 2018, pp. 1624–1631.

[76] Changhong Zhao, Emiliano Dall'Anese, and Steven H Low. "Convex Relaxation of OPF in Multiphase Radial Networks with Delta Connections". In: *Proceedings of the 10th Bulk Power Systems Dynamics and Control Symposium*. 2017.

[77] Changhong Zhao, Emiliano Dall'Anese, and Steven H Low. *Optimal power flow in multiphase radial networks with delta connections*. Tech. rep. National Renewable Energy Lab. (NREL), Golden, CO (United States), 2017.

[78] Zhu Zhongming et al. "Global EV Outlook 2021". In: (2021).

[79] Fengyu Zhou, James Anderson, and Steven H Low. "Differential privacy of aggregated DC optimal power flow data". In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 1307–1314. DOI: 10.23919/ACC.2019.8815257.

[80] Ray Daniel Zimmerman. "Comprehensive distribution power flow: modeling, formulation, solution algorithms and analysis". PhD thesis. Cornell University, 1995.

[81] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. “MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education”. In: *IEEE Transactions on power systems* 26.1 (2010), pp. 12–19.