Technische Universität München

Fakultät für Mathematik

On the Mathematics of Energy System Optimization

Network Models, Decomposition, and Economic Incentives

Paul Melvin Stursberg

Vollständiger Abdruck der von der promotionsführenden Einrichtung Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Michael Ulbrich
Prüfer der Dissertation:	1. Prof. Dr. Peter Gritzmann
	2. Prof. Dr. Thomas Hamacher
	3. Prof. Dr. Alexander Martin (schriftliche Beurteilung)

Die Dissertation wurde am 28.02.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 02.08.2019 angenommen.

Acknowledgements

Parts of the research underlying this thesis were supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81, and by the German Federal Ministry for Economic Affairs and Energy (FKZ 03ET4029) on the basis of a decision by the German Bundestag.

Abstract

In this thesis, we analyze mathematical structures of optimization problems and algorithms in the context of the analysis and design of electrical power systems.

We present a new way to improve the representation of electrical power flows in a simple, network-flow based optimization model. We furthermore analyze the implications with respect to the set of feasible solutions for the two most commonly used network models. We provide a combinatorial characterization of the vertices of the polyhedron describing power flows feasible under the so-called Linearized Load Flow model.

In the context of Benders decomposition, we present a unifying perspective on cut generation that encompasses many of the most common criteria for cut selection. These include cuts based on minimal infeasible subsystems, facet cuts, and pareto-optimal cuts, and we develop a cut selection framework that unites these different criteria. We apply our framework to an archetypical problem from the context of power system analysis and discuss some special questions that arise from this application.

Finally, we perform a game-theoretical analysis of an interdependent scheduling problem, as it may arise in the context of energy infrastructure investment decisions. We discuss welfare-maximizing solutions, as well as Nash equilibria and questions of price of anarchy and price of stability. Besides existence results, we put a special focus on algorithmic questions about the complexity of finding solutions with the stated properties.

Zusammenfassung

In dieser Arbeit untersuchen wir mathematische Strukturen von Optimierungsproblemen und -algorithmen im Kontext der Analyse und des Designs elektrischer Stromsysteme.

Zunächst stellen wir eine neue Möglichkeit vor, das Verhalten elektrischer Stromflüsse in einem einfachen Netzwerk-Fluss-basierten Optimierungsmodell abzubilden. Wir untersuchen zudem die Auswirkungen zweier verschiedener Netzmodelle in Bezug auf das resultierende Polyeder der zulässigen Lösungen. In diesem Zusammenhang liefern wir unter anderem eine kombinatorische Charakterisierung der Ecken des Polyeders, das zulässige Lösungen nach dem sogenannten *Linearized-Load-Flow*-Modell beschreibt.

Bezüglich des Dekompositionsverfahrens *Benders Decomposition* präsentieren wir eine neue Perspektive auf das Problem der Auswahl von Schnittebenen, die einige der meistverwendeten Auswahlkriterien der Literatur vereint. Hierbei gehen wir insbesondere auf Schnitte auf der Basis minimaler Unzulässigkeitssysteme, auf Facettenschnitte und auf Pareto-optimale Schnitte ein. Wir entwickeln ein Auswahlverfahren, innerhalb dessen jedes dieser unterschiedlichen Kriterien abgebildet werden kann, wenden dieses auf ein typisches Problem der Energiesystemanalyse an und diskutieren einige anwendungsspezifische Fragen.

Schließlich nehmen wir eine spieltheoretische Analyse eines *Scheduling*-Problems mit Abhängigkeiten zwischen Aufträgen vor, wie es etwa im Kontext von Infrastrukturinvestitionen im Energiesektor auftritt. Wir diskutieren hierbei wohlfahrtsmaximierende Lösungen, ebenso wie Nash-Gleichgewichte und Fragen nach dem Preis der Anarchie und dem Preis der Stabilität. Neben Existenzresultaten legen wir einen besonderen Fokus auf algorithmische Komplexitätsfragen bezüglich der Suche nach Lösungen mit den entsprechenden Eigenschaften.

Contents

Chapter 1

Introduction

1.1 Economic Background

In the analysis and design of power systems, particularly in Germany and Europe, one of the predominant topics in recent years has been the integration of a higher share of renewable energy sources. This transition, known in Germany under the term "Energiewende", presents all stakeholders in the power system with a series of new challenges. These can be collected under the terms *decentralization*, *flexibilization* and *integration*, and we will describe each of those terms briefly below.

As part of the effort to understand and to address these issues, we have undertaken two research projects together with the Chair of Renewable and Sustainable Energy Systems at TUM: *Integration of Renewable Electricity Generation*[1] took an overall view at methodological improvements in the context of energy system models. Subsequently, in *Modelling Decentralised Electricity Supply by Decomposition of Energy Systems*[2] we focussed specifically on the application of decomposition techniques.

Decentralization generally refers to the idea that a power system dominated by numerous small-scale generation units such as photovoltaic cells or wind turbines is structurally different from one in which power is mostly generated by large, centralized units, such as nuclear, coal or lignite power plants. Decentralization has two distinct effects for the analysis and design of power systems:

Firstly, geography plays a much larger role for all planning decisions. While there is an ongoing competition between the ideas of largely self-sufficient microgrids and highly interconnected supergrids, both approaches have in common that the geographic location of generation and demand needs to be taken into account. Models which represent a large area such as Germany by a single "copper plate" region and ignore the issue of power transmission between different parts of the area are no longer useful. This increases the decision space of the models (asking not only from what type of source to generate electricity, but also where) and adds an entire layer of complexity through the representation of the transmission grid.

[1] funded by the TUM International Graduate School of Science and Engineering
[2] funded by the German Federal Ministry of Economic Affairs and Energy

Secondly, to the extent to which power generation moves towards diverse and independently managed units, decision making at the macro level has to adapt, as well. It is no longer possible to prescribe a detailed schedule to each individual generation unit – both due to their large number and due to ownership and control by a diverse set of agents. Instead, individual agents determine the schedule of the generation units which they control in a way that conforms with their individual priorities. Their objectives might differ between agents from financial profit to communal well-being. These priorities have to be taken into account in planning decisions and methods have to be developed to manage them and align them with central planning objectives.

Flexibilization is the challenge that arises from the inherent unpredictability and fluctuating nature of renewable energy sources. As generation and demand have to be balanced at all times and in every region of the system, the components of the system must be able to react swiftly to any changes in the availability of renewable energy. This can mean adapting demand by the means of *demand side management*, adjusting generation in highly flexible units, using storage devices to shift generation and demand over time and/or using the transmission grid to shift generation and demand across space.

All of these need engineering solutions that provide the respective devices with the required flexibility, but in addition to this, flexibility has to be taken into account in the analysis of the power system, as well. Well-established aggregated characteristics of power systems such as the number of full load hours or sorted annual load curves have in the past provided methods to analyze large scale energy systems with acceptable computational effort. Yet these methods are unable to adequately represent many important aspects of flexibilization. Resorting to full-scale high-resolution optimization models, on the other hand, severely limits the maximal system size that can be analyzed on a given computational infrastructure.

Integration finally addresses the requirement to join the analysis of different sectors of the system that have previously been considered independently from each other. This can refer to both geographical sectors and sectors characterized by the used resources or equipment. For instance, a highly interconnected European supergrid mandates the integration of power system analysis for different European countries. Analogously, the large-scale deployment of combined heat and power devices (CHP) to increase energy efficiency requires a joint analysis of electricity, heat and natural gas systems.

As mentioned above, one effect of the need for integration is that optimization models which are employed to take all of these aspects into account grow in size tremendously, stressing the limits of available computational resources. With each added sector of the power system, the number of independent units multiplies and with it the number of related decisions that have to be made. Furthermore, each sector might bring with

it a very distinct structure of the underlying optimization problems. The established approach of adding all aspects into a single, gigantic mathematical model in this context quickly becomes impractical. Specialized algorithms might be much more capable of solving the optimization problems associated with a given sector, raising the need for a reliable method to connect different algorithms and to direct them to a global system optimum.

Furthermore, integration increases the number of agents controlling distinct parts of the infrastructure that have to be accounted for in the analysis of the energy system: Companies managing, e. g., electricity and natural gas infrastructure face different incentives, yet are increasingly depending upon each other as the degree of integration between the individual systems increases.

In forcing stakeholders to include more and more aspects of the power system into their models in order for them to remain relevant, this combination of challenges stretches the computational limits of the field. Despite ongoing progress in computer hardware and general solution algorithms for mathematical programming problems, the increased complexity forces stakeholders to make unpleasant compromises in order to keep computation times and memory demands within reasonable bounds.

In this thesis, we tackle this challenge in three ways: We suggest a new model to represent certain features of the power system with high accuracy while reducing model complexity. We furthermore propose improvements to established solution algorithms, which reduce computation times but also open an avenue to connect specialized models for different subsystems. Finally, we investigate some of the incentive structures that arise in a decentralized market and evaluate their effect on the global outcome.

1.2 Mathematical Background

From a mathematical perspective, this thesis builds upon three main strands of theory which mainly originated around the middle of the 20th century but remain active areas of research until today. They cover a wide spectrum of important theoretical developments in the area of optimization. The main theoretical building blocks are *network flow theory*, *decomposition techniques* and *algorithmic game theory*.

Network Flow Theory The theory of network flows is one of the oldest areas in mathematical optimization. A standard problem is the *maximum flow problem*: Given a network with capacities on every link, compute a flow of maximal magnitude between two distinguished nodes in the network. This basic setting has been extended in many ways, e. g. to flows between more than two nodes, circulations, and by modifications to the concept of a network, such as *generalized networks*, where losses (or gains) are incurred by flow along a link. A large portion of the field is dedicated to cost-minimizing

flows, where the use of a link (or the generation/consumption of flow at a node) is associated with some cost.

A broad overview of basic network flow theory and its applications is given by Ahuja, Magnanti, and Orlin [AMO93]. The book also mentions some applications from the area of electric circuits, see also [Chr+11] for an interesting connection between electrical flows and maximal flows in a graph. Two works from this area are particularly relevant for the results in this thesis: The first is a seminal paper by Truemper [Tru77] on the connection between cost-minimizing flows in ordinary networks and feasible flows in generalized networks. The second is a book by Rockafellar [Roc84] which covers a very extensive theory of duality between different types of network flow problems.

Decomposition Techniques Since the development of the simplex algorithm by George Dantzig, linear programming has become a standard procedure for all kinds of optimization problems. Numerous improvements (or indeed replacements) of the original algorithm have been developed, improving its performance in various settings and incorporating additional requirements for the feasibility of solutions (e. g. integer linear programming). One approach in this context are decomposition techniques, which aim to solve a large optimization problem by repeatedly solving smaller and simpler problems, while guaranteeing that the final solution is optimal for the original problem.

The most prominent representatives of this approach are known as *Lagrangian Relaxation* (see [Geo74]), *Dantzig-Wolfe-Decomposition* (see [DW61]) and *Benders Decomposition*. In this thesis, we focus on the latter method, which was introduced by Benders [Ben62]. This approach has received an enormous amount of attention recently (see, e. g., [Rah+17]) due to its usefulness in many practical settings, such as stochastic optimization, and has found wide applications in the engineering community.

Algorithmic Game Theory The origins of game theory can be traced back to Von Neumann and Morgenstern [VM44] and Nash [Nas51]. The field reached popular recognition during the Cold War as the theoretical foundation of nuclear deterrence. It can be understood as an attempt to supplement classical optimization problems with the concept of selfish behavior by individual agents. In the last twenty years, the field has received a new boost of attention with a focus on the algorithmic aspects of game theoretical problems, such as their computational complexity. A good overview over this new era of interest is given by the collection Nisan et al. [Nis+07].

In this thesis, we analyze a problem motivated from applications which builds upon the extensive theory of scheduling problems (see, e. g., [LRB77]) and extends these with elements of selfish behavior in a very natural way.

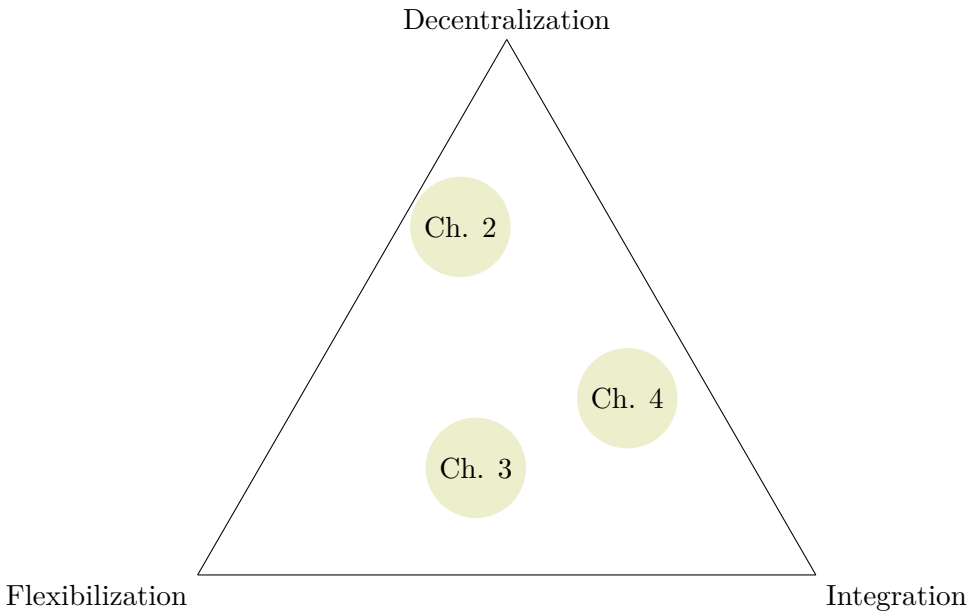


Figure 1.1: A sketch of the focus of each chapter with respect to the individual challenges laid out in Section 1.1.

1.3 Contributions

This thesis is divided into three chapters broadly along the lines of the blocks of mathematical theory presented above. While there obviously exist connections between all three areas, we have tried to keep each chapter as self-contained as possible. Some theoretical foundations which are required by all chapters can be found in the Appendix. At the same time, each chapter addresses a subset of the challenges facing power system analysis and design as outlined in Section 1.1. A rough sketch of the degree to which each chapter contributes to the different challenges is provided in Fig. 1.1.

Within each chapter, our main contributions are the following:

Network Flow Models for Transmission Capacity Expansion

- We prove the equivalence of TR and DC model, two widely used models for the representation of electrical power flows, under certain loss functions in networks without capacity constraints and empirically evaluate the validity of the above result in practical networks *with* capacity constraints.

- We derive some approximation results regarding the feasible regions (the sets of feasible *injections*) of TR and DC model.

- We present a combinatorial interpretation of extremal solutions for the DC model and prove sufficient conditions for when the resulting structure completely characterizes all extremal solutions.
- We characterize the class of networks that allow a complete characterization of extremal solutions regardless of the network parameters and prove that if we take into account network parameters, then the problem of identifying such networks becomes computationally difficult.

Parts of these results under the first bullet point have previously been published in a preliminary version as [AS13]. Major parts of the results under the last two bullet points are currently being prepared for publication as [BS19a].

Benders Decomposition for Energy System Optimization

- We prove the exact connection between two representations of the set of all possible cut normals at a given iteration, the alternative polyhedron (commonly used in Benders Decomposition) and the reverse polar set (a concept from convex geometry and disjunctive programming).
- We refine a common selection procedure for Benders cuts and obtain a characterization in terms of the objective function used for common quality criteria from the literature such as minimal infeasible subsystems and Pareto-optimal cuts.
- In particular, we prove that if the objective is chosen from a certain linear subspace, then we generally obtain facet-defining Benders cuts.
- We present a number of practical enhancements to the Benders decomposition algorithm, most notably a new method to generate valid upper bounds from infeasible subproblems.
- We present a reference implementation of Benders decomposition for a standard problem from Energy System Analysis and empirically evaluate the performance impacts of our theoretical contributions.

Major parts of the results under the first three bullet points are currently being prepared for publication as [BS19b].

Interdependent Scheduling Games

- We present a natural extension of common scheduling problems to the case of agents controlling both tasks and machines and analyze the tractability of best-response as well as welfare-maximizing schedules for both the weighted and unweighted case.

- We prove existence and efficient computability of pure Nash equilibria in the unweighted case as well as $\mathbb{NP}$-hardness of deciding existence in the weighted case.
- We prove asymptotically tight upper and lower bounds for the Price of Anarchy and Price of Stability and analyze the effect of different formulations of the objective function in this context.

Some of the results in this chapter have previously appeared in the proceedings of the ICJAI 2016 conference as [Abe+16].

Chapter 2

Network Flow Models for Transmission Capacity Expansion

Contents

In this chapter, we analyze the properties of different mathematical models that can be used to represent electrical power flows in a network. It can broadly be divided into two parts:

In Sections 2.1 and 2.2, we develop some improvements to existing *network-flow-based* models for the problem of designing the optimal infrastructure layout for an energy

system in order to satisfy a given demand. In order to do this, we first have to go through at least some of the physical details of electrical power transmission in order to set up the necessary background to present our results. The general idea of these sections is similar to our previous work on network models published as [AS13] and we use adaptations of two figures from that work (which are both referenced explicitly). All of the mathematical results in this chapter, however, are independent from [AS13].

In Sections 2.3 to 2.5, we consider a more abstract setting that allows us to isolate the fundamental differences between two common network models. We establish a link to the existing mathematical theory of network flows and extend it to our setting.

2.1 Network Models in Energy System Optimization

Since a special focus in this chapter lies on the representation of transmission networks, we begin by discussing a few concepts and issues that play a role in this context. When we talk about transmission networks (or transmission lines), we generally have in mind networks that operate at *extra high voltage*, such as the European system of 380kV lines. These typically consist of overhead alternating current (AC) lines, although underground cables and *high-voltage direct current (HVDC)* may be included, as well.

We take a high-level view at the energy system, which means that we will typically aggregate generation and demand from certain geographical regions. Accordingly, a transmission line in this context is not necessarily a single cable or circuit, rather the set of transmission lines that connect two regions are aggregated to a single equivalent transmission line. Such a line is characterized by a number of different electrical properties, most notably for our purposes the line's complex admittance (the reciprocal of its impedance) with its two components conductance and susceptance.

Since in designing an optimal infrastructure layout, we inherently have to deal with "hypothetical" transmission lines, for the electrical parameters of which no data is available, we assume that the *effective* conductance and susceptance of a transmission line of given length is linear in the line's capacity. This assumption is reasonable if we think of a line in the power network as a connection between two regions that results from aggregating multiple physical transmission lines. If the technical equipment (e. g. the kind of conductor system used etc.) is identical for these lines, then the capacity of the connection is determined by the number of such systems that are used in parallel to connect the two regions. The electrical parameters of the aggregate transmission line are then approximately proportional to the line's capacity.

Note, however, that the constant of variation between admittance and capacity depends not only on the technical equipment used, but also on the line's length. Two aspects play a role here: First, the admittance of a transmission system of given type changes with the length of the transmission line. Both susceptance and conductance are approximately inversely proportional to the length of the line. Furthermore, the

maximal load of a given transmission system (its capacity) changes depending on the length of the connection, as well. This relation is captured by the so-called *St. Clair curves* [StC53; GMD79], which specify a line's *loadability factor* as a function of the line's length. The loadability factor represents the capacity of a transmission line of the given length as a multiple of the transmission system's reference capacity, the *surge impedance loading (SIL)*. In other words, the *actual* capacity of a line is the product of its SIL and the loadability factor corresponding to the line's length.

Overall, we can write a line's effective conductance g^{eff} and susceptance b^{eff} as a function of its actual capacity in the following way:

$$g^{\text{eff}} = g^{\text{base}} \cdot \frac{1}{\text{length}} \cdot \text{SIL} = g^{\text{base}} \cdot \frac{1}{\text{length}} \cdot \frac{\text{capacity}}{\text{pfactor}(\text{length})} \tag{2.1}$$

$$b^{\text{eff}} = b^{\text{base}} \cdot \frac{1}{\text{length}} \cdot \text{SIL} = b^{\text{base}} \cdot \frac{1}{\text{length}} \cdot \frac{\text{capacity}}{\text{pfactor}(\text{length})} \tag{2.2}$$

Here, pfactor denotes the line's *loadability factor* as specified by the St. Clair curves and g^{base} and b^{base} are the base conductance and susceptance per unit of SIL for a line of length 1 km as determined exclusively by the type of technical equipment used.

To simplify the notation, we will aggregate all the factors that do not depend on a line's capacity into a single parameter and define a line's susceptance and conductance per unit of capacity by

$$g := \frac{g^{\text{base}}}{\text{length} \cdot \text{pfactor}(\text{length})} \tag{2.3}$$

$$b := \frac{b^{\text{base}}}{\text{length} \cdot \text{pfactor}(\text{length})}. \tag{2.4}$$

Within a certain subnetwork (e. g. the German 380kV transmission grid), it is furthermore reasonable to assume that the technical equipment used does not differ too much between different transmission lines and that the electrical properties of different lines with the same capacity hence depend only on the line's physical length. For instance, Egerer et al. [Ege+14] assume that in the European 380kV transmission grid, every line's susceptance and conductance per unit length is the same.

We will sometimes make the following somewhat weaker assumption about similarity of technical equipment across the network: We define a line's *material parameter* as the quotient of is effective conductance and susceptance. Note that by (2.1) to (2.4),

$$\frac{g^{\text{eff}}}{b^{\text{eff}}} = \frac{g}{b} = \frac{g^{\text{base}}}{b^{\text{base}}},$$

the material parameter is thus independent of the line's length and capacity. We call a power network *uniform*, if the material parameter is the same for all lines in the network. A power network is formally defined as follows:

Definition 2.1 (Power Network)
Let $(\mathcal{R}, \mathcal{L})$ be a directed graph with vertex set $\mathcal{R}$ and edge set $\mathcal{L} \subset \mathcal{R} \times \mathcal{R}$. Let $g, b \in \mathbb{R}^{\mathcal{L}}$ be two edge weights representing the specific electrical properties (susceptance and conductance per unit of capacity) of each individual line. We call $(\mathcal{R}, \mathcal{L}, g, b)$ a *power network* and refer to the vertices in $\mathcal{R}$ as *regions* and to the edges in $\mathcal{L}$ as *lines*.

We call a power network *uniform* if there exists a constant μ such that $g_l/b_l = \mu$ for all lines $l \in \mathcal{L}$.

Note that by $g, b \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$, we mean that g and b are vectors of dimension $|\mathcal{L}|$, the components of which are indexed directly by the elements of the set $\mathcal{L}$ (see Appendix A.1).

Furthermore, note that the edge orientations in a power network are completely arbitrary and will have no effect on the feasibility of a power flow in the network. They only serve as a reference to fix the direction of flow associated with a positive/negative flow value on the respective line.

Finally, the set $\mathcal{L}$ of lines might include both existing and *potential* lines. We will later assign to each line a capacity variable that can be set to zero if a potential line should not be built (see, e. g., Problem 2.7).

A typical set of parameters for the German 380kV transmission network can be found in [KNK11, Table 3.2]: For a typical 380kV line (562-AL1/49-STIA) of length 100 km, we obtain

$$g_l \approx 2.54 \frac{\mu S}{MW} \tag{2.5}$$

$$b_l \approx -28.91 \frac{\mu S}{MW}. \tag{2.6}$$

Similar values are obtained by Hewes et al. [Hew+16] as capacity-weighted averages for the European power grid. For a line of length 100 km, their results imply (assuming a loadability factor of 2.5) that

$$g_l \approx 2.07 \frac{\mu S}{MW} \tag{2.7}$$

$$b_l \approx -23.22 \frac{\mu S}{MW}. \tag{2.8}$$

In general, we can always assume that $g > 0$ and in a network of overhead transmission lines that $b < 0$. Note that our definition of a power network so far is independent of the transmission capacities, which are actually installed. A model of any actual power network thus requires in addition a capacity vector f^+ to be specified. Now, an energy system can be written as a power network together with a demand vector and a vector of generation cost functions:

Definition 2.2 (Energy System)
Let $(\mathcal{R}, \mathcal{L}, g, b)$ be a power network and $D \in \mathbb{R}^{\mathcal{R}}_{\geq 0}$. For every $i \in \mathcal{R}$, let $c_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ be a closed convex function. We call $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ an *energy system* and refer to D as the *demand vector* and to c_i as the *production cost function* in region i. This function maps the total power p_i produced in region i to the associated production cost $c_i(p_i)$ for sourcing this power *locally* in the cheapest possible way.

We call an energy system *non-trivial* if $D \neq 0$ and for every region $i \in \mathcal{R}$ it holds that $c_i(p_i) > c_i(0)$ for all $p_i > 0$, i. e., there is some demand for power in the system and no region can produce power at no cost.

Note that we allow $c_i(p_i) = \infty$ for certain values of p_i, which can be interpreted as "it is impossible to produce the amount p_i of power in region i" (e. g. due to limited production capacities).

Given a power network, the specific susceptance and conductance of its transmission lines together with a vector of installed capacities f^+ determine the set of electrical flows that are admissible in the network. This set of feasible electrical flows in a power network can be described in several ways and to different levels of detail, depending on which mathematical *grid model* is being used.

The arguably simplest such model, the Transport model, is based on the classical concept of network flows, where flows in a network are constrained only by edge capacities. The Transport model is widely applied because of its simplicity and exceptional computational performance. On the other hand, it often fails to represent important aspects of electrical power flows and thus severely lacks in accuracy. As a consequence, more elaborate network models are increasingly being used. Since these are computationally more demanding, their use often greatly reduces the scope of optimization models that can be solved within a reasonable amount of time.

In this chapter, we will discuss possible improvements to the accuracy of the Transport model while aiming to maintain its theoretical simplicity and practical performance. We will introduce the most commonly used grid models in Section 2.1.1 below and we provide a more detailed discussion of their physical interpretation and the relations between them in Section 2.1.2.

2.1.1 Transmission Capacity Expansion Problems

Before we proceed with the definition of the concepts of feasibility, as well as the optimization problems that this chapter is concerned with, we introduce *loss functions* which are used to model the transmission losses incurred by transporting a certain amount of power across a particular transmission line. We will see some common examples of loss functions in Definition 2.10.

Definition 2.3 (loss functions)
Let $S := \{(b_l, g_l, f_l^+, f_l) \in \mathbb{R}^2 \times \mathbb{R}^2_{\geq 0} \,|\, f_l \leq f_l^+\}$ denote the set of possible line parameters and flows satisfying the line's capacity. We call a function $\eta : S \to \mathbb{R}$ defined on that set *loss function*. For a vector $x_l = (b_l, g_l, f_l^+, f_l)$ of a transmission line's specific susceptance and conductance, its capacity and the amount of power transmitted, the value $\eta(x_l)$ represents the power lost on a line with these particular parameters when the line receives a power injection of magnitude f_l.

Given a loss function η and a power network $(\mathcal{R}, \mathcal{L}, g, b)$, we define the line-specific loss functions $\eta_l : \mathbb{R}^2 \to \mathbb{R}$ of a line $l \in \mathcal{L}$ with specific susceptance b_l and conductance g_l by

$$\eta_l(f_l^+, f_l) := \eta(b_l, g_l, f_l^+, f_l).$$

For a power network $(\mathcal{R}, \mathcal{L}, g, b)$ and a transmission line $l \in \mathcal{L}$ with capacity f_l^+ and power flow f_l, the value $\eta_l(f_l^+, |f_l|)$ is the power lost on this line when it transmits the amount $|f_l|$ between its two endpoints (the losses are independent of the direction of flow). Note that, while in reality losses occur along the entire length of a line and the amount of power transmitted across a particular segment of the line thus changes over the course of the line, it makes sense from a modeling point of view to deduct the total amount of losses at one or both of the two endpoints of a line.

In principle, we could use any arbitrary distribution of losses between the two endpoints. However, the amount of losses that is deducted at the beginning of the line no longer registers as flow across the line. On the other hand, we define losses as a function of the flow across the line. Depending on the exact distribution of losses between the two endpoints, a different loss function thus has to be used to obtain the same level of losses.

For reasons of symmetry, we have decided to use the following distribution: One half of the incurred losses is subtracted directly at the outgoing endpoint (thereby reducing the load on the line), the other half first has to be transmitted to the other endpoint and is deducted there.

For the sake of completeness, the most natural alternative to dividing the loss between both ends of each line would be to assign them entirely to either the outgoing or the incoming endpoint. Indeed, it turns out that in both cases, all the results from this chapter can be translated (replacing $\eta^{\log}$ by another suitable loss function where required). Our choice for the model presented above is based on the cleaner notation that it admits, due to the fact that we do not need to distinguish flows in different directions.

We can now formally define the *Transmission Capacity Expansion Problem*, that we will focus on in this chapter. We start by specifying the conditions for a production vector and a flow to be a feasible solution in our setting. Recall that we denote the set of incoming and outgoing edges in a vertex i by $\delta^{\text{in}}(i)$ and $\delta^{\text{out}}(i)$, respectively (see Definition A.1).

Definition 2.4 (Demand-satisfying Flow)
Let $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ be an energy system and η a loss function. A pair (p, f) of a production vector $p \in \mathbb{R}^{\mathcal{R}}_{\geq 0}$ and a flow $f \in \mathbb{R}^{\mathcal{L}}$ is called *demand-satisfying* with respect to the loss function η and a capacity vector $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$ if

$$p_i + \sum_{l \in \delta^{\text{in}}(i)} \left(f_l - \frac{\eta_l(f_l^+, |f_l|)}{2} \right) - \sum_{l \in \delta^{\text{out}}(i)} \left(f_l + \frac{\eta_l(f_l^+, |f_l|)}{2} \right) \geq D_i \qquad \forall\, i \in \mathcal{R}. \quad (2.9)$$

The *cost* of (p, f) is given by

$$c(p, f) := \sum_{i \in \mathcal{R}} c_i(p_i).$$

If there exists a pair (p^*, f^*) with $c(p^*, f^*) < \infty$ that strictly satisfies all the inequalities (2.9), then we call $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ *strictly feasible for* η.

Whenever the loss function is clear from the context, we omit it and say simply that (p, f) is demand-satisfying for f^+. While a demand-satisfying solution provides every region with enough electricity, it does not necessarily respect the physical constraints imposed by the transmission grid (e. g. line capacities). In addition to demand-satisfaction, we hence require the following properties for a pair (p, f) to be feasible. As these depend on the network model used, we differentiate two degrees of feasibility (see Fig. 2.1 for an example of the differences):

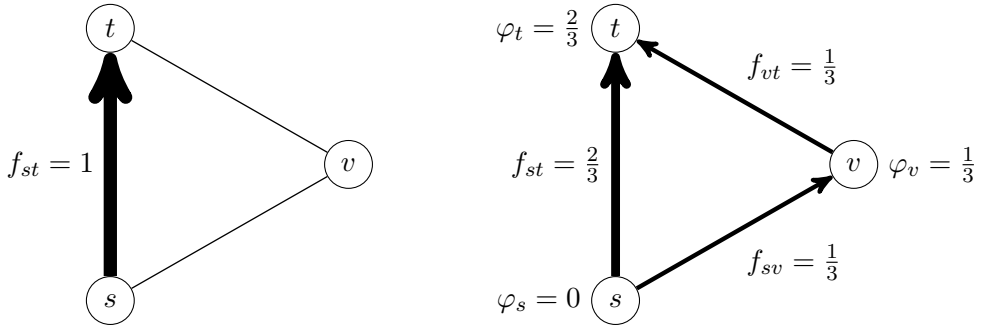
Definition 2.5
Let an energy system $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ and a loss function η be given. A pair (p, f) is *TR-feasible* for a capacity vector $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$, if it is demand-satisfying w.r.t. η and f^+ and satisfies

$$\forall\, l \in \mathcal{L} : |f_l| \leq f_l^+. \quad (2.10)$$

TR-feasibility thus requires only that capacity constraints are respected. DC-feasibility adds to this the requirement that the flow on each line is proportional to the difference in *phase angles* φ_i between its endpoints with respect to a suitably chosen vector φ.

Definition 2.6
Let an energy system $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ and a loss function η be given. A pair (p, f) is *DC-feasible* for a capacity vector $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$, if it is TR-feasible for f^+ and there exists a vector $\varphi \in \mathbb{R}^{\mathcal{R}}$ that satisfies

$$\forall\, (i, j) \in \mathcal{L} : f_{ij} = b_{ij} f_{ij}^+ \cdot (\varphi_j - \varphi_i). \quad (2.11)$$



(a) A flow that is TR-feasible, but not DC-feasible.

(b) A flow that is both TR- and DC-feasible.

Figure 2.1: A small example to illustrate the difference between TR- and DC-feasibility: Three nodes with demand values $D_t = 1$ and $D_s = D_v = 0$ are connected by identical (lossless) lines with $b_l = 1$ and capacities $f_l^+ = 1$ for all lines l. Both flow realizations shown are TR-feasible (together with the production vector $p_s = 1, p_v = p_t = 0$), since they respect the capacity constraints. However, only the flow in (b) is also DC-feasible (as certified by the vector φ which is also shown). In (a), there is no assignment of values φ_i to the nodes that satisfies the equation (2.19) for each pair of vertices with respect to the flow f.

The equations (2.11) ensure that power flows in the network behave in a way similar to current flows in a *direct current (DC)* circuit. TR-feasibility, on the other hand, ignores any interdependence between flows on different lines and only requires that capacity limits are respected on each individual line. While Definitions 2.5 and 2.6 are sufficient for our work in this chapter from a mathematical perspective, we provide a brief interpretation of the respective requirements in the context of alternating current power transmission networks in the following Section 2.1.2.

Using the above terminology, we can now formulate the following two optimization problems, the Transmission Capacity Expansion Problem (TCEP) and the Optimal Power Flow Problem (OPF).

Problem 2.7 (Transmission Capacity Expansion Problem (TCEP))
Let an energy system $(\mathcal{R}, \mathcal{L}, g, b, D, c)$, a loss function η and a capacity cost vector $c^f \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$ be given. Determine a capacity vector $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$ and a pair (p, f) DC-feasible/TR-feasible for f^+ such that

$$c^f \cdot f^+ + c(p, f)$$

is minimal. Depending on the network model used to determine feasibility, we use the abbreviation DC-TCEP and TR-TCEP, respectively.

It is easy to see that, since DC-feasibility implies TR-feasibility, the optimal objective value for the problem TR-TCEP is a lower bound for the objective value of DC-TCEP.

We may also consider the following corresponding problem that asks for an optimal flow (p, f) for a given, fixed capacity vector f^+:

Problem 2.8 (Optimal Power Flow Problem (OPF))
Let an energy system $(\mathcal{R}, \mathcal{L}, g, b, D, c)$, a loss function η and a capacity vector $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$ be given. Determine a pair (p, f) DC-feasible/TR-feasible for f^+ such that $c(p, f)$ is minimal. As above, we use the abbreviation DC-OPF and TR-OPF, respectively, depending on the network model used to determine feasibility.

Note that both problems as defined above are abstracted versions of problems as they arise in an engineering context, which typically include many more details, e. g. with respect to the constraints for power generation inside an individual region. These versions are sufficient for our work in this chapter, since they include all the relevant aspects of the problems while keeping the notational overhead minimal.

One particular aspect that we ignore in Problem 2.7 is that we would typically be interested in a cost-minimizing capacity vector *over time*. With a suitable discretization of the time horizon under consideration into time steps $t \in T$ and demand vectors D_t as well as production cost functions c_t for each time step t, the problem then consists of finding a capacity vector f^+ and pairs (p^t, f^t) for every time step t, that are all DC-/TR-feasible for f^+ and minimize

$$c^f \cdot f^+ + \sum_{t \in T} c_t(p^t, f^t). \tag{2.12}$$

Furthermore, it might be desirable to include some security margin in the optimal capacities obtained from the TCEP. One could, e. g., require that only 80 % of the installed capacity on each line should ever be used. This can easily be incorporated into Problem 2.7 by using for f^+ the capacities that should actually be respected by the power flow, e. g. 80 % of the installed capacity for the example above (and adapting the cost vector etc. accordingly).

2.1.2 Network Models

In this section, we give a brief overview over the archetypical network models that are used most commonly in the context of Energy System Optimization. The remainder of this chapter will only be based on the definitions from Section 2.1.1 above, in this sense the contents of this section are entirely optional. However, they provide some context, in particular to put the definitions of TR- and DC-feasibility in relation. Furthermore, we would like to provide a very brief introduction into the description of alternating current power networks and link the models that we use in this chapter to the physical equations that a reader might be familiar with from other contexts.

We focus on three particular models for the description of power networks: the Transport model and the DC model, which are most widely used in the field of

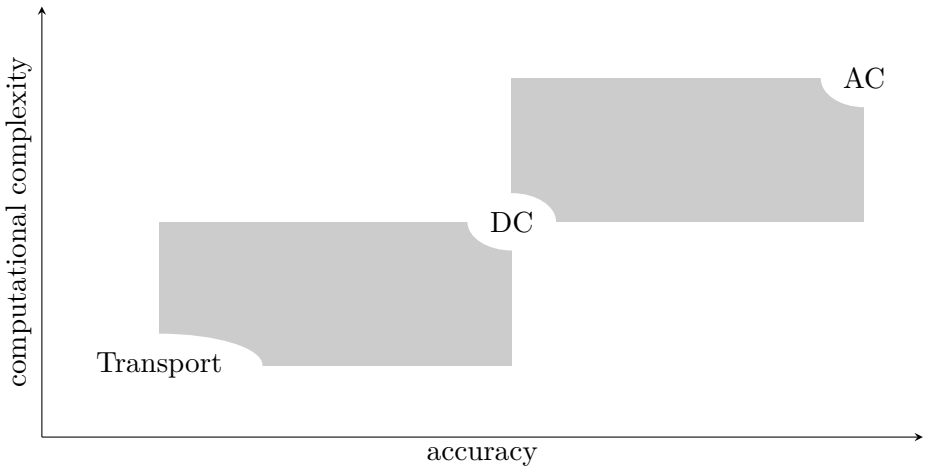


Figure 2.2: Qualitative sketch of the properties of different lossless grid models: The Transport model is rather crude, but computationally very efficient, the AC model represents an adequately accurate representation of power flows in a grid. The DC model strikes a middle ground on both dimensions. Different refinements and simplifications of these models can be developed, as indicated by the grey rectangles.

investment planning and on which our definitions of TR-/DC-feasibility (Definitions 2.5 and 2.6) are based, as well as the AC model, which can be seen as the most *physically correct* and is generally considered a suitable reference to compare the accuracy of other models against (see, e. g., [Rom+02; SJA09]).

The three models can be distinguished from one another by specific advantages and disadvantages with respect to computational effort and accuracy that are qualitatively depicted in Figure 2.2. Naturally, computational performance is not only a function of the type of network model. Instead it varies depending on the application and on the specific variant of the respective model that is being used. For instance, all models can be enhanced by different representations of transmission losses, increasing their accuracy and (potentially) their computational complexity. However, these effects are not the same in all models. One result from this chapter will be that an appropriate choice for the representation of transmission losses can in particular greatly increase the accuracy of the Transport model, moving it much closer to the DC model without significantly increasing the computational complexity.

We will revisit these computational complexity issues in more detail in Section 2.1.1. For now we shall focus on their different physical interpretations. For the following presentation of the different network models, note that the AC model shall serve as a reference point only, hence it will be covered only to the degree necessary to distinguish it from the other models and to understand their relation. The exposition of the AC

model is inspired by [SJA09]. The DC model and Transport model are also covered in [Rom+02], a study of network models used in network expansion planning.

In accordance with the literature (e. g., [SJA09]) we will throughout the remainder of this chapter represent all parameters and values in the *per-unit* system with base power value 1 MVA and base voltage magnitude equal to that of the grid under consideration (e. g. 380 kV for the German extra high voltage transmission grid). This allows for a cleaner notation since we can include certain constants (such as the base voltage magnitude) into the parameters and do not have to carry them around explicitly. For example, the values from (2.5) and (2.6) can be rewritten on this basis as follows:

$$g_l \approx 2.54 \cdot 10^{-6} \frac{S}{MW} = 2.54 \cdot 10^{-6} \cdot \frac{380000^2}{10^6} \frac{p.u.}{MW} = 0.366776 \frac{p.u.}{MW} \quad (2.13)$$

$$b_l \approx -28.91 \cdot 10^{-6} \frac{S}{MW} = -28.91 \cdot 10^{-6} \cdot \frac{380000^2}{10^6} \frac{p.u.}{MW} = -4.174604 \frac{p.u.}{MW} \quad (2.14)$$

AC model. Long-distance power transmission networks usually work using *high-voltage alternating current*. For this, the entire network is oscillating at a defined frequency (typically, in the EU, 50 Hz), similar to a *resonant circuit* (or *LC circuit*). This means that both voltage and current at any point in the network oscillate in the form of a sine wave. Ohmic resistance in devices connected to the network (as well as the network itself) reduces the amplitude of the oscillation over time, as electrical energy is converted into heat. Such devices are known as *active loads* and they are said to consume *active power*.

For a given amplitude of the oscillation, active power, as the product of voltage and current, is highest if both oscillate perfectly in sync. If, on the other hand, they oscillate with a phase shift of 90°, the same amount of power that is delivered by the network at one point in time has to be fed back into the network over the course of one cycle of the oscillation. This means that, in total, no active power can be consumed.

Such a phase shift can result from certain components in the connected devices (and, again, the network itself) behaving to some extent as electrical capacitors or inductive coils. This is true even for a simple overhead line: As an electrical current is sent through the conductor, both a magnetic and an electric field form around it. In the case of alternating current, both collapse and are then rebuilt as the direction of current reverses. This behavior, known as *reactive load*, does not decrease the amplitude of the oscillation, instead it leads to a phase shift between the oscillations of voltage and current, respectively. This, as mentioned above, reduces the amount of active power that is available in the network (the power is not *consumed*, or converted into heat, but becomes inaccessible since it is stored in the respective electric and magnetic fields).

The AC model represents a reasonably accurate model of the behavior described above: Every line (i, j) is characterized by its capacity f_{ij}^+ and the two parameters g_{ij} and b_{ij}, corresponding to the inverse of its active and reactive resistance which

quantify the extent to which the line itself behaves as an active load and/or a reactive load when it is used to transport active power between its endpoints.

If these parameters are known, then the power flow across a transmission line can be described by the voltage magnitude (the amplitude of the oscillation) at the beginning (v_i) and the end (v_j) of the line, as well as the line angle $\Delta\varphi_{ij} = \varphi_j - \varphi_i$, where φ_i is the voltage phase shift at network node i compared to some reference node.

The (active) power injections at the beginning (p_i) and the end (p_j) of the line are then given by the following set of equations (for a detailed explanation, see, e. g., [Kni72; WW84]):

$$\begin{aligned} p_i &= -v_i v_j f_{ij}^+ b_{ij} \sin\Delta\varphi_{ij} + f_{ij}^+ v_i g_{ij}(v_i - v_j \cos\Delta\varphi_{ij}) \\ p_j &= +v_i v_j f_{ij}^+ b_{ij} \sin\Delta\varphi_{ij} + f_{ij}^+ v_j g_{ij}(v_j - v_i \cos\Delta\varphi_{ij}) \end{aligned} \tag{2.15}$$

Note that if several lines meet in one node, the same voltage and phase shift variables appear in the equations associated with each of those lines, thus linking the power flows on different lines.[1]

The first summand on the right hand side of the above equations can be understood as representing the amount of active power that is transmitted, while the second term represents losses that occur across the line. We can thus reformulate the above equations to obtain for each line (i, j) a real power flow from i to j of magnitude

$$f_{ij} := v_i v_j f_{ij}^+ b_{ij} \sin\Delta\varphi_{ij}. \tag{2.16}$$

If f_{ij} is negative, it corresponds to a real power flow from j to i. These equations constitute the lossless AC model mentioned in Fig. 2.2. Similarly, the transmission losses between vertices i and j are captured in the term

$$f_{ij}^+ g_{ij} \cdot (v_i^2 + v_j^2 - 2 v_i v_j \cos\Delta\varphi_{ij}). \tag{2.17}$$

A similar set of equations holds for the flow of reactive power.

The AC model is used in different contexts, one common example being security-constrained unit commitment (see, e. g., [FSL05]), but also network contingency analysis. However, since the equations (2.15) represent a system of non-convex nonlinear equations, a solution can in most cases only be approximated numerically. This restriction makes the model methodologically hard for global optimization and furthermore computationally unsuitable in many cases.[2]

[1] This also motivates the traditionally more common *bus-injection model*, where power flows along transmission lines are given only implicitly. We use the above representation because it is more in line with mathematical network flow problems, see, e. g., [Low14a] on the equivalence of both models.

[2] Lavaei and Low have recently sparked some effort to come up with and analyze convex relaxations, at least for certain network topologies [LL12; LTZ14; MSL15].

Transport model. Faced with the conflict between physical accuracy and computational complexity sketched in Fig. 2.2, the Transport model (on which our definition of TR-feasibility (Definition 2.5) is based) strikes the opposite tradeoff to that represented by the AC model: The only restriction that the Transport model imposes on network flows are (thermal) line capacities. These are represented by an upper bound on the absolute value of the flow on the line: For every line (i, j) the flow f_{ij} has to respect $|f_{ij}| \leq f_{ij}^+$ where f_{ij}^+ represents the thermal capacity of the line (i, j). A power flow according to the Transport model can thus be understood as a simple network flow that is bound merely by the capacity of edges in a graph.

Since, major parts of the underlying physical laws are ignored, power flows feasible under this model can deviate quite substantially from physically "correct" flows. In exchange for its lower accuracy, the Transport model has a number of advantages from a computational point of view which we will discuss in more detail in Section 2.2. These are the primary reason why the Transport model is still widely being used, especially in contexts that either involve very large instances (large number of timesteps and many regions/units) or require a basic problem to be solved a significant number of times (e. g. in the context of a branch-and-bound scheme for integer unit commitment problems), see,e. g., [Tuo+09; Fri12; SSH12].

DC model (Linearized Load Flow). As indicated in Figure 2.2, the DC model (on which our definition of DC-feasibility (Definition 2.6) is based) can be seen as a compromise between AC model and Transport model with respect to the tradeoff between accuracy and computational complexity. Most authors present it as a simplification of the AC model (see, e. g., [SJA09], [WW84, Ch. 4.1.3]): It is assumed that the voltage magnitude remains unchanged across the network and thus all v_i can be normalized to 1. Furthermore, the function $\sin \Delta\varphi$ is approximated by $\Delta\varphi$, as the occurring angles are typically rather small.

If we furthermore ignore transmission losses, then this reduces (2.15) to

$$\begin{aligned} p_i &= -f_{ij}^+ b_{ij}(\varphi_j - \varphi_i) \\ p_j &= +f_{ij}^+ b_{ij}(\varphi_j - \varphi_i) \end{aligned} \tag{2.18}$$

or, analogously to (2.16),

$$f_{ij} := f_{ij}^+ v_i v_j b_{ij} \cdot (\varphi_j - \varphi_i). \tag{2.19}$$

Instead of viewing the DC model as a simplified version of the AC model, one may also interpret it as enhancing the Transport model by adding exactly those constraints that make sure that a flow respects an adaptation of Kirchhoff's voltage law for direct current (DC) circuits (see [HK86, Ch. 2-3]).

In terms of our definitions from Section 2.1.1, this is captured by the following proposition, for which we introduce some shorthand notations: If convenient, we

can interpret a line $(i,j) \in \mathcal{L}$ as its reverse line. This does not change the line's parameters, but reverses the sign of the flow. We therefore write $b_{ji} := b_{ij}$, $f^+_{ji} := f^+_{ij}$ and $f_{ji} := -f_{ij}$.

Proposition 2.9
A solution (p,f) is DC-feasible if and only if it is TR-feasible and for every cycle C in the power network visiting the vertices $r_0, r_1, \dots, r_k = r_0$,

$$\sum_{i=0}^{k-1} \frac{f_{r_i r_{i+1}}}{f^+_{r_i r_{i+1}} b_{r_i r_{i+1}}} = 0. \tag{2.20}$$

Proof. Let $\varphi \in \mathbb{R}^{\mathcal{R}}$ be a vector for which f satisfies (2.11). Let C be a cycle visiting the vertices $r_0, r_1, \dots, r_k = r_0$. Then,

$$\sum_{i=0}^{k-1} \frac{f_{r_i r_{i+1}}}{f^+_{r_i r_{i+1}} b_{r_i r_{i+1}}} = \sum_{i=0}^{k-1} \varphi_{r_{i+1}} - \varphi_{r_i} = 0.$$

Conversely, assume that (2.20) holds for all cycles. We construct a vector $\varphi \in \mathbb{R}^{\mathcal{R}}$ that satisfies (2.11) as follows: For each connected component of the power network, choose an arbitrary vertex $r_0 \in \mathcal{R}$ and set φ_{r_0} to some arbitrary value. For every vertex $r \in \mathcal{R}$ that belongs to the same connected component, choose an undirected path that connects r_0 with r and denote the vertices encountered along this path by $r_0, r_1, r_2, \dots, r_k = r$. Now, set

$$\varphi_r := \varphi_{r_0} + \sum_{i=0}^{k-1} \frac{f_{r_i r_{i+1}}}{f^+_{r_i r_{i+1}} b_{r_i r_{i+1}}}.$$

Now, let $(r, r') \in \mathcal{L}$ and consider the paths connecting r and r' to r_0 as defined above. Together with the edge (r, r'), these paths form a (not necessarily simple) cycle. By (2.20), we obtain that $\frac{f_{rr'}}{f^+_{rr'} b_{rr'}} = \varphi_{r'} - \varphi_r$, which implies (2.11). □

The above proposition can also be verified in the example from Fig. 2.1: The flow shown on the right satisfies (2.20) with respect to the unique cycle in the network, which is not true for the flow shown on the left.

Proposition 2.9 thus implies an alternative characterization of DC-feasibility: Instead of (2.11), we could also enforce (2.20) for all cycles in the power network. In fact, it would be sufficient to enforce (2.20) for a subset of cycles that form a *cycle basis* of the underlying network [Kir47]. These bases are always of size linear in the number of edges and a basis with a minimum total number of edges can be computed in polynomial time (e. g., [BGdV04; Kav+04]).

In view of the three models presented above, each with its particular tradeoff between precision and computation time, a natural question is whether other tradeoffs can be found: How much precision can be achieved without sacrificing the computational advantages of a simpler model? The grey areas in Figure 2.2 indicate such regions of interest.

This question has been addressed to some extent by Alguacil, Motto, and Conejo [AMC03], who present an improved representation of transmission losses in the context of the DC model. Also, Stott, Jardim, and Alsaç [SJA09] analyze the different results obtained from DC model and AC model. The authors point out conditions under which the assumptions underlying the DC model are unjustified and lead to misleading results. Finally, a recent study by Coffrin and Van Hentenryck [CV14] has attempted with some success to include relaxed versions of some constraints from the AC model into the framework of the DC model. This approach leads to a much better accuracy of the resulting flows while maintaining a large portion of the computational advantage of the DC model over the AC model.

In the area between Transport model and DC model, on the other hand, the only attempt that we are currently aware of to increase accuracy beyond the Transport model, while keeping complexity below the DC model, is the *Hybrid Model* (see, e. g., [VGS85]) where existing lines are represented using the DC model while expansion options use the Transport model.

In this chapter, we will more carefully investigate the differences between Transport model and DC model and present an approach that substantially improves the representation of power flows under the Transport model without sacrificing its computational advantages. In contrast to the Hybrid Model, our approach will be applicable to a greenfield setting without any existing infrastructure, as well. Furthermore, it can at least to some extend be combined with the Hybrid Model approach: A Hybrid Model can be thought of as a Transport model network with some additional DC lines (those that already exist). This means that all of our results with respect to individual transmission lines can be used for the underlying Transport Model network of a Hybrid model, as well. This covers in particular our work on loss functions (Section 2.1.3) and their piecewise-linear approximation (Section 2.2.1).

2.1.3 Loss Functions

One important aspect to note is that the AC power flow equations (2.15) include a term that corresponds to the transmission losses incurred, depending on the power flow on a line as well as its physical parameters. Hence, the AC model comes with a very natural model to handle transmission losses using the expression (2.17). In the DC model and Transport model on the other hand power flows are lossless by default, which gives us some choice in how to add transmission losses to the model.

Many ways to represent transmission losses can be found in the literature. The

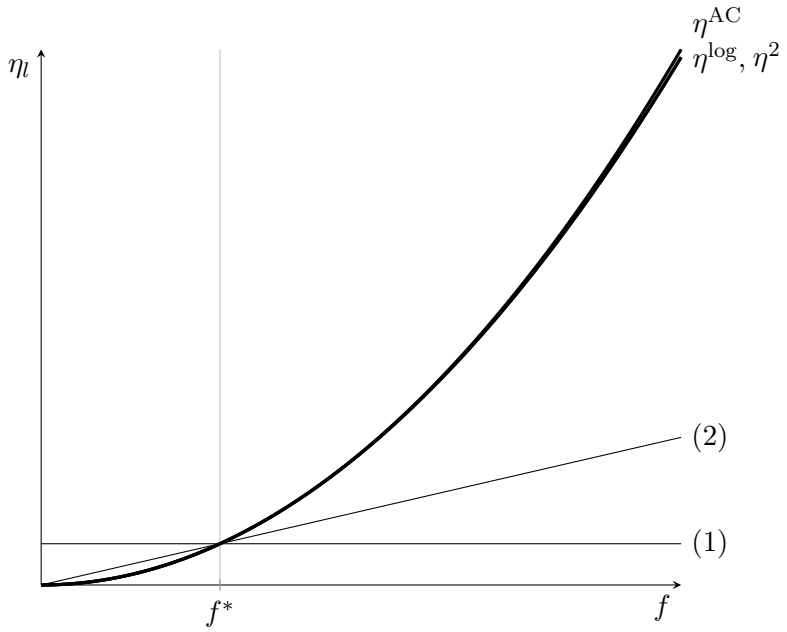


Figure 2.3: Different models of transmission losses for fixed capacity f^+ in a typical transmission line. The thick plots represent the reference point-free loss functions from Definition 2.10 (η^2 and $\eta^{\log}$ are indistinguishable in this plot). Constant and linear local approximations with respect to the reference point f^* are shown as (1) and (2), respectively. Based on a figure from [AS13].

simplest idea is to use empirical experience or expert judgement to determine some fixed amount of energy lost during transmission (globally or per line) and add the result to the demand vector (e. g., [LF92]). This approach (depicted as plot (1) in Fig. 2.3) is very simple and does not affect the complexity of the used optimization model. However, the choice of the level of losses is very prone to error and in particular assumes that the system will not deviate substantially from a reference point defined *a priori*.

This drawback may be alleviated by using some local approximation of transmission losses around a predefined operating point. The most common method is to define losses as a constant fraction of the power transmitted on a line (e. g., [Tuo+09], [SSH12]). This approach is depicted as plot (2) in Fig. 2.3. This fraction must be fixed *a priori* and is typically obtained by measuring or estimating the properties of a line in some reference point (see Fig. 2.3).

Since any choice of such reference points invariably ties the model to a certain predefined operating point, we will instead focus on transmission loss models which, in contrast to the two approaches mentioned above, are *reference point-free* (or *cold-start*

models in the terminology of [SJA09]). By this we mean that they do not rely on any previously determined reference point. In the terminology of Definition 2.3, we define the following loss functions, which are visualized in Fig. 2.3 as functions of f for a fixed line and fixed capacity f^+:

Definition 2.10 (particular loss functions)

$$\eta^{\mathrm{AC}}(b_l, g_l, f_l^+, f_l) := 2g_l f_l^+ \left(1 - \cos\left(\arcsin\left(\frac{f_l}{b_l f_l^+}\right)\right)\right) \tag{2.21}$$

$$= 2g_l f_l^+ \left(1 - \sqrt{1 - \left(\frac{f_l}{b_l f_l^+}\right)^2}\right)$$

$$\eta^2(b_l, g_l, f_l^+, f_l) := \frac{g_l}{b_l^2 f_l^+} f_l^2 \tag{2.22}$$

$$\eta^{\log}(b_l, g_l, f_l^+, f_l) := \frac{2b_l^2 f_l^+}{g_l} \left(\log\left(\exp\left(\frac{2g_l f_l}{b_l^2 f_l^+}\right) + 1\right) - \log 2\right) - 2f_l \tag{2.23}$$

$$\eta^0(b_l, g_l, f_l^+, f_l) := 0 \tag{2.24}$$

Furthermore, we define $\eta^{\mathrm{AC}}(b_l, g_l, 0, 0) = \eta^2(b_l, g_l, 0, 0) = \eta^{\log}(b_l, g_l, 0, 0) := 0$ for all $b_l, g_l \in \mathbb{R}$.

The loss function η^{AC} corresponds to the AC loss term (2.17) for voltages fixed to 1 *p.u.* and resolved to the amount of electrical power transmitted (2.16). This is the most accurate representation of transmission losses that can be derived from the models presented above. On the opposite end, the function η^0 corresponds to the most simplistic assumption of lossless transmissions. The function η^2 represents an often-used quadratic approximation from the area of direct current electrical flows (see, e. g., [HK86, Ch. 2-2]) that is also used in [AMC03] to improve the DC model.

Finally, the function $\eta^{\log}$ does not commonly appear in the literature, but we shall see below that it represents an approximation of η^{AC} very similar to η^2 (Remark 2.11), maintaining important properties of η^2 that make them easy to handle in the context of optimization problems (Theorem 2.16). In addition, it will allow us to theoretically prove the equivalence of optimal flows in the DC model and the Transport model under certain conditions (Theorem 2.14).

In the following, the loss functions (2.21) to (2.23) will be used as approximations of each other. This is justified by the following observation:

Remark 2.11
Let $(\mathcal{R}, \mathcal{L}, g, b)$ be a power network, let $l \in \mathcal{L}$ and $f_l^+ \in \mathbb{R}$ be fixed and consider the loss functions $\eta_l^{\mathrm{AC}}, \eta_l^{\log}, \eta_l^2$ from Definition 2.10. Then, for any $\alpha, \beta \in \{\mathrm{AC}, \log, 2\}$ and $f_l \in \mathbb{R}_{\geq 0}$ we have

$$\eta_l^\alpha(f_l^+, f_l) = \eta_l^\beta(f_l^+, f_l) + o(|f_l|^2).$$

2.1.4 TR-optimality and DC-feasibility

In light of the tradeoff between accuracy and computational complexity that was outlined in Section 2.1.2 (and will be made more precise in Section 2.2), it is worthwhile to compare the solutions to the different versions of the Optimal Power Flow Problem (Problem 2.8) and to investigate their differences. More specifically, we will ask how optimal solutions of DC- and TR-OPF relate to each other and under which circumstances a TR-feasible solution also satisfies the more restrictive concepts of DC-feasibility. We start by observing that the set of demand-satisfying flows (a superset of DC-/TR-feasible flows) is convex under mild conditions regarding the loss function.

Lemma 2.12
Let $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$ be a capacity vector and η a loss function that is non-decreasing and convex in the fourth argument. Then, the set of demand-satisfying flows is convex, i. e., if (p, f) and (p', f') are demand-satisfying for f^+ w.r.t. η and $\lambda \in (0, 1)$, then $\lambda(p, f) + (1 - \lambda)(p', f')$ is demand-satisfying.

Proof. As η is non-decreasing and convex in the fourth argument and $|\cdot|$ is convex, $-\eta_l(f_l^+, |f_l|)$ is concave in f_l for all $l \in \mathcal{L}$ and $f_l^+ \in \mathbb{R}_{\geq 0}$. Thus, the left hand side of (2.9), as a sum of linear and concave expressions in (p, f) is itself concave, which proves the statement. □

The following lemma provides the basis for the main theorem in this section: If we use the loss function $\eta^{\log}$, then any cost-minimal demand-satisfying flow satisfies (2.11). In order for this lemma to be true, we have to assume a uniform power network, i. e., one where the material parameter $\mu_l = g_l/b_l$ is the same for all lines. This can be thought of as a *similar material assumption* for all transmission lines in the system:

Assuming that μ_l is the same for all lines is equivalent to assuming that lines only differ in length and capacity, not in the properties of the technical components that are used. While this is certainly a simplification, we have argued in Section 2.1 that it is a reasonable assumption for models that cover only a single voltage level such as the 380kV level typically used in European transmission networks. Moreover, it will turn out in our empirical evaluation in Section 2.2.2 that our result is relatively stable with respect to small variations in the material parameter.

Lemma 2.13
Let $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ be a non-trivial energy system with $g > 0$ and $b \neq 0$ that is strictly feasible for the loss function $\eta^{\log}$ and let the underlying power network $(\mathcal{R}, \mathcal{L}, g, b)$ be uniform with material constant $\mu = \frac{g_l}{b_l}$ for all lines $l \in \mathcal{L}$. Furthermore, let $f^+ > 0$ be a capacity vector (we can delete lines $l \in \mathcal{L}$ with $f_l^+ = 0$ from the network) and let (p, f) be cost-minimizing among all flows demand-satisfying under $\eta^{\log}$. Then, there exists a vector $\varphi \in \mathbb{R}^{\mathcal{R}}$ such that (p, f) satisfies (2.11).

Proof. Let (p, f) be a cost-minimizing solution among all demand-satisfying flows. Then, (p, f) is an optimal solution of the following optimization problem:

$$\min \sum_{i \in \mathcal{R}} c_i(p_i) \tag{2.25}$$

$$\text{s.t. } p_i + \sum_{l \in \delta^{\text{in}}(i)} \left(f_l - \frac{\eta_l^{\log}\left(f_l^+, |f_l|\right)}{2} \right) - \sum_{l \in \delta^{\text{out}}(i)} \left(f_l + \frac{\eta_l^{\log}\left(f_l^+, |f_l|\right)}{2} \right) \geq D_i \quad \forall\, i \in \mathcal{R} \tag{2.26}$$

$$p \geq 0 \tag{2.27}$$

Note that $g > 0$ implies the function $\eta^{\log}$ is convex in the fourth argument, the above optimization problem is thus convex by Lemma 2.12. As $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ is strictly feasible and the solution value is non-negative (because $(c_i \geq 0)$, we can obtain a KKT vector $\sigma \in \mathbb{R}^{\mathcal{R}}_{\geq 0}$ for the above optimization problem ([Roc70, Corollary 28.2.1]). If we write $\eta_l'(f_l) := \frac{\partial}{\partial f}\eta_l^{\log}(f_l^+, |f_l|)$, then [Roc70, Theorem 28.3 (c)] yields (for some $\tau \geq 0$)

$$\begin{aligned} 0 &\in \partial c_i(p_i) - \sigma_i - \tau_i && \forall\, i \in \mathcal{R} \\ 0 &= \left(1 + \frac{\eta_l'(f_l)}{2}\right)\sigma_i - \left(1 - \frac{\eta_l'(f_l)}{2}\right)\sigma_j && \forall\, l = (i,j) \in \mathcal{L} \end{aligned}$$

which simplifies to

$$\sigma_i \leq \max \partial c_i(p_i) \qquad \forall\, i \in \mathcal{R} \tag{2.28}$$

$$0 = \left(1 + \frac{\eta_l'(f_l)}{2}\right)\sigma_i - \left(1 - \frac{\eta_l'(f_l)}{2}\right)\sigma_j \qquad \forall\, l = (i,j) \in \mathcal{L}. \tag{2.29}$$

Note that if the function c_i is differentiable, the subdifferential is unnecessary, since in that case $\max \partial c_i(p_i) = c_i'(p_i)$. From the definition of $\eta^{\log}$, we have for every $l \in \mathcal{L}$ that

$$\begin{aligned} \eta_l'(f_l) &= \frac{\partial \eta_l^{\log}(f_l^+, |f_l|)}{\partial f} = \frac{\partial \eta^{\log}(b_l, g_l, f_l^+, |f_l|)}{\partial f} \\ &= \frac{b_l^2 f_l^+}{g_l} \cdot \left(\frac{2\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) \cdot \frac{2g_l}{b_l^2 f_l^+}}{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) + 1} - \frac{2g_l|f_l|}{b_l^2 f_l^+} \right) \cdot \operatorname{sign}(f_l) \\ &= 2 \cdot \left(\frac{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) - 1}{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) + 1} \right) \cdot \operatorname{sign}(f_l) \end{aligned}$$

and thus in particular $|\eta_l'(f_l)| < 2$.

We first observe that $\sigma_i > 0$ for all $i \in \mathcal{R}$: Suppose that there exists i^* with $\sigma_{i^*} = 0$. Then, (2.29) implies for all neighbors j of i^* that $0 = (1 - \frac{1}{2}\eta'_{i^*j}(f_{i^*j}))\sigma_j$ or $0 = (1 + \frac{1}{2}\eta'_{ji^*}(f_{ji^*}))\sigma_j$, i. e., $\sigma_j = 0$. By induction, it follows that $\sigma \equiv 0$. In this case, by [Roc70, Theorem 28.4], the optimal solution is $\operatorname{argmin}_{(p,f)\geq 0} \sum_{i\in\mathcal{R}} c_i(p_i)$, which in a non-trivial energy system implies that $p \equiv 0$. Summing up all inequalities (2.26) yields $\sum_{i\in\mathcal{R}} p_i \geq \sum_{i\in\mathcal{R}} D_i + \sum_{l\in\mathcal{L}} \eta_l^{\log}(f_l^+, f_l) \geq \sum_{i\in\mathcal{R}} D_i$ and hence $D \equiv 0$, a contradiction with non-triviality of the energy system.

For every $i, j \in \mathcal{R}$ with $l = (i, j) \in \mathcal{L}$, we can now conclude that

$$\frac{\sigma_j}{\sigma_i} = \frac{1 + \frac{1}{2}\eta'_l(f_l)}{1 - \frac{1}{2}\eta'_l(f_l)}.$$

Using an idea from Truemper's comparison of min-cost flows and flows with gains [Tru77], we can take logarithms to obtain

$$\log(\sigma_j) - \log(\sigma_i) = \log\left(\frac{1 + \frac{1}{2}\eta'_l(f_l)}{1 - \frac{1}{2}\eta'_l(f_l)}\right) \tag{2.30}$$

$$= \log\left(\frac{1 + \left(\frac{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) - 1}{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) + 1}\right) \cdot \operatorname{sign}(f_l)}{1 - \left(\frac{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) - 1}{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) + 1}\right) \cdot \operatorname{sign}(f_l)}\right) \tag{2.31}$$

$$= \log\left(\frac{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) + 1 + \left(\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) - 1\right) \cdot \operatorname{sign}(f_l)}{\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) + 1 - \left(\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right) - 1\right) \cdot \operatorname{sign}(f_l)}\right) \tag{2.32}$$

$$= \log\left(\frac{2\exp\left(\frac{2g_l|f_l|}{b_l^2 f_l^+}\right)}{2}\right) \cdot \operatorname{sign}(f_l) = \frac{2g_l|f_l|}{b_l^2 f_l^+} \cdot \operatorname{sign}(f_l) \tag{2.33}$$

$$= 2\mu \frac{f_l}{b_l f_l^+}. \tag{2.34}$$

This yields that

$$\left(\frac{\log(\sigma_j)}{2\mu} - \frac{\log(\sigma_i)}{2\mu}\right) \cdot b_l f_l^+ = f_l.$$

We can conclude that for the choice of $\varphi_i := \frac{\log \sigma_i}{2\mu}$ for all $i \in \mathcal{R}$, the flow (p, f) satisfies (2.11), which proves the lemma. □

Note that while we drew inspiration from [Tru77], we actually apply the crucial idea in a more general setting than it was originally formulated for. In this sense, the above lemma can be seen as generalizing Truemper's observations about generalized flows and min-cost flows to the setting of variable *gain factors* and non-linear minimum cost flow problems, somewhat similar in spirit to [Tru78] but applied specifically to the setting of optimal power flows.

Using the above lemma, we now prove that under the loss function $\eta^{\log}$, TR-optimality implies DC-feasibility, as long as the capacity constraints are not active. It is a well-known fact about DC electric circuits that the resulting current in a circuit minimizes thermal losses. The following theorem can be seen as an adaptation of that fact to the objective function of cost-minimization typically used in OPF and TCEP problems.

Theorem 2.14
Let $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ be a non-trivial energy system with $g > 0$ that is strictly feasible for the loss function $\eta^{\log}$ and let the underlying power network $(\mathcal{R}, \mathcal{L}, g, b)$ be uniform with material constant $\mu = \frac{g_l}{b_l}$ for all lines $l \in \mathcal{L}$. Let $f^+ > 0$ be a capacity vector. Let $(p, f) \in \mathbb{R}^{\mathcal{R}}_{\geq 0} \times \mathbb{R}^{\mathcal{L}}_{\geq 0}$ such that $|f_l| < f_l^+$ for all $l \in \mathcal{L}$. Then, (p, f) is cost-minimizing among all TR-feasible flows if and only if it is cost-minimizing among all DC-feasible flows.

Proof. We prove both directions separately.

"⇒" Let (p, f) be cost-minimizing among all TR-feasible flows. As $|f_l| < f_l^+$ for all $l \in \mathcal{L}$, it holds that (p, f) is also cost-minimizing among all demand-satisfying flows. From Lemma 2.13, we therefore obtain that (p, f) is DC-feasible. As (p, f) is cost-minimizing among all demand-satisfying flows, it is in particular cost-minimizing among those flows that are furthermore DC-feasible.

"⇐" Now, let (p, f) be cost-minimizing among all DC-feasible flows. Then, (p, f) is obviously TR-feasible. Suppose, therefore, that (p, f) is not cost-minimizing among TR-feasible flows. Let (p', f') be TR-feasible with $c(p', f') < c(p, f)$. Furthermore, denote the cost-minimizing demand-satisfying solution by (p^*, f^*). As every TR-feasible solution is also demand-satisfying, we obtain $c(p^*, f^*) \leq c(p', f') < c(p, f)$.

Since (p, f) is DC-feasible, there exists a vector of vertex angles φ that satisfies (2.11) together with (p, f). Similarly, we can obtain from Lemma 2.13 a vector of vertex angles φ^* that satisfies (2.11) together with (p^*, f^*). As (2.11) is linear, the same holds for any convex combination $\lambda(p^*, f^*) + (1 - \lambda)(p, f)$ with $\lambda \in [0, 1]$ (using the vector of vertex angles $\lambda\varphi^* + (1 - \lambda)\varphi$). Furthermore, $\lambda(p^*, f^*) + (1 - \lambda)(p, f)$ is also demand-satisfying. We may now choose $\lambda > 0$ small enough to obtain a point where $|\lambda f_l^* + (1 - \lambda) f_l| \leq f_l^+$ for all $l \in \mathcal{L}$ and $\lambda(p^*, f^*) + (1 - \lambda)(p, f)$ is hence DC-feasible. Due to the convexity of the cost function c, we have $c(\lambda(p^*, f^*) + (1 - \lambda)(p, f)) < c(p, f)$, a contradiction. □

The above theorem suggests that using the loss function $\eta^{\log}$ (which is an approximation of η^{AC} that is asymptotically as good as the more common η^2) allows us to drop the equations (2.11), leading to a much simpler model with the same accuracy. However, the restriction of $|f_l| < f_l^+$ is substantial, restricting ourselves in effect to the case where our problem does not contain any capacity constraints. While other constraints (such as demand satisfaction) remain part of the problem, this represents a drastic simplification and indeed, especially in the case of the TCEP, removes a core part of the problem.

Using the loss function $\eta^{\log}$ can thus change significantly the optimal solution within the set of TR-feasible flows, leading to a better representation of actual (DC model) load flows than in the original (lossless) Transport model. On the other hand, the set of TR-feasible flows remains almost the same (and substantially different from the set of DC-feasible flows), no matter which loss function is being used.

We approach the above-mentioned issue in two ways: In Section 2.2.2, we evaluate the accuracy of Transport model power flows under the loss function $\eta^{\log}$ empirically in a number of practical instances. We vary the values of a set of important parameters to obtain a better understanding of the sensitivity of our results and their usefulness in the context of TCEP problems. Then, in Sections 2.3 to 2.5, we obtain some theoretical insights into the differences between the sets of DC- and TR-feasible flows.

In preparation for our empirical evaluation, the next section focusses on issues that arise in implementing an approach according to the suggestions from Theorem 2.14. We explore ways to apply the above results (which use non-linear loss functions) to Linear Programming based approaches for the Transmission Capacity Expansion Problem.

2.2 LP Models for Transmission Capacity Expansion

The constraints (2.9) and (2.10) already provide a foundation for a Mathematical Programming model to solve the Transmission Capacity Expansion Problem. We can compute an optimal solution for the TR-TCEP (and also for the TR-OPF, by setting f^+ constant) using the following optimization problem:

$$\min\ {c^f}^\top f^+ + \sum_{i \in \mathcal{R}} c_i(p_i) \tag{2.35}$$

$$\text{s.t. } p_i + \sum_{l \in \delta^{\mathrm{in}}(i)} \left(f_l - \frac{\eta_l}{2}\right) - \sum_{l \in \delta^{\mathrm{out}}(i)} \left(f_l + \frac{\eta_l}{2}\right) \geq D_i \qquad \forall\ i \in \mathcal{R} \tag{2.36}$$

$$\eta_l \geq \eta_l(f_l^+, |f_l|) \qquad \forall\ l \in \mathcal{L} \tag{2.37}$$

$$f_l \leq f_l^+ \qquad \forall\ l \in \mathcal{L} \tag{2.38}$$

$$p \geq 0 \tag{2.39}$$

We have seen in Lemma 2.12 that for a given capacity vector f^+, the set of demand-satisfying flows is convex provided that the loss function η is convex in the fourth argument. If it is jointly convex in the third and fourth argument, then the above optimization problem is convex and hence in practically all cases theoretically solvable in polynomial time (see, e. g., [GLS12, Ch. 4], [NN94, Theorem 3.2.1]).

This is in stark contrast to the DC-TCEP, for which no such statement holds in general, even under the above convexity assumptions on η: It can be formulated as a mathematical programming problem by adding variables φ_i for each node i and the constraints (2.11) for every line to the optimization problem (2.35) to (2.39):

$$f_{ij} = b_{ij} f_{ij}^+ \cdot (\varphi_j - \varphi_i) \qquad \forall\ (i,j) \in \mathcal{L} \tag{2.40}$$

In the case of the DC-OPF (with f^+ fixed), the φ_i are the only variables in the above equation, which is thus linear and we obtain again a convex optimization problem. For variable f^+ (as in the case of the DC-TCEP), however, (2.40) is a non-convex quadratic equality constraint. Hence, the DC-TCEP is typically approached by discretizing f^+ and using a Mixed-Integer LP formulation (see, e. g., [Bah+01; AMC03; Zha+12]), which dramatically increases the computation time.

The TR-TCEP, on the other hand, consisting only of (2.35) to (2.39), can generally be solved in polynomial time (if the above convexity assumptions are made). This theoretical assessment however comes with serious drawbacks from a practical perspective: All of the above problems are typically considered over a large numbers of interconnected timesteps covering long time horizons (see our discussion at the end of Section 2.1.1), which results in very large-scale mathematical programming problems. For these cases, the practical performance of (theoretically efficient) general convex optimization algorithms is often insufficient, even for the TR-TCEP. Almost all case studies in the above problems therefore rely on Linear Programming models, for which highly developed and thus extremely powerful solution algorithms are available.

If we assume that c_i is piecewise-linear and $\eta_l(f_l^+, |f_l|) := \eta_l \cdot |f_l|$ is a constant fraction of $|f_l|$ that does not depend on f_l^+, then (2.35) to (2.39) can easily be written as a Linear Program. The first condition can be justified by assuming that in each region, a set of production units are available, each with linear production cost function, a very common simplification in large-scale TCEP studies. The second condition is a very common assumption that corresponds to approach (2) in Fig. 2.3.

Under the above assumptions, a solution to the TR-TCEP can be computed via linear programming using the above optimization problem. For optimal power flows under the Transport model, even more efficient combinatorial algorithms are available (both practically and theoretically, an optimal solution can even be computed in strongly polynomial time, see [GT89]). We will now discuss under which conditions other (not necessarily linear) loss function η can be approximated in a suitable way for use in a Linear Program. In particular, this will enable us to include an approximated version of $\eta^{\log}$ into a Linear Programming model.

2.2.1 Piecewise-linear Loss Approximations

In order to incorporate higher-order approximations of actual transmission losses into a linear-programming based model, piecewise-linear functions are a common tool. For instance, Alguacil, Motto, and Conejo [AMC03] include a piecewise-linear approximation of the quadratic loss function η^2 for fixed capacities f^+ into a DC model. Compared to a simple linear approximation, this approach achieves a much closer fit to the actual loss function that is furthermore reference point-free, i. e., independent of any previously estimated operating point (see Fig. 2.3). In terms of the above LP model, this corresponds to replacing (2.37) by

$$\eta_l \geq \bar{\eta}_l(f_l^+, |f_l|) \qquad \forall\, l \in \mathcal{L}$$

for some suitable piecewise-linear approximation $\bar{\eta}_l$ of the loss function corresponding to line l. While a good piecewise-linear approximation is easy to find for one-dimensional loss functions as they result from fixing f^+ (see,e. g., [HC94]), this is in general not necessarily the case for the two-dimensional functions that appear in the case of variable capacities [DLM10].

We will see, however, that if a potential loss functions η satisfies a property called *homogeneity*, then the case of variable capacities reduces to the much simpler case of fixed capacities. It will turn out that the required property is satisfied in particular by the loss functions η^2 and $\eta^{\log}$ mentioned above.

Definition 2.15
Let $n \in \mathbb{N}$ and let $C \subset \mathbb{R}^n_{\geq 0}$ be a cone. We call a function $h : C \to \mathbb{R}_{\geq 0}$ *homogenous*, if for all $x \in \mathbb{R}^n_{\geq 0}$ and $\lambda \geq 0$ it satisfies $h(\lambda x) = \lambda h(x)$.

Theorem 2.16
For any values of b_l and g_l, both $\eta_l^{\log}$ and η_l^2 are homogenous.

Proof. Let $f_l^+, f \in \mathbb{R}_{\geq 0}$. If $\lambda = 0$, then $\eta_l^{\log}(\lambda f_l^+, \lambda f_l) = \eta_l^{\log}(0,0) = 0 = 0 \cdot \eta_l^{\log}(f_l^+, f_l)$ (and analogously for η^2).

If $\lambda > 0$, then

$$\begin{aligned}\eta_l^{\log}(\lambda f_l^+, \lambda f_l) &= \frac{2b_l^2 \lambda f_l^+}{g_l}\left(\log\left(\exp\left(\frac{2g_l \lambda f_l}{b_l^2 \lambda f_l^+}\right)+1\right)-\log 2\right) - 2\lambda f_l \\ &= \frac{2b_l^2 \lambda f_l^+}{g_l}\left(\log\left(\exp\left(\frac{2g_l f_l}{b_l^2 f_l^+}\right)+1\right)-\log 2\right) - 2\lambda f_l = \lambda \eta_l^{\log}(f_l^+, f_l)\end{aligned}$$

and

$$\eta_l^2(\lambda f_l^+, \lambda f_l) = \frac{g_l}{\lambda f_l^+ b_l^2}(\lambda f_l)^2 = \frac{g_l \lambda}{f_l^+ b_l^2} f_l^2 = \lambda \eta_l^2(f_l^+, f_l). \qquad \square$$

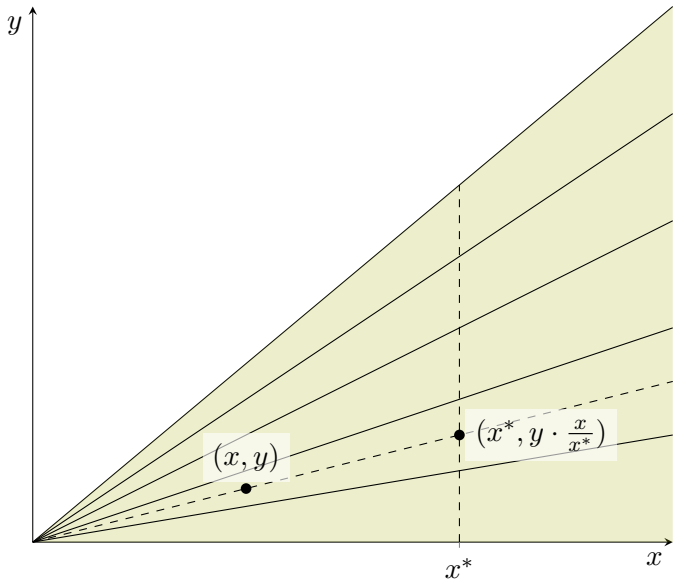


Figure 2.4: The situation in Theorem 2.17. The set C is shown in green.

Theorem 2.17
Let $C := \{(x, y) \in \mathbb{R}^2_{\geq 0} \mid y \leq x\}$ and let $h : C \subset \mathbb{R}^2_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a two-dimensional convex homogenous function. For some fixed value $x^* > 0$, let $\bar{h}(y) := \max_{i \in \mathcal{I}} a_i \cdot y - b_i$ be a piecewise linear approximation of $h(x^*, \cdot)$ that satisfies $\bar{h}(y) = 0$ for all y with $h(x^*, y) = 0$ and $\frac{|\bar{h}(y) - h(x^*, y)|}{h(x^*, y)} \leq \alpha$ for all $y \leq x^*$ with $h(x^*, y) > 0$. Then

$$h^*(x, y) := \max_{i \in \mathcal{I}} a_i \cdot y - \frac{b_i}{x^*} \cdot x \tag{2.41}$$

is a piecewise-linear approximation of h that satisfies $h^*(x, y) = 0$ for all $y < x$ with $h(x, y) = 0$ and $\frac{|h^*(x,y) - h(x,y)|}{h(x,y)} \leq \alpha$ for all $(x, y) \in C$ with $h(x, y) > 0$.

Proof. Let $(x, y) \in C$. If $x = 0$ then $y = 0$ and, since h is homogenous, $h(x, y) = 0$. On the other hand, (2.41) implies that $h^*(x, y) = 0$.

Now, let $x > 0$. Then

$$\begin{aligned}
|h^*(x, y) - h(x, y)| &= \left| h(x, y) - \max_{i \in \mathcal{I}} a_i \cdot y - \frac{b_i}{x^*} \cdot x \right| \\
&= \left| \frac{x}{x^*} \cdot h\left(x^*, y \cdot \frac{x^*}{x}\right) - \frac{x}{x^*} \cdot \left(\max_{i \in \mathcal{I}} a_i \cdot y \cdot \frac{x^*}{x} - b_i \right) \right| \\
&= \frac{x}{x^*} \cdot \left| h\left(x^*, y \cdot \frac{x^*}{x}\right) - \bar{h}\left(y \cdot \frac{x^*}{x}\right) \right|.
\end{aligned}$$

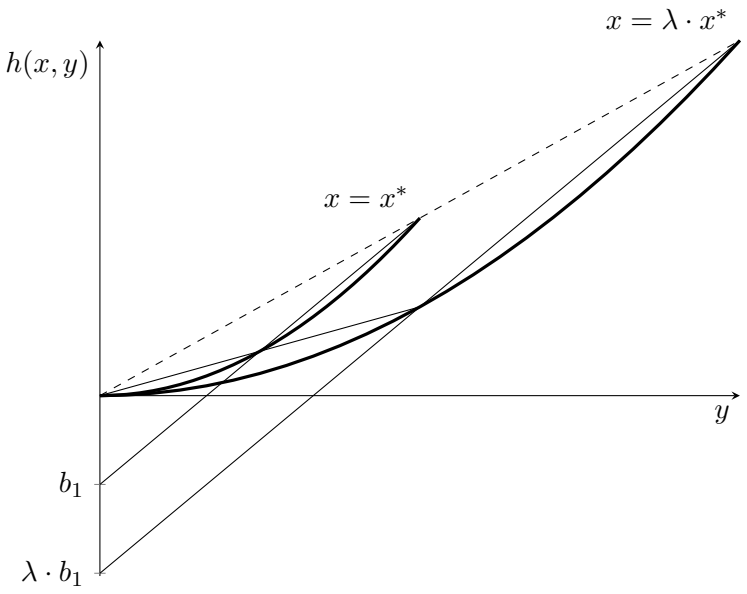


Figure 2.5: The situation in Theorem 2.17 projected along the x-axis. Given a piecewise-linear approximation of h for any fixed value of $x = x^*$ (shown in thin), this approximation can be extended to all other values of x: The segments' slopes remain the same and the segments' intercepts change linearly in x (see (2.41)). Based on a figure from [AS13].

If $h(x, y) = 0$ then by homogeneity, $h(x^*, y \cdot \frac{x^*}{x}) = 0$. But then $\bar{h}\left(y \cdot \frac{x^*}{x}\right) = 0$, which by the above equality implies that $h^*(x, y) = 0$.

If, on the other hand, $h(x, y) \neq 0$, then we reduce everything to the one-dimensional case where $x = x^*$ (see Fig. 2.4):

$$\begin{aligned}
\frac{|h^*(x,y) - h(x,y)|}{h(x,y)} &= \frac{x}{x^* \cdot h(x,y)} \cdot \left| h\left(x^*, y \cdot \frac{x^*}{x}\right) - \bar{h}\left(y \cdot \frac{x^*}{x}\right) \right| \\
&\leq \frac{1}{h(x,y)} \cdot \frac{x}{x^*} \cdot h\left(x^*, y \cdot \frac{x^*}{x}\right) \cdot \alpha \\
&= \frac{1}{h(x,y)} \cdot h(x,y) \cdot \alpha = \alpha
\end{aligned}$$

□

An interpretation of Theorem 2.17 is given in Fig. 2.5: For each value of x, we can compute a corresponding piecewise-linear approximation of $h(x, \cdot)$ such that the slopes of each linear segment are the same for all of these piecewise-linear approximations. The only difference between the different piecewise-linear approximations are the values of the intercepts, these change linearly in x.

In particular, the piecewise-linear approximations that result from Theorem 2.17 are *optimal* in the sense that there can be no piecewise-linear approximation of the function f with $|\mathcal{I}|$ linear segments that globally achieves a better approximation ratio: It would in particular have to guarantee that ratio for every fixed x^* and can use no more than $|\mathcal{I}|$ (one-dimensional) segments there. This, however, allows us to construct a piecewise-linear approximation from Theorem 2.17 with the same approximation ratio.

The above results provide us with a method to easily incorporate optimal approximations of certain loss functions into a Linear Programming based model for the TR-TCEP. In particular, we can use loss functions that depend on both the flow and the capacity of a line, an aspect that to our knowledge is neglected by all available applications of the Transport model in this context.

It should also be noted again that the approximated loss functions are reference-point free (if this is true for the underlying loss function): The accuracy of the approximation of transmission losses is the same, regardless of what initial capacity was used to define the model constraints. The model is thus independent of any assumptions about a likely future infrastructure or operating point.

By using a piecewise-linear approximation of the loss function $\eta^{\log}$, we can also attempt to leverage Theorem 2.14 in the context of our LP model (2.35) to (2.39). While Theorem 2.14 only holds for the non-linear loss function $\eta^{\log}$, the next section provides some empirical evidence that the result carries over to piecewise-linear approximations of $\eta^{\log}$, provided that the approximation is sufficiently fine. In fact, an approximate statement along the lines of Theorem 2.14 could be proved using piecewise-linear approximations of the *transmission input/output functions* $1 \pm \eta^{\log}$. These are a more direct analogue of (non-linear) gain factors as used in [Tru78], but move us conceptually further away from their interpretation as approximations of ohmic losses in electrical networks.

2.2.2 Empirical Evaluation

While Theorem 2.14 guarantees a perfect match between power flows under the Transport model and the DC model, this guarantee applies only if all the prerequisites of that theorem are met. While non-triviality and strict feasibility represent minor technical requirements that impose no real restriction in a practical scenario, three important caveats remain:

Piecewise-linear approximation The statement of Theorem 2.14 applies to the (convex non-linear) loss function $\eta^{\log}$. However, as argued in Section 2.2, practical considerations require us to use Linear Programming based models in many cases. Thanks to our observations from Section 2.2.1, we can translate the loss function $\eta^{\log}$ into the realm of Linear Programming using a piecewise-linear

approximation with very advantageous properties. Nonetheless, this approach entails an approximation error with the potential to compromise the applicability of Theorem 2.14.

Uniformity In Theorem 2.14, we require the underlying network to be uniform. As we have mentioned in Section 2.1, this assumption is generally reasonable while we concentrate on a subnetwork which covers only a single voltage level. Nonetheless, small deviations even between transmission lines of the same type must be expected. Furthermore, lines operating at different voltage levels do interact in reality and analyzing this interaction would require our model to include different types of equipment.

Capacity constraints Most significantly, Theorem 2.14 holds only if none of the capacity constraints associated with transmission lines in the network are binding. This assumption is almost certain to be violated in any practically relevant instance of the TCEP. It is therefore important to understand the behavior of Theorem 2.14 with respect to this limitation, i. e., the sensitivity of the result under partial violation of our assumption.

In order to address these potential limitations of Theorem 2.14, Widmann has undertaken an empirical study in the context of his Bachelor's thesis [Wid18], the results of which we summarize below. In the study, the accuracy of flows under the Transport model is compared to the result obtained using the DC model.

The study uses three different base networks for the analysis: the *Garver 6-Bus-System* on eight lines, a small but commonly used test system from [Gar70], a network on 24 lines derived from the *IEEE Reliability Test System* [IEEE79] released as a benchmark instance by the *Institute of Electrical and Electronics Engineers (IEEE)* and a version of the model ELMOD-DE [Ege16] on 333 lines, an open-source model of the German electricity system. These networks cover a broad range of system sizes, a detailed description of the test systems (and the broader test setup) can be found in [Wid18].

In each of the above networks different parameters were varied independently in order to gain a better understanding of each of the above-mentioned potential sources of error individually. In each case, the *error*, which is defined as the (absolute) difference between DC- and Transport flow, is summed over all lines and scaled by the total magnitude of the DC flow:

$$\mathrm{Error}(f^{\mathrm{TR}}, f^{\mathrm{DC}}) = \frac{\left\| f^{\mathrm{TR}} - f^{\mathrm{DC}} \right\|_1}{\left\| f^{\mathrm{DC}} \right\|_1}$$

Widmann [Wid18] begins by computing an optimal solution to the TR-OPF using the loss function $\eta^{\log}$ and comparing the result to an optimal solution using a linear

Table 2.1: Relative error against the DC model for the Transport model using the loss function $\eta^{\log}$ and a Transport model using a linear loss function η^{lin} (from [Wid18]).

		Garver	IEEE	ELMOD-DE
Error (%)	$\eta^{\log}$	0.7	0.6	0.6
	η^{lin}	27.1	49.3	59.8

loss function η^{lin} that has been scaled to produce the same magnitude of losses across the entire network. In both cases, the resulting error is measured against the optimal solution to the corresponding DC-OPF using the loss function $\eta^{\log}$. While all networks are almost uniform by default, transmission capacities were relaxed in order to obtain an unrestricted power flow in all networks. Similarly, a fine piecewise-linear approximation using 100 linear pieces was used with the effect that all of the potential sources of error mentioned above are minimized.

The results are shown in Table 2.1: In the case of all three test networks, the modified Transport model using the loss function $\eta^{\log}$ represents a tremendous improvement over the base Transport model using a linear loss function. While the accuracy of the latter model varies substantially across the three networks, the result from the modified Transport model is always within 1 % of the DC model, representing an improvement by a factor of 30 to 100. This demonstrates the effect of Theorem 2.14 in the case of all three networks and, since all of the prerequisites of the theorem are (almost) satisfied, this is exactly what we expected.

Piecewise-linear approximation On the other hand, the size of the Linear Program used to solve the Optimal Power Flow problem grows approximately linearly in the number of linear pieces used. In order to keep computation times low, it is therefore advisable to only use as many linear pieces as are required to achieve the desired accuracy. Figure 2.6 shows the effect of the granularity of piecewise-linear approximations on the accuracy of the result.

While the results vary a lot for smaller networks (especially for the 8-line Garver system), the error generally drops very quickly: In all three networks, an error level of around 5-10 % can already be reached for a small number of linear pieces. From the error level of the linear loss function η^{lin} (which corresponds to a piecewise-linear approximation with a single piece), the error drops to 5-10 % for around eight linear pieces. From there on, the accuracy can be increased further to below 0.5 % for 200 linear pieces, but the increase in computational effort that has to be spent in return becomes increasingly outsized.

In the right-hand part of Fig. 2.6, it can be seen that using a rather coarse piecewise-linear approximation of the loss function $\eta^{\log}$ already, we can obtain a reasonable approximation of the result of the DC-OPF from solving the corresponding TR-OPF.

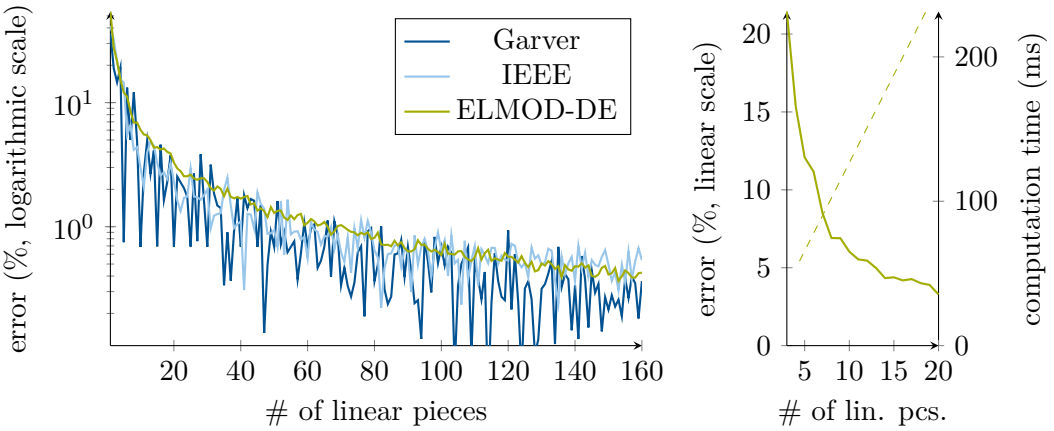


Figure 2.6: Error for the Transport model using different piecewise-linear approximations of the loss function $\eta^{\log}$. On the left, the error is shown on a logarithmic scale. The behavior is quite similar in all three networks, with variation reducing as the size of the model grows. On the right, the error for a smaller number of linear pieces in the ELMOD-DE network (the largest of the analyzed networks) is shown in detail, this time on a linear scale together with a linear regression plot (dashed) of the computation time (data from [Wid18]). It can be seen that a decent approximation can be achieved by a small number of linear pieces already, but any increase in accuracy beyond ~ 5 % comes at a high computational cost.

Uniformity Regarding uniformity, starting from the original material parameter μ (which is almost uniform in the IEEE network and perfectly uniform in the Garver and ELMOD-DE network), Widmann draws a new material parameter uniformly at random from the range $[\mu, \lambda \cdot \mu]$, independently for each individual line. Ten different values for λ between 1 and 2 were used and the experiment was repeated 1000 times for each value of λ (see Fig. 2.7).

In all cases it can be observed that, as long as a network is sufficiently close to uniform, the error compared to the DC model also remains limited. Broadly speaking, it seems that a material parameter that deviates slightly from uniformity does not create any substantial issues unless the deviation is distributed in a particularly problematic fashion. This is what occasionally happens in the smaller networks Garver and IEEE, where the smaller number of lines means that any worst-case (and best-case) constellation is very likely to be covered by a given draw of parameters, as reflected by the range of error values which consequently reduces as the size of the network increases.

Capacity Constraints Finally, to measure the effect of binding capacity constraints, line capacities were reduced in each of the networks to values between 35 % and

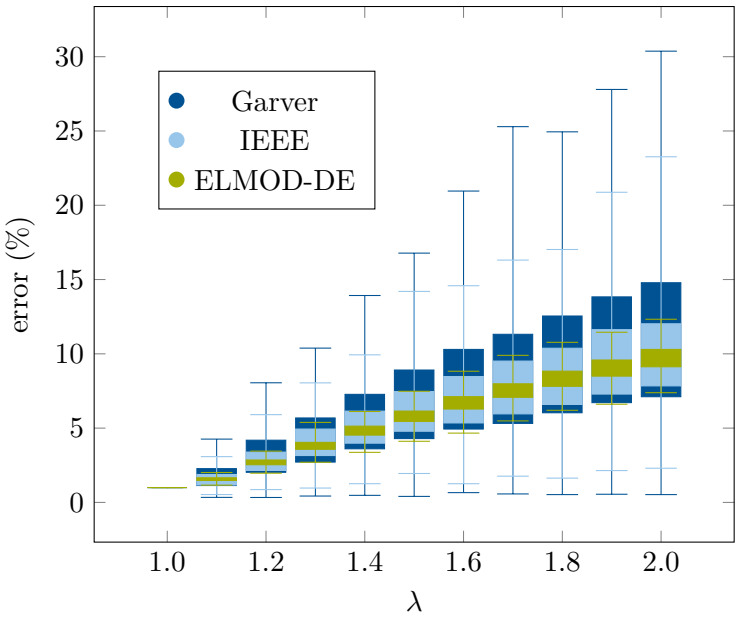


Figure 2.7: Distribution of errors for the three networks with the material parameter μ drawn uniformly at random for each individual line from the range $[\mu, \lambda \cdot \mu]$. The boxes cover half of all data points around the median while the whiskers cover the whole range of data points (data from [Wid18]).

100 % of the original capacity (Fig. 2.8). The results in this respect are less promising: While the accuracy of the modified Transport model always remains higher than the original Transport model (as was to be expected), the difference changes substantially as capacities become more and more limiting: On the one hand, the error of the original Transport model is reduced slightly, since the reduced capacities limit the possibilities for different flows in the network. On the other hand, the violation of the respective prerequisite of Theorem 2.14 becomes more and more apparent, reducing the difference between flows from both versions of the Transport model.

This effect begins as soon as the first capacity constraint becomes binding but, again, is more dramatic in the smaller networks: In the networks Garver and IEEE, the error grows to around 70 % of the error in the original Transport model almost immediately after the first capacity constraint becomes binding. In the network ELMOD-DE, on the other hand, capacities can be reduced by an additional 20 % without dramatically increasing the error value, even after the first capacity constraint becomes binding. At line capacities of around 40 % of the original value, the error in the modified Transport model reaches almost the level of the error in the original Transport model in all three networks.

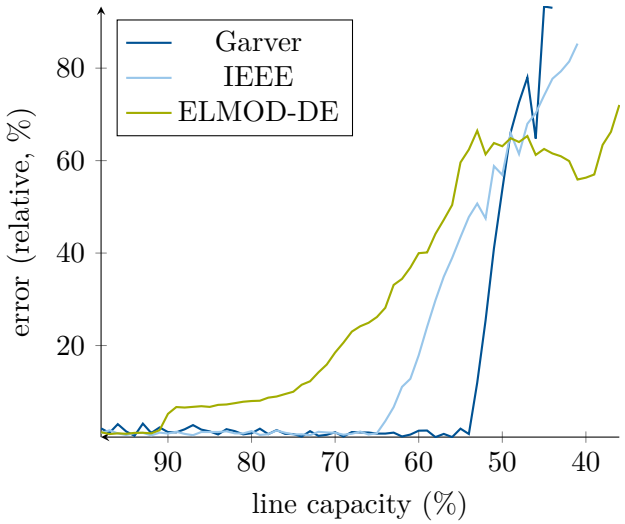


Figure 2.8: Relative error in the Transport model using the loss function $\eta^{\log}$ as a percentage of the error in the Transport model using the linear loss function η^{lin} for reduced line capacities. The error remains negligible while no capacity constraints are binding, it then grows to almost the size of the error in the original Transport model just before the DC model becomes infeasible (data from [Wid18]).

If capacities are reduced even further, the DC model becomes infeasible, which means that we no longer have a benchmark to compute error values for either the original or the modified Transport model (which remain feasible). This reflects our observation from the end of Section 2.1.4: The loss function has only a minor effect on the feasibility region of the underlying OPF problem, which means that regardless of the loss function used, the fundamental difference between feasibility regions of the DC model and the Transport model remains. Obtaining a better understanding of this difference is the aim of the following section.

Nonetheless, as Widmann [Wid18] points out, the optimal solution of a Transmission Capacity Expansion Problem using the modified Transport model can provide valuable information about the region of the solution space in which to look for an optimal solution for the DC-TCEP: Restricting the possible capacities in a discretized Mixed-Integer LP model for the DC-TCEP (see Section 2.2) to a maximum of ~ 120 % of the optimal capacity under the modified Transport model yields a solution within 5 % of the optimal solution to the full DC-TCEP. At the same time, computation time in the ELMOD-DE network (the largest of the three networks) is reduced by ~ 99 %.

2.3 The Sets of TR- and DC-feasible Solutions

Theorem 2.14 leaves three possible reasons why the stated equivalence of optimal TR- and DC-feasible flows might not hold in practice if a piecewise-linear approximation of the loss function $\eta^{\log}$ according to Section 2.2.1 is used: The error incurred by the piecewise-linear approximation, differences in the material parameter μ or binding capacity constraints. Of these, as the examples in Section 2.2.2 suggest, the last source of error is the most severe. This is true particularly since it is also the most likely to occur in practice: Capacity constraints can certainly not be assumed to always be non-binding in general. This becomes particularly relevant in the context of Transmission Capacity Expansion problems, where in any optimal solution, all capacity constraints will be binding (at least in some timestep, otherwise a lower-cost feasible solution could be obtained by reducing the capacity on lines where the constraint is not binding).

In this section, we will therefore focus on the sets of TR- and DC-feasible solutions themselves, for a fixed capacity vector and irrespective of the objective or the loss function used. We will try and understand the structure of both sets and compare them with each other. More specifically, we will first obtain some general statements about the relative size of both sets. We will then develop a combinatorial description of extremal solutions for the DC model and finally in Section 2.5.6 we will characterize a family of graphs on which the DC model is structurally very similar to the Transport model. This family in particular covers certain benign network topologies that are currently very much in the focus of research into feasible regions with respect to the AC model sparked by [LL12].

Let $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ be an energy system. Recall from Definition 2.5 that a pair (p, f) is TR-feasible for a capacity vector $f^+ \in \mathbb{R}^{\mathcal{L}}_{\geq 0}$ if it satisfies

$$p_i + \sum_{l \in \delta^{\text{in}}(i)} \left(f_l - \frac{\eta_l(f_l^+, |f_l|)}{2} \right) - \sum_{l \in \delta^{\text{out}}(i)} \left(f_l + \frac{\eta_l(f_l^+, |f_l|)}{2} \right) \geq D_i \quad \forall\, i \in \mathcal{R} \tag{2.42}$$

$$|f_l| \leq f_l^+ \qquad \forall\, l \in \mathcal{L}. \tag{2.43}$$

It is DC-feasible if in addition there exists a vector φ of vertex angles satisfying

$$f_{ij} = b_{ij} f_{ij}^+ \cdot (\varphi_j - \varphi_i) \qquad \forall\, (i,j) \in \mathcal{L}. \tag{2.44}$$

As we want to focus on the core differences between TR- and DC-feasibility in this chapter, we will make the following simplifying assumptions: First, we use the loss function η^0, thereby eliminating the term $\eta_l(f^+, |f_l|)$ above. In the previous section, we have argued that the *optimal* solution can be heavily influenced by the choice of the loss function and that this can be used e. g. to reduce to some extent the differences between Transport model and DC model. In this section, we now want to analyze in more detail the exact nature of these differences, which are most apparent in the lossless version.

Furthermore, we assume that b_l has the same sign for all $l \in \mathcal{L}$ (typically, $b < 0$ for all overhead transmission lines). we also allow some upper bound p_i^+ to be given on the net production (or *injection*) in vertex i, i. e., $p_i - D_i \leq p_i^+$. This constraint can be used e. g. to capture the domain where $c_i(p)$ is finite (shifted by the demand D_i).

We will also assume that the demand satisfaction inequality (2.42) always holds with equality. This assumption is without loss of generality with respect to feasibility, since for any (p, f) strictly satisfying (2.42) there exists a vector p' with $p'_i < p_i$ such that equality holds and $p'_i - D_i \leq p_i^+$, i. e., (p', f) is still feasible (if necessary, we allow $p_i < 0$).

To further simplify the notation, we assume that $f_l^+ > 0$ for all $l \in \mathcal{L}$ (otherwise, remove the line l from the power network), that the network is *anti-symmetric* (since edge directions were arbitrary to begin with, we can choose them so that edges between two vertices point in the same direction and aggregate parallel lines to one equivalent line) and that the power network is weakly connected (otherwise we can consider all connected components independently).

On the other hand, we will allow two slight generalizations from the setting that we have considered until now: We may require a lower bound p_i^- for the injection in vertex $i \in \mathcal{R}$ (in addition to the upper bound p_i^+) and we may specify a lower bound f_l^- for the flow on line $l \in \mathcal{L}$ independently of the upper bound f^+. These generalizations have no effect on most of the results in the remainder of this chapter (unless specified otherwise). In the context of power networks, however, these lower bounds can always be understood as $p_i^- = -\infty$ and $f_i^- = -f_i^+$.

A (fixed-capacity) network graph is now defined as follows:

Definition 2.18 (Network Graph)
A *network graph* $(V, E, b, \bar{p}, \bar{f})$ consists of a weakly connected, anti-symmetric directed graph (V, E) together with an *elasticity vector* $b \in \mathbb{R}^E$ with $b > 0$ and pairs of upper and lower bounds $\bar{p} = (p^-, p^+) \in (\mathbb{R} \cup \{-\infty\})^V \times (\mathbb{R} \cup \{\infty\})^V$ and $\bar{f} = (f^-, f^+) \in (\mathbb{R} \cup \{-\infty\})^E \times (\mathbb{R} \cup \{\infty\})^E$ for vertices and edges, respectively, such that $p^- \leq p^+$ and $f^- \leq f^+$.

Note that in a network graph, we have chosen to require $b > 0$ instead of $b < 0$, which would typically hold in a power transmission network. This is merely a technical assumption which makes the notation in this chapter easier to read and gives us a clearer view on some connections to known problems in the context of network flows. We will see in Remark 2.19 that inverting the sign of b has no effect on the resulting feasible region, which means that requiring $b > 0$ is effectively equivalent to merely requiring that all b_l have the same sign.

Note furthermore that we write, e. g., $p^+ = \infty$ to denote that $p_e^+ = \infty$ for all $e \in E$. To denote that no bounds apply at all, we abbreviate $\bar{p} = (-\infty, \infty)$ by $\bar{p} = \infty$ (and analogously for $\bar{f}$). Given a *network graph* $N = (V, E, b, \bar{p}, \bar{f})$, let A be the incidence

matrix of the graph (V, E) as defined in (A.1). We define the *elasticity matrix* B of N by $B = A \cdot \operatorname{diag}(b)$, i. e.,

$$B_{ie} := \begin{cases} b_e & e \in \delta^{\text{in}}(i) \\ -b_e & e \in \delta^{\text{out}}(i) \\ 0 & \text{else.} \end{cases} \tag{2.45}$$

We can now summarize the constraints for TR- and DC-feasibility in vector-matrix-notation: With p as the vector of net productions (or *injections*), the system of linear equations (2.42) (assuming, as above, that the inequalities always hold with equality) can be written as $p + Af = 0$. This system of linear equations is well-known from the literature on network flows, see, e. g., [PS98, Chs. 3.4 and 7.3] or [Sch03, Ch. 13.2]. Similarly, with f as the vector of flows and the entries of B as the effective susceptance (since $\bar{f}$ is constant, the resulting effective susceptance is constant, as well), we can write (2.44) as $B^\top \varphi = f$: With $e = (v_1, v_2)$,

$$f_e = (B^\top \varphi)_e = \sum_{v \in V} B_{ve} \varphi_v = \sum_{\substack{v \in V \\ e \in \delta^{\text{in}}(v)}} b_e \varphi_v - \sum_{\substack{v \in V \\ e \in \delta^{\text{out}}(v)}} b_e \varphi_v = b_e \cdot (\varphi_{v_2} - \varphi_{v_1}).$$

For a given network graph $G = (V, E, b, \bar{f}, \bar{p})$, we define the polyhedra

$$Q^{\text{TR}}(G) := \left\{ (p, f) \in \mathbb{R}^V \times \mathbb{R}^E \;\middle|\; \begin{array}{c} p + Af = 0 \\ f^- \le f \le f^+ \\ p^- \le p \le p^+ \end{array} \right\}$$

and

$$Q^{\text{DC}}(G) := \left\{ (p, f, \varphi) \in \mathbb{R}^V \times \mathbb{R}^E \times \mathbb{R}^V \;\middle|\; \begin{array}{c} B^\top \varphi = f \\ p + Af = 0 \\ f^- \le f \le f^+ \\ p^- \le p \le p^+ \end{array} \right\}.$$

Under the assumptions above, these polyhedra capture exactly the TR- and DC-feasible points that we are interested in (translated by the demand vector D), as the following remark shows:

Remark 2.19
Let $(\mathcal{R}, \mathcal{L}, g, b, D, c)$ be an energy system with $b < 0$ and let $f^+ \in \mathbb{R}^{\mathcal{L}}, f^+ > 0$ be a capacity vector. Now, define $b'_l := -b_l \cdot f_l^+$ for all $l \in \mathcal{L}$, $p_i^+ := \max\{p \in \mathbb{R} \mid c_i(p) < \infty\}$ for all $i \in \mathcal{R}$ and $f_l^- = -f_l^+$ for all $l \in \mathcal{L}$. Now, consider the network graph $G = (\mathcal{R}, \mathcal{L}, b', (-\infty, p^+), (f_l^-, f_l^+))$.

Then, $(p, f) \in Q^{\text{TR}}(G)$ if and only if $(p + D, f)$ is TR-feasible for f^+ and the loss function η^0. Analogously, there exists $\varphi' \in \mathbb{R}^{\mathcal{R}}$ such that $(p, f, \varphi') \in Q^{\text{DC}}$ if and only if

$(p + D, f)$ is DC-feasible for f^+ and the loss function η^0 (using the vector of vertex angles $\varphi = -\varphi'$): Let B denote the elasticity matrix for the network graph G and let $l := (i, j) \in \mathcal{L}$. Then,

$$f_l = (B^\top \varphi')_l = \sum_{i' \in \mathcal{R}} B_{il} \varphi'_{i'} = b'_l \cdot (\varphi'_j - \varphi'_i) = b_l f_l^+ \cdot (\varphi_j - \varphi_i). \tag{2.46}$$

Note that, as indicated above, although we have $b < 0$ and use $b' > 0$ to define $Q^{\mathrm{DC}}(G)$, the resulting feasible solutions are DC-feasible with respect to the original values of b (the only difference is the inverted sign of the vector of vertex angles φ).

As mentioned above, our definition of a network graph is slightly more general than required by Remark 2.19. In particular, we do generally not assume that $-f^- = f^+$ or $p^- = -\infty$. Where necessary, we will occasionally explicitly invoke these requirements.

In addition, note that in a network graph, elasticity vector and edge capacities are independent, which is not true in a power network: By (2.2), increasing the capacity of a line also increases its susceptance. Throughout most of the remainder of this chapter, it makes sense to ignore this dependence, since we only deal with fixed-capacity networks (and it simplifies the notation significantly). However, we sometimes use a network graph with modified capacities in our arguments. In these cases we will explicitly discuss the implications of changing the elasticity along with the capacities.

We conclude this introduction by some observations about the polyhedra $Q^{\mathrm{TR}}(G)$ and $Q^{\mathrm{DC}}(G)$. First, note that while $Q^{\mathrm{DC}}(G)$ lives in a higher-dimensional space than $Q^{\mathrm{TR}}(G)$, the additional constraint $B^\top \varphi = f$ also restricts the set of values feasible for (p, f): In a similar way as DC-feasibility requires the existence of a vector of vertex angles φ that satisfies (2.44), there must exist a ($|V|$-dimensional) vector φ which satisfies the set of $|E|$ additional equations $B^\top \varphi = f$ for a point $(p, f) \in Q^{\mathrm{TR}}(G)$ in order for (p, f, φ) to lie in $Q^{\mathrm{DC}}(G)$.

In other words, we require that $f \in \mathrm{im}(B^\top)$, which means that the $|E|$-dimensional vector f is restricted to a lower-dimensional subspace. Regarding the dimension of that subspace, the following proposition yields in particular that $\dim(\mathrm{im}(B^\top)) = |V| - 1$, which means that whenever a network graph contains more edges than the minimum of $|V| - 1$ required for weak connectivity, then the constraint $B^\top \varphi = f$ represents a real restriction of the values that the vector f (and as a consequence also p) can take.

We conclude this introduction by observing some useful properties of the incidence and elasticity matrices A and B as defined above, and of the matrix $AB^\top$ that appears if we substitute $B^\top \varphi = f$ into $p + Af = 0$ (as, e. g., in (2.47) below).

Proposition 2.20
Let (V, E) be a weakly connected directed graph with incidence matrix A. Let $b \in \mathbb{R}^E$, $b > 0$ and let B be the elasticity matrix as defined in (2.45)*. Then,*

$$\ker(A^\top) = \ker(B^\top) = \ker(AB^\top) = \mathbb{R} \cdot \mathbb{1}.$$

Proof. It is a well-known fact that $\operatorname{rank}(A) = |V| - 1$ (this can be found, e. g., in [Bap14, Lemma 2.2]) and since $b > 0$ and $B = A \cdot \operatorname{diag}(b)$, it is clear that $\operatorname{rank}(B) = \operatorname{rank}(A) = |V| - 1$. On the other hand, observe that $A^\top(\varphi + \alpha \cdot \mathbb{1}) = A^\top\varphi$ (and analogously for B). Hence $\ker(A^\top) = \ker(B^\top) = \mathbb{R} \cdot \mathbb{1}$.

Since obviously $\ker(B^\top) \subset \ker(AB^\top)$, it only remains to show that $\ker(AB^\top) \subset \ker(B^\top)$: Let $\varphi \notin \ker(B^\top)$, i. e., $f := B^\top\varphi \neq 0$. For contradiction, suppose that $AB^\top\varphi = 0$. Choose an edge e with $f_e > 0$ (or analogously $f_e < 0$ if $f \leq 0$). Let $e = (v_1, v_2)$, then $\varphi_{v_2} > \varphi_{v_1}$. As $Af = 0$, there must be an edge $e' = (v_2, v_3)$ with $f_{e'} > 0$ carrying flow away from v_2 (or $e' = (v_3, v_2)$ and $f_{e'} < 0$) and hence $\varphi_{v_3} > \varphi_{v_2}$. Iterate this argument to obtain a cycle along which f contains a positive flow. This implies that φ increases strictly along the cycle, a contradiction. □

The matrix $AB^\top$ is known in electrical engineering as the *nodal admittance matrix*. If $b_e = 1$ for all $e \in E$, then the matrix $AB^\top = AA^\top$ is known as the *Laplacian matrix* of (V, E), which captures many interesting properties of the graph (see, e. g., [Chu97]).

The following sections will consider the sets Q^{TR} and Q^{DC} from two different perspectives (see Fig. 2.9): First with respect to the projection onto the space of p-variables (the set of *feasible power injections*), then with respect to the projection onto the space of f-variables (the set of *feasible differential flows*, the origin of this term will become clear in Section 2.5).

This approach is motivated by the fact that the projections of $Q^{\mathrm{DC}}(G)$ have the same combinatorial structure as the original polyhedra, i. e., no information is lost by the projection:

First, note that $\dim(Q^{\mathrm{DC}}(G)) \leq |V|$ since the polyhedron is contained in the $|V|$-dimensional linear subspace defined by $B^\top\varphi = f, p + Af = 0$. If we identify this linear subspace with $\mathbb{R}^V$, we obtain the following alternative representation of $Q^{\mathrm{DC}}(G)$:

$$Q^{\mathrm{DC}}_\varphi(G) := \left\{ \varphi \in \mathbb{R}^V \;\middle|\; \begin{array}{c} f^- \leq B^\top\varphi \leq f^+ \\ p^- \leq -AB^\top\varphi \leq p^+ \end{array} \right\} \tag{2.47}$$

Rather than as orthogonal projections of $Q^{\mathrm{DC}}(G)$, the set of feasible power injections and the set of feasible differential flows can now be seen as linear transformations of $Q^{\mathrm{DC}}_\varphi(G)$, more specifically $AB^\top Q^{\mathrm{DC}}_\varphi(G)$ and $B^\top Q^{\mathrm{DC}}_\varphi(G)$, respectively. Finally, note that $Q^{\mathrm{DC}}_\varphi(G)$ contains the 1-dimensional linear subspace $\mathbb{R}\cdot\mathbb{1}$. To eliminate this lineality space, we can select a vertex $v_0 \in V$ and define

$$Q^{\mathrm{DC}}_{\varphi_0}(G) := Q^{\mathrm{DC}}_\varphi(G) \cap \{\varphi \in \mathbb{R}^V \mid \varphi_{v_0} = 0\}.$$

From Proposition 2.20, we have that $\ker(B^\top) = \ker(AB^\top) = \mathbb{R} \cdot \mathbb{1}$ and hence $\ker(B^\top) \cap Q^{\mathrm{DC}}_{\varphi_0}(G) = \ker(AB^\top) \cap Q^{\mathrm{DC}}_{\varphi_0}(G) = \{0\}$. The transformations, restricted to $Q^{\mathrm{DC}}_{\varphi_0}(G)$, are thus invertible and by composition with the (equally invertible) function mapping $Q^{\mathrm{DC}}(G)$ to $Q^{\mathrm{DC}}_\varphi(G)$, we obtain the desired statement.

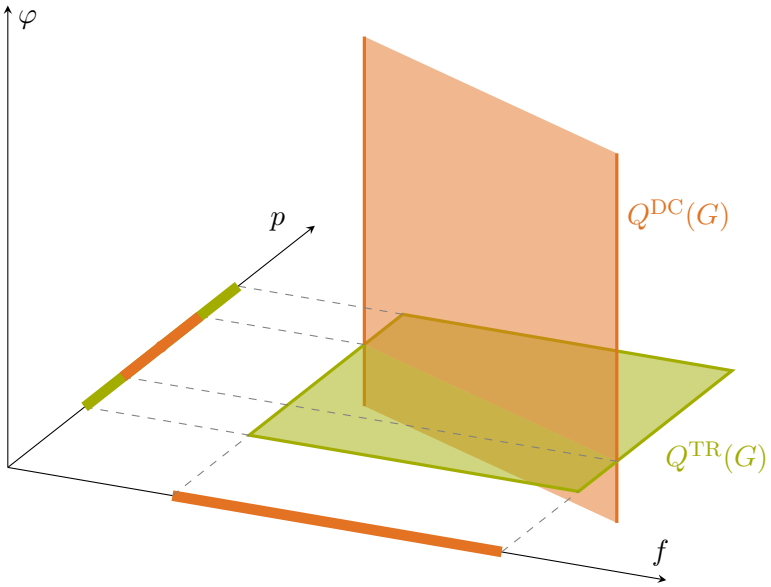


Figure 2.9: A sketch of the sets $Q^{\mathrm{TR}}(G)$ (shown in green) and $Q^{\mathrm{DC}}(G)$ (shown in orange), together with their respective projections onto the space of f-variables and the space of p-variables. Note that, while $Q^{\mathrm{DC}}(G)$ lives in a higher-dimensional space, its projection into the space of (p, f)-variables is actually of *lower* dimension than the set $Q^{\mathrm{TR}}(G)$. Furthermore, once we remove the lineality space contained in $Q^{\mathrm{DC}}(G)$, it is isomorphous to its projections into the space of p- and f-variables, respectively (see Lemma 2.22). Note, however, that in the case where $|E| \geq |V|$ the projection of $Q^{\mathrm{TR}}(G)$ into the space of f-variables is actually of higher dimension than the projection of $Q^{\mathrm{DC}}(G)$ (which cannot be shown in this sketch).

Hence, if we remove the lineality space contained in $Q^{\mathrm{DC}}(G)$ by fixing the value of φ to 0 in some arbitrary vertex v_0, then there exists a 1-to-1-correspondence of vertices (and also higher-dimensional faces) between $Q^{\mathrm{DC}}(G) \cap \{(p, f, \varphi) \mid \varphi_{v_0} = 0\}$ and its projections into the space of p- and f-variables, respectively (as well as $Q^{\mathrm{DC}}_{\varphi_0}(G)$), which we prove (for the space of p-variables) after formally introducing the corresponding projection below (Lemma 2.22).

The case of the polytope $Q^{\mathrm{TR}}(G)$ is somewhat simpler: By the equality $p + Af = 0$, the values of p are uniquely determined by f, already. The projection of $Q^{\mathrm{TR}}(G)$ onto the space of f-variables thus maintains the entire structure of $Q^{\mathrm{TR}}(G)$, whereas the projection onto the space of p-variables (which is generally of lower dimension) loses all information that is associated with flows along cycles in G.

2.4 Feasible Power Injections

In this section, we are interested in the sets of feasible power injections, in other words the projection of the sets Q^{TR} and Q^{DC} onto the space of p-variables. Remember that we always assume in this chapter that the network graph is weakly connected and anti-symmetric.

Definition 2.21
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph. Denote by

$$\begin{aligned} Q_p^{\mathrm{TR}}(G) &:= \mathrm{proj}_p(Q^{\mathrm{TR}}(V, E, b, \infty, \bar{f})) \\ &= \left\{ p \in \mathbb{R}^V \;\middle|\; \exists f \in \mathbb{R}^V : \begin{array}{r} p + Af = 0 \\ f^- \le f \le f^+ \end{array} \right\} \\ Q_p^{\mathrm{DC}}(G) &:= \mathrm{proj}_p(Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f})) \\ &= \left\{ p \in \mathbb{R}^V \;\middle|\; \exists f \in \mathbb{R}^E, \varphi \in \mathbb{R}^V : \begin{array}{r} B^\top \varphi = f \\ p + Af = 0 \\ f^- \le f \le f^+ \end{array} \right\} \end{aligned}$$

the set of TR-feasible *power injections* and the set of DC-feasible *power injections*, respectively. If G is fixed and clear from the context, we might omit the argument and write Q_p^{TR} and Q_p^{DC}, respectively.

The set of feasible power injections can be interpreted as the set of possible ways to offset production below demand in some regions (negative values) with production above demand in other regions (positive values) over the network. This set is restricted by the capacities of the individual vertices, given by $\bar{p}$, as well as by constraints imposed by the network.

Since we want to study specifically the constraints imposed by the network, we only consider network graphs with $\bar{p} = \infty$ in this section (and in the above definition). This corresponds to ignoring the box constraints $p^- \le p \le p^+$, which appear in both $Q^{\mathrm{TR}}(G)$ and $Q^{\mathrm{DC}}(G)$, in order to isolate those restrictions imposed by the network itself. Note that the set of power injections that are actually feasible under the capacity restrictions can be easily recovered as $Q_p^{\mathrm{TR}}(G) \cap [p^-, p^+]$ and $Q_p^{\mathrm{DC}}(G) \cap [p^-, p^+]$, respectively.

As mentioned above, there is a 1-to-1-correspondence of vertices between $Q_p^{\mathrm{DC}}(G)$ and $Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f}) \cap \{(p, f, \varphi) \mid \varphi_{v_0} = 0\}$:

Lemma 2.22
Let $(V, E, b, \bar{p}, \bar{f})$ be a network graph and $v_0 \in V$. Then there exist f, φ such that (p, f, φ) is an extremal point in $Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f}) \cap \{(p, f, \varphi) \mid \varphi_{v_0} = 0\}$ if and only if p is an extremal point in $Q_p^{\mathrm{DC}}(G)$.

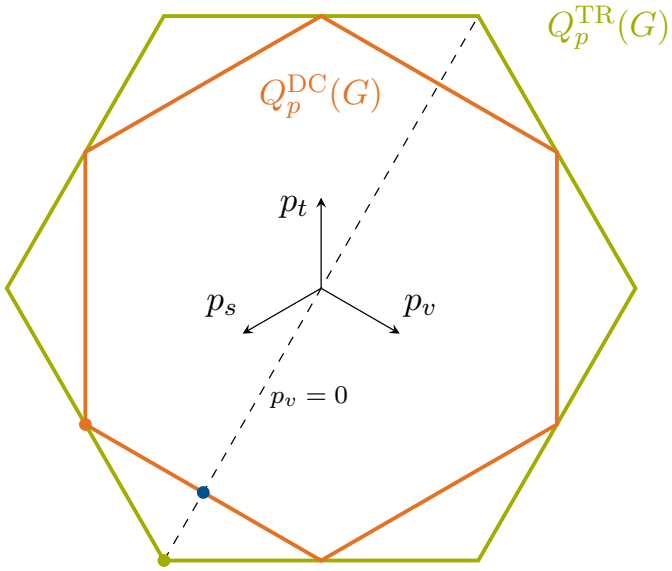


Figure 2.10: The sets $Q_p^{\mathrm{DC}}(G)$ and $Q_p^{\mathrm{TR}}(G)$ for the triangular network from Fig. 2.1. While both sets are actually (2-dimensional) subsets of $\mathbb{R}^3$, the viewing angle was chosen such that their common affine hull (defined by $p_s + p_t + p_v = 0$) coincides with the drawing plane. The point shown in orange corresponds to the maximal power injection possible in the vertex s (restricted by the sum of capacities on edges incident with s). Within $Q_p^{\mathrm{DC}}(G)$, the only feasible such injection vector is the one where both v and t receive half of the injection, saturating the edges connecting them with s and leaving the edge (v, t) without any flow. In $Q_p^{\mathrm{TR}}(G)$, on the other hand, all distributions between v and t are feasible, e. g. the point shown in green, where the entire injection is transferred to t (shown in Fig. 2.1a). The configuration shown in Fig. 2.1b is represented by the blue point.

Proof. Let $(p, f, \varphi) \in Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f}) \cap \{(p, f, \varphi) \mid \varphi_{v_0} = 0\}$ be extremal. Suppose for contradiction that p is not extremal in Q_p^{DC}. Then, there exist $p^1, p^2 \in Q_p^{\mathrm{DC}}$ with $p^1 \neq p^2$ such that $p = \lambda p^1 + (1 - \lambda)p^2$ with $\lambda \in (0, 1)$. By definition of Q_p^{DC}, this means that there exist potentials $\varphi^1, \varphi^2 \in \mathbb{R}^V$ such that $p^i = -AB^\top \varphi^i$ for $i \in \{1, 2\}$. By Proposition 2.20, since $\mathbb{1} \in \ker(B^\top)$, these potentials can be chosen such that $\varphi^1_{v_0} = \varphi^2_{v_0} = 0$.

Furthermore, since $AB^\top(\lambda\varphi^1 + (1 - \lambda)\varphi^2 = -\lambda p^1 + (1 - \lambda)p^2 = -p = AB^\top\varphi$, we have that $\lambda\varphi^1 + (1 - \lambda)\varphi^2 - \varphi \in \ker(AB^\top)$. By Proposition 2.20 and since $\varphi^1_{v_0} = \varphi^2_{v_0} = 0 = \varphi_{v_0}$, this means that $\lambda\varphi^1 + (1 - \lambda)\varphi^2 = \varphi$. Hence, $(p, f, \varphi) = \lambda(-AB^\top\varphi^1, B^\top\varphi^1, \varphi^1) + (1 - \lambda)(-AB^\top\varphi^2, B^\top\varphi^2, \varphi^2)$ and $-AB^\top\varphi^1 = p^1 \neq p^2 = -AB^\top\varphi^2$, a contradiction to extremality of (p, f, φ).

Conversely, let $p \in Q_p^{\mathrm{DC}}$ be extremal. By definition of Q_p^{DC}, this means that there exists a potential $\varphi \in \mathbb{R}^V$ such that $p = -AB^\top\varphi$. By Proposition 2.20, since $\mathbb{1} \in \ker(B^\top)$, this potentials can be chosen such that $\varphi_{v_0} = 0$. Then, $(p, B^\top\varphi, \varphi) \in Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f}) \cap \{(p, f, \varphi) \mid \varphi_{v_0} = 0\}$.

Now suppose that there exist $(p^1, f^1, \varphi^1), (p^2, f^2, \varphi^2) \in Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f}) \cap \{(p, f, \varphi) | \varphi_{v_0} = 0\}$ with $(p^1, f^1, \varphi^1) \neq (p^2, f^2, \varphi^2)$ and $\lambda(p^1, f^1, \varphi^1) + (1 - \lambda)(p^2, f^2, \varphi^2) = (p, B^\top\varphi, \varphi)$. But then, $p^1 \neq p^2$: Otherwise, if $p^1 = p^2$ then $\varphi^1 - \varphi^2 \in \ker(AB^\top)$ and since $\varphi^1_{v_0} = \varphi^2_{v_0} = 0$ we obtain by Proposition 2.20 that $\varphi^1 = \varphi^2$ and thus $f^1 = f^2$, a contradiction with $(p^1, f^1, \varphi^1) \neq (p^2, f^2, \varphi^2)$. This, however, implies that $p = \lambda p^1 + (1 - \lambda)p^2$ with $p^1, p^2 \in Q_p^{\mathrm{DC}}$ and $p^1 \neq p^2$, a contradiction with extremality of p. □

A corresponding statement with respect to the projection on the space of f-variables can be obtained along the same lines.

The set of feasible power injections plays an important role in the context of Optimal Power Flow (OPF) problems: Given an initial operating point, a network operator might want to know, by how much she can alter the injection in some node (e. g. by increasing production in some power plant) without threatening the stability of the network. A planner might ask if the network can cope with the additional injection of power generated by a new power plant in a particular node. Correspondingly, there exists a significant thread of work on the *feasible injection region* of Optimal Power Flow (OPF) problems:

Most literature is concerned with the more complex feasibility region of the AC model because of its higher practical relevance, given that most transmission lines in fact operate using alternating current. While the DC model is used in the context of AC power flow problems, for instance as a local linearization in the context of the common Decoupled Load Flow method (see Ilić [Ili92]), it is inaccurate as a global solution method in many important cases (see, e. g., the comparison of AC and DC model by Stott, Jardim, and Alsaç [SJA09]).

One particular area of interest is the question of convexity of the AC injection region. A comprehensive overview of the related literature is given in a two-part review article by Low [Low14a; Low14b]. Lesieutre and Hiskens [LH05] investigate the convexity of the feasible injection region and review some examples to show that the feasible injection region is in general not convex, but "close to convex".

This result motivates a major line of work that considers convex relaxations of Optimal Power Flow problems using the AC model. While such relaxations have been known since at least 2006 (see, e. g., [Jab06; Bai+08]), Lavaei and Low [LL12] provide some answers in a seminal paper to the question of when these relaxations are tight. They start from the observation that the Optimal Power Flow problem using the AC model, while non-convex in general, is *weakly dual* (in the sense that it can be bounded from below) to a semidefinite program. Whenever the duality gap between

both problems is zero, then the Optimal Power Flow problem can hence be solved optimally in polynomial time, even using the AC model. The authors provide necessary and sufficient criteria for this (which can be checked *a priori*) and perform case studies to demonstrate that the duality gap is zero in many established benchmark networks.[3]

Sojoudi and Lavaei [SL12] and Bose et al. [Bos+11] further investigate network topologies for which the *a priori* conditions mentioned above always hold. These include acyclic networks (called *radial* by Sojoudi and Lavaei) and those where controllable phase-shifters are installed in all edges outside of a spanning tree. Consequently, most positive results in this field are limited to acyclic networks (e. g., [LTZ14; TCL15]) and networks consisting of cycles and trees [ZT13].

In the context of the network models from Section 2.1.2, the literature listed above can be understood in the following way: On network topologies with certain nice properties (e. g., acyclicity), the AC model is theoretically no more complex than the DC model. However, two important caveats remain: First, in the context of large-scale energy system optimization there is good reason to consider even semidefinite (or second-order cone) optimization problems as too difficult to solve in practice. This is true regardless of the underlying network topology and becomes even more significant if the model is extended by additional constraints. Secondly, Transmission Capacity Expansion Planning (TCEP) introduces a new kind of nonlinearity (analogously to the case of the DC model mentioned in Section 2.2) that cannot (currently) be handled by a convex relaxation according to [SL14].

In light of the above, the aim of this section is twofold: Since the DC-OPF is obviously solvable in polynomial time (see Section 2.2), we go a step further. We want to obtain a deeper understanding of the structure of solutions to the DC-OPF and how they compare with solutions to the TR-OPF. Secondly, since no efficient approach is currently known to solve the DC-TCEP in general, we try to characterize DC-feasible points in a way that can provide us with some information about the required production and transmission capacities. This approach could provide a basis for a similar technique aiming at the Transmission Capacity Expansion Problems using the AC model, characterizing the solutions of convex relaxations as mentioned above in a similar fashion.

[3] The authors have subsequently generalized these results to a special type of quadratically constrained optimization problems for which the coefficients can be interpreted with respect to an underlying *generalized weighted graph* in which the signs of edge weights obey a particular constraint for each cycle in the graph [SL14].

2.4.1 Injection Regions of Transport Model and DC Model

We begin by making some simple observations about the polyhedra Q_p^{DC} and Q_p^{TR}:

Proposition 2.23
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph with $f^- \leq 0 \leq f^+$. Then,

a) $0 \in Q_p^{\mathrm{DC}}(G) \subset Q_p^{\mathrm{TR}}(G)$,

b) if (V, E) is a tree then $Q_p^{\mathrm{DC}}(G) = Q_p^{\mathrm{TR}}(G)$.

Furthermore, let $k \geq 0$, $G' = (V, E, k \cdot b, \bar{p}, \bar{f})$ and $G'' = (V, E, b, \bar{p}, k \cdot \bar{f})$. For any $X \in \{\mathrm{DC}, \mathrm{TR}\}$ it holds that

c) $Q_p^X(G') = Q_p^X(G)$

d) $Q_p^X(G'') = k \cdot Q_p^X(G)$.

Parts c) and d) can be rephrased as follows: The polyhedra Q_p^{DC} and Q_p^{TR} are invariant under scaling of the elasticity vector b. On the other hand, they scale linearly with the capacity vector $\bar{f}$. Note in particular that this partially answers the question about the implications of ignoring the interdependence of elasticity and edge capacities: As long as we scale all capacities uniformly, it does not matter whether we keep the elasticity vector constant or scale it along with the capacity vector.

In the following, we will consider different subclasses of network graphs for which a statement can be made with respect to the quality of $Q_p^{\mathrm{TR}}(G)$ as an approximation of $Q_p^{\mathrm{DC}}(G)$. We start by showing that for the most general case, $Q_p^{\mathrm{TR}}(G)$ is not a good approximation.

Theorem 2.24
There is no constant k such that $Q_p^{\mathrm{TR}}(G) \subset k \cdot Q_p^{\mathrm{DC}}(G)$ for all network graphs G.

Proof. Consider the network graph $G = (V, E, b, \bar{p}, \bar{f})$ with (V, E) as shown in Fig. 2.11, $b = \mathbb{1}$, $\bar{f} = (-\mathbb{1}, \mathbb{1})$, and $\bar{p} \in (\mathbb{R} \cup \{-\infty\})^V \times (\mathbb{R} \cup \{\infty\})^V$ chosen arbitrarily. Let $p^* \in \mathbb{R}^V$ with $p_s^* = n + 1$, $p_t^* = -(n + 1)$ and $p_v^* = 0$ for all $v \in V \setminus \{s, t\}$. Note that there exists $f^* \in \mathbb{R}^E$ such that (p^*, f^*) is TR-feasible and hence $p^* \in Q_p^{\mathrm{TR}}(G)$.

On the other hand, let $\bar{f}' := k \cdot \bar{f} = (-k \cdot \mathbb{1}, k \cdot \mathbb{1})$ and $G' = (V, E, b, \bar{p}, \bar{f}')$. Let $p \in Q_p^{\mathrm{DC}}(G')$ and let f, φ be corresponding vector of flows and vertex angles such that $(p, f, \varphi) \in Q^{\mathrm{DC}}(G')$. Note that, due to $p - Af = 0$ and since $p_v = 0$ for all $v \in V \setminus \{s, t\}$, it must hold for every $i \in [n]$ that $f_{sv_{1i}} = f_{v_{1i}v_{2i}} = \cdots = f_{v_{(n-1)i}v_{ni}} = f_{v_{ni}t}$ and hence $\varphi_{v_{1i}} - \varphi_s = \varphi_{v_{2i}} - \varphi_{v_{1i}} = \cdots = \varphi_{v_{ni}} - \varphi_{v_{(n-1)i}} = \varphi_t - \varphi_{v_{ni}}$. This implies that $\varphi_t - \varphi_s = (n + 1) \cdot \varphi_t - \varphi_{v_{ni}}$ and since $\varphi_t - \varphi_s = f_{st} \leq f'^+_{st} = k$, it follows that

$$\sum_{i=1}^{n} f_{v_{in}t} + f_{st} = \left(\sum_{i=1}^{n} (\varphi_t - \varphi_{v_{ni}}) + (\varphi_t - \varphi_s) \right) = \left(\frac{n}{n+1} + 1 \right) \cdot (\varphi_t - \varphi_s) < 2k.$$

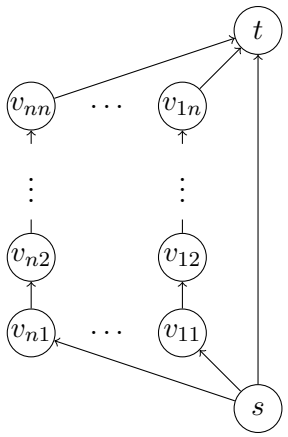


Figure 2.11: There is no k such that $Q_p^{\mathrm{TR}}(G) \subset k \cdot Q_p^{\mathrm{DC}}(G)$ for all network graphs with the structure depicted, with $n \to \infty$.

For every $p \in Q_p^{\mathrm{DC}}(G')$, it thus holds that $p_t^* = -(\sum_{i=1}^n f_{v_i n t} + f_{st}) > -2k$. In particular, $p^* \notin Q_p^{\mathrm{DC}}(G') = k \cdot Q_p^{\mathrm{DC}}(G)$ if $n \geq 2k$. □

This result seems to suggest that $Q_p^{\mathrm{DC}}(G)$ is "much smaller" than $Q_p^{\mathrm{TR}}(G)$. We will see however, that this is not really true. The following lemma provides us with a method to investigate this question.

Lemma 2.25
Let $(V, E, b, \bar{p}, \bar{f})$ be a network graph with $f^- = -f^+$. For a subset $I \subset E$, let flow values f_e^* be given for all edges $e \in I$. Let $S := \bigcup_{(v,w)\in E}\{(v,w),(w,v)\}$ be a new set of edges which includes reverse edges for all edges in E and for $(w,v) \in S \setminus E$ define $b_{wv} = b_{vw}$ as well as $f_{wv}^+ := f_{vw}^+$ and let $I^{\mathrm{rev}} := \{(w,v) \mid (v,w) \in I\} \subset S$ be the set of reverse edges for edges in I.

Consider the directed graph (V, S) with edge weights d given by

$$d_e = \begin{cases} \frac{f_e^*}{b_e} & e \in I \\ -\frac{f_e^*}{b_e} & e \in I^{\mathrm{rev}} \\ \frac{f_e^+}{b_e} & e \in S \setminus (I \cup I^{\mathrm{rev}}). \end{cases}$$

If (V, S, d) contains no directed cycle of negative weight, then $Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f})$ contains a point (p, f, φ) with $f_e = f_e^*$ for all $e \in I$.

Proof. We prove the contrapositive of the above statement. Assume that the set $Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f})$ does not contain a point (p, f, φ) with $f_e = f_e^*$ for all $e \in I$.

Then, the polyhedron

$$\left\{ (p, f, \varphi) \in \mathbb{R}^V \times \mathbb{R}^E \times \mathbb{R}^V \;\middle|\; \begin{array}{rl} b_{vw}(\varphi_w - \varphi_v) = f^*_{vw} & \forall\ (v,w) \in I \\ b_{vw}(\varphi_w - \varphi_v) \leq f^+_{vw} & \forall\ (v,w) \in E \setminus I \\ -b_{vw}(\varphi_w - \varphi_v) \leq -f^-_{vw} = f^+_{vw} & \forall\ (v,w) \in E \setminus I \end{array} \right\}$$

is empty. From Farkas' Lemma (e. g., the variant from [Roc70, Theorem 22.1] where the coefficients corresponding to $b_{vw}(\varphi_w - \varphi_v) \leq f^*_{vw}$ and $b_{vw}(\varphi_w - \varphi_v) \geq f^*_{vw}$ for all $(v,w) \in I$ are combined into a single coefficient without sign restriction), we obtain the existence of a coefficient vector $\nu \in \mathbb{R}^I$ (corresponding to the equality constraint) as well as two vectors $\nu^+_e, \nu^-_e \in \mathbb{R}^{E\setminus I}_{\geq 0}$ such that the corresponding linear combination of left-hand sides is 0, whereas the right-hand sides evaluate to a value strictly less than 0.

In other words,

$$\begin{aligned} 0 &= \sum_{e \in \delta^{\text{in}}(r) \cap I} b_e \nu_e + \sum_{e \in \delta^{\text{in}}(r) \cap (E \setminus I)} b_e(\nu^+_e - \nu^-_e) - \sum_{e \in \delta^{\text{out}}(r) \cap I} b_e \nu_e - \sum_{e \in \delta^{\text{out}}(r) \cap (E \setminus I)} b_e(\nu^+_e - \nu^-_e) \\ 0 &> \sum_{e \in I} \nu_e f^*_e + \sum_{e \in E \setminus I} (\nu^+_e f^+_e + \nu^-_e f^+_e) > \sum_{e \in I} \nu_e f^*_e + \sum_{e \in E \setminus I} (|\nu^+_e - \nu^-_e| \cdot f^+_e). \end{aligned}$$

Writing $\nu'_e := b_e \nu_e$ for $e \in I$ and $\nu'_e := b_e(\nu^+_e - \nu^-_e)$ for $e \notin I$, we obtain

$$\sum_{e \in \delta^{\text{in}}(r)} \nu'_e - \sum_{e \in \delta^{\text{out}}(r)} \nu'_e = 0 \tag{2.48}$$

$$\sum_{e \in I} \nu'_e \frac{f^*_e}{b^*_e} + \sum_{e \in E \setminus I} |\nu'_e| \frac{f^+_e}{b^*_e} < 0. \tag{2.49}$$

Assigning negative values of ν' to the corresponding reverse edges, we obtain a proper (non-negative) flow in the network (V, S, d) of weight as given in (2.49). By (2.48), the flow is a circulation. However, a circulation can only have negative weight if it contains at least one cycle with negative weight, which proves the lemma. □

Note that in the graph that is constructed in the above lemma, the arcs from I^{rev} are the only ones that have negative weight. The set of cycles, that might potentially have negative weight is quite restricted, especially if the set I is small.

As an example, we can use Lemma 2.25 to prove, under reasonable assumptions for a network graph G derived from an energy network, that despite Theorem 2.24, $Q^{\text{DC}}_p(G)$ is not even strictly contained in $Q^{\text{TR}}_p(G)$, i. e., $Q^{\text{TR}}_p(G)$ and $Q^{\text{DC}}_p(G)$ share some points on their respective boundary. This means that it is not possible to obtain a tighter approximation of $Q^{\text{DC}}_p(G)$ on the basis of $Q^{\text{TR}}_p(G)$ simply by scaling the polyhedron with an appropriate factor or, in terms of our example from Fig. 2.10, that points like the one shown in orange always exist:

Theorem 2.26
Let $G = (V, E, \bar{p}, \bar{f})$ be a network graph with $f^- = -f^+$. For any $(s,t) \in E$ and undirected s-t-path $\mathcal{P}$, let

$$\frac{f^+_{st}}{b_{st}} \leq \sum_{e \in \mathcal{P}} \frac{f^+_e}{b_e}. \tag{2.50}$$

Let $v \in V$ and define $p^v_{\max} = \sum_{e \in \delta(v)} f^+_e$. Then, there exists an injection vector p^v satisfying

$$p^v_v = p^v_{\max}$$

such that $\pm p^v$ lies on the boundary of both $Q^{\mathrm{TR}}_p(G)$ and $Q^{\mathrm{DC}}_p(G)$.

The inequality (2.50) can be thought of as a triangle inequality for the completed graph with respect to the edge weight f^+/b (in the case where (V, E) is a complete graph, it is equivalent to the triangle inequality). Looking at this from the perspective of power networks, recall from Remark 2.19 that the inverse of the edge weight is the absolute value of each line's *susceptance per unit of capacity*. If we assume that all material-related parameters are the same for all lines, then by (2.2) the inequality (2.50) is equivalent to

$$\mathrm{length}_{st} \cdot \mathrm{pfactor}(\mathrm{length}_{st}) \leq \sum_{e \in \mathcal{P}} \mathrm{length}_e \cdot \mathrm{pfactor}(\mathrm{length}_e). \tag{2.51}$$

As $\mathrm{pfactor}(x)$ is non-decreasing in x, inequality (2.51) is implied in particular by the ordinary triangle inequality with respect to the length of lines:

$$\begin{aligned} \mathrm{length}_{st} \cdot \mathrm{pfactor}(\mathrm{length}_{st}) &\leq \sum_{e \in \mathcal{P}} \mathrm{length}_e \cdot \mathrm{pfactor}\left(\sum_{e \in \mathcal{P}} \mathrm{length}_e \right) \\ &\leq \sum_{e \in \mathcal{P}} \mathrm{length}_e \cdot \mathrm{pfactor}(\mathrm{length}_e) \end{aligned}$$

The inequality (2.50) hence also follows from the ordinary triangle inequality with respect to the length of lines, which makes it a very reasonable assumption.

Proof (Proof of Theorem 2.26). We first show that a point p^v with the desired properties exists in $Q^{\mathrm{DC}}_p(G)$. To see this, we use Lemma 2.25 to obtain a point $(p, f, \varphi) \in Q^{\mathrm{DC}}(V, E, b, \infty, \bar{f})$ where p has the desired properties, which by the definition of $Q^{\mathrm{DC}}_p(G)$ means that $p^v := p \in Q^{\mathrm{DC}}_p(G)$.

Let $f^*_e = f^+_e$ for all $e \in \delta^{\mathrm{out}}(v)$ and $f^*_e = f^-_e = -f^+_e$ for all $e \in \delta^{\mathrm{in}}(v)$. Now, let $I := \delta^{\mathrm{out}}(v) \cup \delta^{\mathrm{in}}(v)$ and consider the graph $G^* := (V, S, d)$ as defined in Lemma 2.25. Note that the only arcs with negative weight are those ending at v. Furthermore, any (simple) directed cycle in G^* can only contain one such arc.

Suppose therefore that G^* contains a negative cycle that includes the arc (v, u). Then, all other edges on that cycle form a directed u-v-path P^* and they all have positive weight. But negativity of the cycle implies that

$$-d_{vu} > \sum_{e \in P^*} d_e.$$

Let $P := (P^* \cap E) \cup \{(w, w') \mid (w', w) \in P^* \setminus E\}$ denote the subset of edges of G from which the edges in P^* are generated (by the definition in Lemma 2.25). Then,

$$\frac{f^+_{vu}}{b_{vu}} = \frac{f^*_{vu}}{b_{vu}} = -d_{vu} > \sum_{e \in P^*} d_e = \sum_{e \in P} \frac{f^+_e}{b_e},$$

which contradicts our assumption from (2.50).

The set $Q^{\mathrm{DC}}_p(V, E, b, \infty, \bar{f})$ thus contains a point (p, f, φ) with $f_e = f^+_e$ for all $e \in \delta^{\mathrm{out}}(v)$ and $f_e = f^-_e = -f^+_e$ for all $e \in \delta^{\mathrm{in}}(v)$, which in particular implies that $p_v = \sum_{e \in \delta(v)} f^+_e$.

Let $p^v := p \in Q^{\mathrm{DC}}_p(G)$ and note that, by an identical argument, $-p^v \in Q^{\mathrm{DC}}_p(G)$. By Proposition 2.23, it thus holds that $\pm p^v \in Q^{\mathrm{TR}}_p(G)$. To see that p^v lies on the boundary of $Q^{\mathrm{TR}}_p(G)$, observe that for all $(p, f) \in Q^{\mathrm{TR}}_p(G)$ it holds that $p = -Af$ and $f^- \leq f \leq f^+$. In particular, this implies that

$$p_v = -(Af)_v = -\sum_{e \in \delta^{\mathrm{in}}(v)} f_e + \sum_{e \in \delta^{\mathrm{out}}(v)} f_e \leq \sum_{e \in \delta(v)} f^+_e.$$

As this inequality is satisfied with equality by the point p^v, it actually supports $Q^{\mathrm{TR}}_p(G)$ and p^v lies on the boundary of $Q^{\mathrm{TR}}_p(G)$. Again, the argument for the point $-p^v$ is identical, which concludes the proof. □

We conclude this section with a rather simple observation that allow us to bound the factor by which a TR-feasible solution may differ from a DC-feasible solution. While Theorem 2.24 showed that there exists no constant such factor for *all* networks, we will obtain an upper bound on the factor, depending on the network structure. In the following, let $\rho := \min_{e \in E} \min\{|f^-_e|, |f^+_e|\}$ denote the minimum edge capacity in the network.

Theorem 2.27
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph with $f^- \leq 0 \leq f^+$ and let $p \in \mathbb{R}^V$. If there exists a partition $S \dot\cup T = V$ of the set of vertices such that $p_v \geq 0$ for all $v \in S$, $p_v \leq 0$ for all $v \in T$, and $\sum_{v \in S} p_v = -\sum_{v \in T} p_v \leq \rho$ then $p \in Q^{\mathrm{DC}}_p(G)$.

Proof. Note that, by Proposition 2.20, p induces a unique (but not necessarily feasible) flow f via $p = AB^\top \varphi$ and $f = B^\top \varphi$. We can see that f is indeed feasible by the

following argument: If $|f_e| > \rho$ for any edge e, then a decomposition of f into flows along paths from S to T and flows along cycles yields that f must contain a flow along a cycle. However, f is induced by the potential φ and hence cannot contain a positive flow along any cycle. Therefore, $|f_e| \leq \rho \leq \min\{f_e^-, f_e^+\}$ for all $e \in E$ and f is indeed feasible. □

This immediately implies a more general bound on the necessary scaling factor k such that $k \cdot Q_p^{\mathrm{DC}}(G)$ contains $Q_p^{\mathrm{TR}}(G)$: We define the operator $\mathrm{cut}(S)$ for a network graph $(V, E, b, \bar{p}, \bar{f})$ and a subset $S \subset V$ of vertices by $\mathrm{cut}(S) := \sum_{v \in V \setminus S} \sum_{w \in N^{\mathrm{in}}(v) \cap S} |f_{wv}^+| + \sum_{w \in \delta^{\mathrm{out}}(v) \cap S} |f_{vw}^-|$.

Corollary 2.28
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph and for any $p \in Q_p^{\mathrm{TR}}(G)$, let $V^{in} := \{v \in V \mid p_v > 0\}$ and $V^{out} := \{v \in V \mid p_v < 0\}$. Let $k := 1/\rho \cdot \min_{V^{in} \subset S \subset V \setminus V^{out}} \mathrm{cut}(S)$. Then, $p \in k \cdot Q_p^{\mathrm{DC}}(G)$.

Proof. From summing up the balance equations in all vertices, it has to hold for any $S \subset V$ that $\sum_{v \in S} p_v = -\sum_{v \in V \setminus S} p_v$. Furthermore, if $V^{\mathrm{in}} \subset S \subset V \setminus V^{\mathrm{out}}$, then $|\sum_{v \in S} p_v| = |\sum_{v \in V \setminus S} p_v| \leq \mathrm{cut}(S)$. This implies in particular that $|\sum_{v \in S} p_v| = |\sum_{v \in V \setminus S} p_v| \leq \rho \cdot k$. With $\lambda^* := 1/k$, it follows that $\sum_{v \in S} \lambda^* p_v = -\sum_{v \in V \setminus S} \lambda^* p_v \leq \rho$. By Theorem 2.27, we can conclude that $\lambda^* \cdot p \in Q_p^{\mathrm{DC}}(G)$ and hence $p \in k \cdot Q_p^{\mathrm{DC}}(G)$.□

Note that for single-source-single-sink s-t-flows, we can furthermore bound the term $\min_{V^{\mathrm{in}} \subset S \subset V \setminus V^{\mathrm{out}}} \mathrm{cut}(S)$ by $\max_{e \in \delta(s) \cup \delta(t)} \max\{|f_e^-|, |f_e^+|\}$, which yields a very simple (but weaker) bound for this case.

Looking back at the example from Theorem 2.24 where all edge capacities were 1, we see that every s-t-cut has size at least $n + 1$. By Corollary 2.28, $p \in k \cdot Q_p^{\mathrm{DC}}(G)$ whenever $k \geq n + 1$ (the single-source-single-sink bound above is identical in this case). However, we also see that the bound is not tight: Since it only takes into account the capacities of edges and not the elasticity values, it must assume the worst case which would be that all s-t-paths except $\{(s, t)\}$ carry practically no flow for any point in $Q^{\mathrm{DC}}(G)$.

2.5 The Differential Flow Polytope

In this section, we take the opposite approach to that from Section 2.4 and consider the following projections of the sets Q^{TR} and Q^{DC} onto the space of f-variables:

$$
\begin{aligned}
Q_f^{\mathrm{TR}}(G) :=& \left\{ f \in \mathbb{R}^E \;\middle|\; \begin{array}{r} p + Af = 0 \\ \exists p \in \mathbb{R}^V : f^- \leq f \leq f^+ \\ p^- \leq p \leq p^+ \end{array} \right\} \\
=& \left\{ f \in \mathbb{R}^E \;\middle|\; \begin{array}{c} p^- \leq -Af \leq p^+ \\ f^- \leq f \leq f^+ \end{array} \right\} \\
Q_f^{\mathrm{DC}}(G) :=& \left\{ f \in \mathbb{R}^E \;\middle|\; \begin{array}{r} B^\top \varphi = f \\ \exists p \in \mathbb{R}^V, \varphi \in \mathbb{R}^V : p + Af = 0 \\ f^- \leq f \leq f^+ \\ p^- \leq p \leq p^+ \end{array} \right\} \\
=& \left\{ f \in \mathbb{R}^E \;\middle|\; \begin{array}{r} B^\top \varphi = f \\ \exists \varphi \in \mathbb{R}^V : p^- \leq -Af \leq p^+ \\ f^- \leq f \leq f^+ \end{array} \right\}
\end{aligned}
$$

The above polyhedra bear a close resemblance to the *feasible flow* and *feasible differential* polyhedra that are well-known from the theory of network flows. In this section, major parts of which are currently being prepared for publication as [BS19a], we investigate this similarity in more detail. Rockafellar [Roc84] gives an extensive account of the underlying theory of flows and differentials, based on which we present the following overview.

2.5.1 Optimal Flows and Differentials

Given a directed graph $G = (V, E)$ with incidence matrix A, the (un-capacitated) *feasible flow* polyhedron (for some right-hand-side vector d satisfying $\mathbb{1}^\top d = 0$) can be written as

$$Q_F(G) := \big\{ x \in \mathbb{R}^E \;\big|\; Ax = d \big\}.$$

Note that, while $Q_F(G)$ does not restrict the flow by any edge capacities, we could add capacity constraints without changing the properties that we will discuss below, albeit at the cost of a somewhat more cumbersome notation. Furthermore x is often restricted to lie in the positive orthant, but in line with our assumption that edges in a network graph can also carry negative flow, we do not impose that restriction here. We can, however, split up the vector x into a positive part x^+ and a negative part x^- (corresponding to positive flow in the direction of an edge and in the direction of the reverse edge, respectively).

For instance, we might want to treat positive and negative flows on an edge separately, say by assigning a (possibly different) positive cost $f_e^+ \geq 0$ and $-f_e^- \geq 0$ to positive

and negative flows, respectively. Splitting up the vector x as described above, we obtain the optimal flow problem

$$\min\Big\{ {f^+}^\top x^+ - {f^-}^\top x^- \;\Big|\; x^+, x^- \geq 0, A(x^+ - x^-) = d \Big\}. \tag{2.52}$$

Observe that since $f_e^+, -f_e^- \geq 0$, we can assume without loss of generality that for every edge e, at most one of x_e^+ and x_e^- is strictly positive.

The feasible flow polyhedron (written in the form of (2.52)) is famous for a very neat characterization of its combinatorial structure: A basis of a vertex of this polyhedron is characterized by the set of components of x^+ and x^- for which the sign constraints are binding, e. g. that have value 0. It is convenient now to differentiate between the set of edges e for which both x_e^+ and x_e^- are 0 (call this set $\mathcal{N}$) and its complement, the set of edges e for which at least one of x_e^+ and x_e^- may assume a value strictly greater than 0 (call this set $\mathcal{B} := E \setminus \mathcal{N}$).

By Proposition 2.20, it holds that $\operatorname{rank}(A) = |V| - 1$. For the pair $(\mathcal{B}, \mathcal{N})$ to specify a basis (i. e., for every d to uniquely identify the vector $x := x^+ - x^-$ satisfying $Ax = d, x_{\mathcal{N}} = 0$), it is therefore necessary that $\mathcal{B}$ contains $|V| - 1$ edges corresponding to linearly independent columns of A (in particular, there might be additional edges e for which both x_e^+ and x_e^- are 0, that are not contained in $\mathcal{B}$).

The following characterization now holds with respect to these bases: $(\mathcal{B}, \mathcal{N})$ specifies a basis if and only if the edges of G corresponding to those columns of A indexed by $\mathcal{B}$ form a spanning tree of G. This can be seen as follows: The columns for any undirected cycle in G are always linearly dependent (as the column of any edge can be obtained as the sum of positive/negative columns of the other edges), hence the sets of $|V| - 1$ linearly independent columns are exactly those corresponding to spanning trees.

A similar characterization holds for the optimization problem dual to (2.52), the optimal differential problem. It asks for a potential φ maximizing the linear function $d^\top \varphi$ over the polyhedron $Q_D(G)$ of *feasible differentials* (see [Roc84]) defined as follows:

$$Q_D(G) := \Big\{ \varphi \in \mathbb{R}^V \;\Big|\; f^- \leq A^\top \varphi \leq f^+ \Big\}$$

The variables in this new polyhedron do not have to satisfy any (individual) bounds or sign restrictions. A basis of this polyhedron is hence characterized by the set of components of the vector $A^\top \varphi$ for which one of their respective capacity bounds is binding. Dual to the notation above, we can call the set of edges that index these rows $\mathcal{B}$ and analogously $\mathcal{N} := E \setminus \mathcal{B}$. Again, by Proposition 2.20, $\operatorname{rank}(A) = |V| - 1$ and hence in order to uniquely identify (up to translation along $\mathbb{1}$) the vector φ conforming to the specifications of $(\mathcal{B}, \mathcal{N})$, we need that $\mathcal{B}$ indexes $|V| - 1$ linearly independent rows of $A^\top$ (or equivalently columns of A). As argued above, a set $\mathcal{B}$ satisfies this requirement if and only if the corresponding edges form a tree. Furthermore, the potential φ uniquely determined by $(\mathcal{B}, \mathcal{N})$ (up to translation along $\mathbb{1}$) is feasible if $f_{\mathcal{N}}^- \leq A_{\mathcal{N}}^\top \varphi \leq f_{\mathcal{N}}^+$.

	feasible flows	feasible differentials
feasible set	$Q_F(G)$	$Q_D(G)$
solution conforms to $(\mathcal{B},\mathcal{N})$ if…	$x_\mathcal{N} := x^+_\mathcal{N} - x^-_\mathcal{N} = 0$	all entries of $A^\top_\mathcal{B}\varphi$ at upper/lower bound
$(\mathcal{B},\mathcal{N})$ is a basis if…	$\mathcal{B}$ indexes a spanning tree	$\mathcal{B}$ indexes a spanning tree
or (equivalently) if…	$\mathcal{N}$ is a maximal co-forest	$\mathcal{N}$ is a maximal co-forest
for a non-extremal solution…	$\{e \mid x_e > 0\}$ contains cycle	$\{e \mid (A^\top\varphi)_e \in (f^-_e, f^+_e)\}$ contains a cut
for every extremal solution…	$\{e \mid x_e > 0\}$ forms a forest	$\{e \mid (A^\top\varphi)_e \in (f^-_e, f^+_e)\}$ forms co-forest

Table 2.2: Characterizing properties of bases and extremal solutions for the polyhedra of feasible flows and feasible differentials. A co-forest is a set of edges that does not contain a cut, i. e., its removal does not increase the number of connected components.

Note that the roles of $\mathcal{B}$ and $\mathcal{N}$ in both cases are reversed: While for the feasible flow polyhedron, the entries of x indexed by $\mathcal{B}$ are the ones that are free to take non-zero values bounds, the inequalities indexed by $\mathcal{B}$ in the feasible differential polyhedron are the ones that are forced to be binding. A summary of the characterizations outlined above is provided in Table 2.2.

2.5.2 Feasible Differential Flows

The connections of the above theory of flows and differentials to the polyhedra Q_f^{TR} and Q_f^{DC} defined above are easy to make: The feasible flow polyhedron $Q_F(G)$ (possibly with added edge capacities) is a special case of the polyhedron $Q_f^{\mathrm{TR}}(G)$ for the case where $p^- = p^+ = -d$. On the other hand, $Q_D(G)$ is (up to a full-rank linear transformation) a special case of the polyhedron $Q_f^{\mathrm{DC}}(G)$ where $p^+ = -p^- = \infty$.

This raises the following question, which we will try to answer in this section: Can a similar characterization of extremal points of Q_f^{DC} be recovered, given that Q_f^{DC} imposes both constraints $B^\top\varphi = f$ and $p^- \leq -Af \leq p^+$? Since the points in Q_f^{DC} combine the defining properties of flows and differentials, we refer to them as *differential flows*.

Definition 2.29 (Differential Flow)
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph. We call $f \in \mathbb{R}^E$ a *differential flow* in G if there exists a potential $\varphi \in \mathbb{R}^V$ such that for $(v, w) \in E : f_{vw} = b_{vw} \cdot (\varphi_w - \varphi_v)$. In this case, we say that φ *induces* f. Furthermore, we say that f is *feasible* (in the network graph G) if

a) $f^-_e \leq f_e \leq f^+_e$ for all $e \in E$, and

b) $p^-_v \leq \sum_{e\in\delta^{\mathrm{out}}(v)} f_e - \sum_{e\in\delta^{\mathrm{in}}(v)} f_e \leq p^+_v$ for all $v \in V$.

As intended, the set of feasible differential flows is exactly the polyhedron $Q_f^{\mathrm{DC}}(G)$. We refer to the first set of constraints $B^\top\varphi = f$ as *differential constraints* and to the

other two sets of constraints $f^- \leq f \leq f^+$ and $p^- \leq -Af \leq p^+$ as *edge constraints* (or *edge capacity constraints*) and *vertex constraints* (or *relaxed flow conservation constraints*), respectively. Note that the vertex constraints represent upper and lower bounds on the flow that *is created* in the respective vertex, the upper bound p_v^+ thus represents a *lower* bound on the *excess* in the vertex v as defined, e. g., in [AMO93, Ch. 7.6].

A differential flow is thus an ordinary flow on the edges of the graph (V, E) with relaxed flow conservation constraints, edge capacity constraints and the additional *differential constraints* $f_{vw} = b_{vw} \cdot (\varphi_w - \varphi_v)$ for a suitable choice of the potential φ. Conversely, the notion of a feasible differential flow generalizes that of a feasible differential (see, e. g., [Roc84]), which is recovered if $-p^- = p^+ = \infty$ and $b_e = \mathbb{1}$.

Note further that if $p^- = p^+ = 0$, then $f \equiv 0$ is the only feasible differential flow (if $0 \in [f^-, f^+]$): In this case, b) is an ordinary flow conservation constraint. In particular, it is well-known that all flows satisfying condition b) with $p^- = p^+ = 0$ must be sums of flows along cycles. On the other hand, the differential constraints require that no cycle can carry a nonzero flow (otherwise, φ would increase strictly along the cycle).

In this section, we want to understand the combinatorial structure of the polyhedron $Q_f^{\mathrm{DC}}(G)$. Given the close similarity, as already mentioned, to the feasible flow and feasible differential polyhedra Q_F and Q_D, one might be tempted to try and express the set of feasible differential flows as feasible flows or differentials on a modified graph.

However, we will see that this does not work: In order to replace the relaxed flow conservation constraints by strict flow conservation constraints, which are more common in network-flow-related settings, we can enforce strict flow conservation for every vertex in the original graph and replace the relaxed flow conservation constraints by edge capacity constraints on a new artificial edge, connecting each vertex to some artificial source, which would supply the excess flow which is required to satisfy flow conservation. We can think of two ways to do this:

a) All artificial edges are connected to the same artificial source s. In this case, since the artificial edge (s, v) carries flow p_v and $p - Af = 0$, we know that (strict) flow conservation also holds in s. However, we have already observed above that in the case of $p^- = p^+ = 0$, the only feasible differential flow is $f \equiv 0$. We would thus have to relax some other constraint, e. g. the differential constraint at least in the artificial source vertex s.

b) Each artificial edge is connected to its own artificial source. In this case, we can use arbitrary values of b for the artificial edges, but we cannot enforce flow conservation in the artificial sources. We thus end up with an only seemingly more restrictive version, where strict flow conservation holds for all $i \in V$ and no flow conservation at all holds for the artificial vertices.

These observations emphasize the idea that the differential flow polyhedron jointly

generalizes different notions from the theory of network flows and somehow lies *in between* the concepts of flows and differentials. This further motivates the question for a suitable characterization of extremal points of Q_f^{DC}, which was raised at the beginning of this section.

2.5.3 α-forests and α-trees

In the following, we prove that under certain conditions, such a characterization is indeed possible: Using a suitable notion of acyclicity, extremal points of $Q_f^{\mathrm{DC}}(G)$ can be identified with maximally acyclic collections of edges *and vertices* wich, again, correspond to the set of active inequalities. To motivate the following definition, consider the following observations, which will be proven in a more general context in Theorems 2.36 and 2.41 in this section:

- If $-p^- = p^+ = \infty$, then the problem reduces to the original feasible differential problem (as mentioned above). Extremal solutions of this problem can be characterized by spanning trees for which edge capacity constraints are binding for every edge in the tree (see Table 2.2).
- If in a solution f the edge capacity constraints are active in all edges from a spanning tree *except one*, and additionally the relaxed flow conservation constraint is binding in one of the endpoints of that missing edge, then f is extremal.
- If $-f^- = f^+ = \infty$, then Proposition 2.20 implies that in order to uniquely determine a solution, the relaxed flow conservation constraint has to be binding in exactly $|V| - 1$ vertices.
- If in a solution f an edge capacity constraint is active on an edge (v, w), then f is uniquely determined by any $|V| - 2$ binding relaxed flow conservation constraints that do not contain both v and w.

These observations motivate the following definition of a structure that can be used to characterize extremal solutions.

Definition 2.30 (α-Forest)
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph. A pair $F = (E_F, V_F)$ consisting of a set of edges $E_F \subset E$ and a set of vertices $V_F \subset V$ in G is called an *α-forest* in G if there exists an injective function $\alpha_F : V_F \to E \setminus E_F$, mapping each vertex from V_F to a neighboring edge which is not already in E_F, such that the set $E_F \cup \alpha_F(V_F)$ does not contain a undirected cycle. We call such an α a *vertex map* for F. The *size* of F is defined by $|F| := |E_F| + |V_F|$.

If there is no α-forest F' with $|F'| > |F|$, then we call F *maximal*. Since we assume (for reasons of simplicity) that G is weakly connected, a maximal α-forest is called

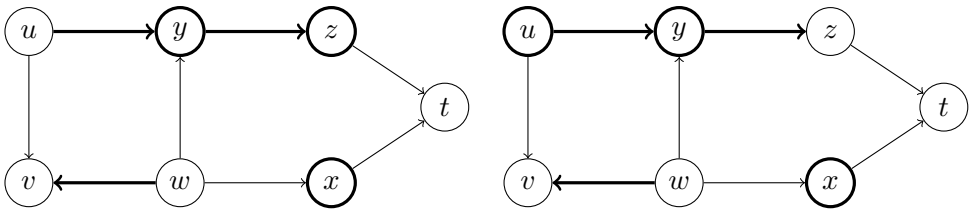


Figure 2.12: We visualize α-forests by marking the sets of active vertices and edges in bold. The structure on the left is an α-forest, as we can choose, e. g., $\alpha_F(y) = (w, y)$, $\alpha_F(x) = (w, x)$ and $\alpha_F(z) = (z, t)$ (in fact, since $|E_F| + |V_F| = |V| - 1$, the α-forest is maximal and hence an α-tree). The structure on the right, in contrast, is not an α-forest: Since α_F must map to $E \setminus E_F$, we must have $\alpha_F(y) = (w, y)$ and $\alpha_F(u) = (u, v)$, closing an undirected cycle.

an α-tree. Furthermore, we say for an edge $e \in E$ that e is *active in* F if $e \in E_F$. Analogously, for a vertex v, we say that v is *active in* F if $v \in V_F$.

Given a feasible differential flow $f \in Q_f^{\mathrm{DC}}(G)$, we say that an α-forest F *conforms with* f if

a) for all $v \in V_F$ it holds that $(-Af)_v \in \{p_v^-, p_v^+\}$

b) for all $e \in E_F$, it holds that $f_e \in \{f_e^-, f_e^+\}$.

By the definition above, α-forests generalize ordinary forests: Choosing $V_F = \emptyset$, $F = (E_F, V_F)$ is an α-forest if and only if the edges in E_F form a forest. While an α-forest consists of a set of edges and a set of vertices, these generally do not form a graph: E_F can contain edges between vertices which are not in V_F (see Fig. 2.12). The vertices in V_F thus may or may not be incident with edges in E_F, the only requirement being that they can be mapped in an injective way by α_F to a neighboring edge, which is not in E_F and does not close an undirected cycle. In particular, this implies for any α-tree in a network graph $(V, E, b, \bar{p}, \bar{f})$ that $|E_F| + |V_F| \leq |V| - 1$. Finally, note that for a given α-forest F, the function α_F need not be unique.

Note that for the special case of $-p^- = p^+ = \infty$, we have already observed that $Q_f^{\mathrm{DC}}(G) = Q_D(G)$ (with a suitable scaling of f^- and f^+), i. e., a feasible differential flow is just a feasible differential. As one would expect, an α-forest conforming with a flow f in this case is simply a forest consisting of edges that are at their capacity limits. We already know from Section 2.5.1 that a point $f := A^\top \varphi$ is extremal in $Q_D(G)$ if and only if there exists a maximal forest of this characteristic (of course, there might be different forests characterizing the same solution).

We will prove in this section that α-trees can be used to characterize extremal solutions in $Q_f^{\mathrm{DC}}(G)$ for all but some very special instances. We start by discussing which subsets of constraints generally characterize extremal solutions in $Q_f^{\mathrm{DC}}(G)$.

For a selection $\mathcal{B}$ of rows from a matrix A, we write $A_{\mathcal{B}}$ for the sub-matrix consisting of these rows. A point $f \in Q_f^{\mathrm{DC}}$ is now uniquely determined by a selection $\mathcal{B}$ of rows from the matrix $(I, A)^\top$ (corresponding to constraints of type a) or b) in Definition 2.29 satisfied with equality) if and only if the matrix

$$\begin{pmatrix} B^\top & -I \\ 0 & \begin{pmatrix} I \\ -A \end{pmatrix}_{\mathcal{B}} \end{pmatrix}$$

has maximal rank, i. e., if any vector x satisfying

$$\begin{pmatrix} B^\top & -I \\ 0 & \begin{pmatrix} I \\ -A \end{pmatrix}_{\mathcal{B}} \end{pmatrix} x = 0$$

is of the form $x = (\varphi, 0)$ with $\varphi := \lambda \cdot \mathbb{1} \in \mathbb{R}^V$.

Another way to look at this is the following: Recall from Proposition 2.20 that $\operatorname{rank}(B) = |V| - 1$ and hence we can alternatively consider the polyhedron Q_φ^{DC} from (2.47). Then, an extremal point in $Q_f^{\mathrm{DC}}(G)$ is uniquely determined by a subset of active constraints from

$$f^- \leq B^\top \varphi \leq f^+$$
$$p^- \leq -AB^\top \varphi \leq p^+$$

such that the matrix of normal vectors has rank $|V| - 1$, or equivalently (since the signs are irrelevant) by a selection of rows from the matrix $(B^\top, AB^\top)^\top$ such that the corresponding sub-matrix has rank $|V| - 1$.

Definition 2.31
We say that a network graph $G = (V, E, b, \bar{p}, \bar{f})$ is *(α-tree) non-degenerate* if a point $f \in Q_f^{\mathrm{DC}}(G)$ is extremal (i. e., the matrix of active edge and vertex constraints has rank $|V| - 1$) if and only if there exists an α-tree in G that conforms with f. Otherwise, G is *(α-tree) degenerate*.

In order to avoid unnecessarily complicated language, we will simply call a network graph *degenerate* (or *non-degenerate*) if this is unambiguous. In a non-degenerate network graph, we can thus identify every extremal solution with a (not necessarily unique) selection of edges and vertices that form an α-tree. Conversely, every solution for which such a selection of edges and vertices exists must be extremal.

Our examples above of network graphs for which the differential flow polyhedron coincides with the ordinary differential polyhedron show that some network graphs are indeed non-degenerate. On the other hand, the following example shows that degenerate network graphs do exist as well:

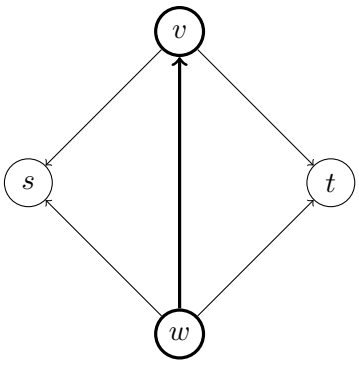


Figure 2.13: If $\frac{b_{wt}}{b_{ws}} = \frac{b_{vt}}{b_{vs}}$ then the shown graph is α-tree degenerate: The α-tree indicated by the active vertices and edges marked in bold conforms with a solution which is not extremal.

Example 2.32
Consider the network graph G in Fig. 2.13 with suitable edge and vertex bounds such that there exists $f \in Q_f^{\mathrm{DC}}(G)$ for which the thick edges and vertices are those where a corresponding constraint is binding. Note that there exists an α-tree F which conforms with f (for example, $F = (\{(w,v)\}, \{v,w\})$ with $\alpha_F(v) = (v,s)$, $\alpha_F(w) = (w,t)$). However, f need not be extremal in $Q_f^{\mathrm{DC}}(G)$. More precisely, f is extremal if and only if $\frac{b_{wt}}{b_{ws}} \neq \frac{b_{vt}}{b_{vs}}$, as can be seen by studying the corresponding matrix of active constraints (rows are indexed by the edge or vertex corresponding to the respective constraint, columns are indexed by the corresponding vertex):

$$\begin{array}{c} \\ (w,v) \\ v \\ w \end{array} \begin{array}{c} \begin{array}{cccc} s & v & w & t \end{array} \\ \begin{pmatrix} 0 & b_{wv} & -b_{wv} & 0 \\ -b_{vs} & b_{vs} + b_{vt} - b_{wv} & b_{wv} & -b_{vt} \\ -b_{ws} & -b_{wv} & b_{ws} + b_{wt} - b_{wv} & -b_{wt} \end{pmatrix} \end{array} \tag{2.53}$$

If $\frac{b_{wt}}{b_{ws}} = \frac{b_{vt}}{b_{vs}}$, then the columns corresponding to vertices s and t are linearly dependent.

If on the other hand a network graph is non-degenerate for all values of $\bar{p}$ and $\bar{f}$, then this implies a more general relation between α-forests and faces of $Q_f^{\mathrm{DC}}(G)$:

Lemma 2.33
Let (V,E) be a weakly connected directed graph with elasticity vector b such that for every $\bar{f}, \bar{p}$ the network graph $G = (V,E,b,\bar{f},\bar{p})$ is non-degenerate. Then, the following are equivalent:

a) the matrix of active constraints in any point $f \in Q_f^{\mathrm{DC}}(G)$ has rank k

b) there exists an α-forest F of size k in G that conforms with f

Proof. Let $f \in Q_f^{\mathrm{DC}}(G)$ such that the matrix M of active constraints has rank k. Choose a vertex set V^* of size $|V^*| = |V| - 1 - k$ such that the matrix

$$M' := \begin{pmatrix} M \\ A_{V^*} \end{pmatrix}$$

has rank $|V| - 1$. Now, consider the network graph G' derived from G by modifying $\bar{p}$ in such a way that for each $v \in V^*$, we have that $-A_v f \in \{p_v^-, p_v^+\}$. As G' is non-degenerate, there exists an α-tree F' in G' which conforms with f. Let F be defined as follows:

$$F := (E_{F'}, V_{F'} \setminus V^*)$$

Then, F is an α-forest in G which conforms with f. Furthermore, $|V^*| = |V| - 1 - k$ and hence $|F| \geq |F'| - (|V| - 1 - k) = k$ which proves the statement.

On the other hand, let $f \in Q_f^{\mathrm{DC}}(G)$ and let F be an α-forest of size k that conforms with f. F can be extended to an α-tree by adding a set E^* of at most $(|V| - 1 - k)$ edges. Consider the graph G'' derived from G by modifying $\bar{f}$ in such a way that for each $e \in E^*$ we have that $f \in \{f_e^-, f_e^+\}$. This adds at most $(|V| - 1 - k)$ additional constraints to the set of constraints which are active in f. As G'' is non-degenerate, f is an extremal point in $Q_f^{\mathrm{DC}}(G'')$. Hence the set of active constraints has full rank $|V| - 1$, which means that the original set of active inequalities must have rank at least k. □

Note that if, in line with our interpretation of electricity networks as network graphs (Remark 2.19), the elasticity vector changes in response to changes in edge capacities, then the flow vector f in the above proof might no longer be feasible. This can be remedied by replacing the auxiliary graph G'' in the proof of Lemma 2.33 with a network graph with modified elasticity vector, such that f remains feasible in combination with the modified edge capacities. If this network graph is also non-degenerate, the same argument can be used to prove Lemma 2.33. This is true in particular in the case where every network graph on (V, E) is non-degenerate, independently of $\bar{f}$, $\bar{p}$ and b. Network graphs with this property will be characterized in detail in Section 2.5.6 below.

2.5.4 Existence of α-trees

Before we delve into the question of when exactly an α-tree conforming with a solution certifies that this solution is extremal, we first settle the inverse question: In this section, we prove that for every extremal solution $f \in Q_f^{\mathrm{DC}}(G)$, there exists an α-tree in G that conforms with f.

Lemma 2.34
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph. Let $V^* := \{v_1, v_2, \ldots, v_k\} \subset V$ denote a set of k vertices and let $E^* \subset E$ denote a set of $|V| - 1 - k$ edges such that

$$\operatorname{rank} \begin{pmatrix} (B^\top)_{E^*} \\ (AB^\top)_{V^*} \end{pmatrix} = |V| - 1. \tag{2.54}$$

Then, the graph $G^* := (V, E^*)$ has $k+1$ weakly connected components $S_1, \ldots, S_{k+1} \subset V$. Furthermore, there exists a $k \times (k+1)$-dimensional matrix $C = (\gamma_{ij})_{\substack{i \in [k] \\ j \in [k+1]}}$ with

$$\gamma_{ij} \begin{cases} > 0 & \text{if } v_i \in S_j \\ < 0 & \text{if } v_i \notin S_j \text{ and } N_G(v_i) \cap S_j \neq \emptyset \\ 0 & \text{else.} \end{cases}$$

such that $\operatorname{rank}(C) = k$ *and* $\mathbb{1} \in \ker(C)$.

Proof. The rank condition (2.54) implies in particular that those rows of $B^\top$ that correspond to the edges in E^* are linearly independent. Since B is (up to linear scaling of columns) the incidence matrix of the graph (V, E), this implies that E^* cannot contain an undirected cycle and hence the graph $G^* = (V, E^*)$ consists of $k + 1$ weakly connected components. Let

$$M := \begin{pmatrix} (B^\top)_{E^*} \\ (AB^\top)_{V^*} \end{pmatrix}. \tag{2.55}$$

We assume w. l. o. g. that the columns of M are ordered in such a way that the columns corresponding to vertices in S_1 come first, followed by the columns corresponding to vertices in S_2 etc. We now scale every row of $(B^\top)_{E^*}$ in such a way that its two entries are 1 and -1 (i. e. we divide the row corresponding to the edge e by b_e). We obtain a matrix of the following structure:

$$M' = \left(\begin{array}{cccc} (A^\top)_{S_1} & 0 & \cdots & 0 \\ 0 & (A^\top)_{S_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (A^\top)_{S_{k+1}} \\ \hline \multicolumn{4}{c}{M^*} \end{array} \right). \tag{2.56}$$

Obviously, we have that $\operatorname{rank}(M) = \operatorname{rank}(M')$ and each of the blocks $(A^\top)_{S_j}$ is the transpose of the incidence matrix of the weakly connected component S_j of G^*, which is an undirected tree spanning the vertices in S_j.

The matrix M^*, on the other hand, is a selection of rows from the nodal admittance matrix $AB^\top$, with each row corresponding to a vertex $v_i \in V^*$. For every $v_i \in V^*, w \in V$, it therefore holds that

$$M^*_{v_i w} = \begin{cases} \sum_{e \in \delta(v_i)} b_e & w = v_i \\ -b_{v_i w} & (v_i, w) \in E \\ -b_{w v_i} & (w, v_i) \in E \\ 0 & \text{else.} \end{cases} \tag{2.57}$$

In particular, $M^*_{v_i v_i} > 0$, $M^*_{v_i w} \leq 0$ for all $w \neq v_i$ and $\sum_{w \in V} M^*_{v_i w} = 0$.

We now select for each component of G^* a representative vertex $w_j \in S_j$. Observe that adding a multiple λ of the row of $(A^\top)$ which corresponds to the edge $(v, w) \in E$ to a row m of M^* subtracts λ from the entry M^*_{mv} and adds λ to the entry M^*_{mw}. Since every S_j is weakly connected, we can move along the edges in S_j, adding a suitable linear combination of rows of $(A^\top)_{S_j}$ to each row m of M^*, thereby eliminating all entries of M^* from columns corresponding to vertices in $S_j \setminus w_j$ and adding the corresponding values to the entry $M^*_{m w_j}$. We obtain a matrix M^{**} such that

$$\operatorname{rank}(M) = \operatorname{rank}(M') = \operatorname{rank} \left(\begin{array}{cccc} (A^\top)_{S_1} & 0 & \cdots & 0 \\ 0 & (A^\top)_{S_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (A^\top)_{S_{k+1}} \\ \hline \multicolumn{4}{c}{M^{**}} \end{array} \right) \tag{2.58}$$

where M^{**} has the form

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_k \end{array} \begin{array}{c} \begin{array}{cccccccccccc} w_1 & & & & w_2 & & & & \cdots & & & w_{k+1} \end{array} \\ \left(\begin{array}{cccccccccccc} \gamma_{11} & 0 & \cdots & 0 & \gamma_{12} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \gamma_{1(k+1)} \\ \gamma_{21} & 0 & \cdots & 0 & \gamma_{22} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \gamma_{2(k+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ \gamma_{k1} & 0 & \cdots & 0 & \gamma_{k2} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \gamma_{k(k+1)} \end{array} \right). \end{array} \tag{2.59}$$

Since the row-sums of all rows of M' were 0, the same holds true for M^{**}, i. e. $\sum_{j \in [p]} \gamma_{ij} = 0$ for all $i \in [k]$. Furthermore, $\gamma_{ij} = \sum_{w \in S_j} M^*_{v_i, w} \leq 0$ for all S_j that do not contain v_i and $\gamma_{ij} = 0$ for all S_j that contain neither v_i itself nor a neighbor of v_i. This implies that, if $v_i \in S_{j^*}$, then $\gamma_{ij^*} > 0$: Otherwise, all entries of the row $M^{**}_{v_i}$ would be 0, a contradiction with M^{**} having full row rank. We can now drop all zero columns from M^{**} to obtain the matrix C with the desired properties. □

The following lemma will serve to capture the relation between the different weakly connected components of G^* and active vertices within these components. It makes

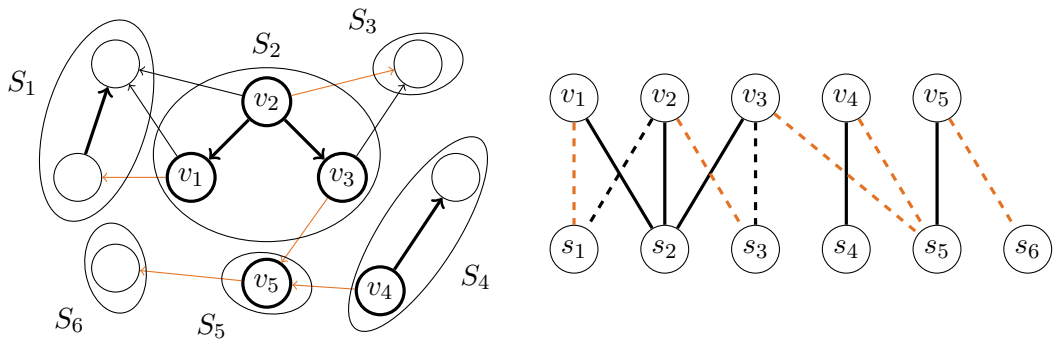


Figure 2.14: On the left, a network graph with the sets V^* and E^* shown in bold, as well as the partition of vertices induces by the weakly connected components of the graph $G^* = (V, E^*)$. On the right, the corresponding bipartite graph $H := (V \dot\cup S, E)$ satisfying the conditions from Lemma 2.35. Solid edges are those from the set R, dashed edges are those from the set U. In orange, a selection of edges U^* with the properties guaranteed by Lemma 2.35 is shown. On the left, a possible translation into a function α (see Theorem 2.36) is indicated by the orange edges (α maps every active vertex to a neighboring orange edge).

a general statement about a bipartite graph $H := (V \dot\cup S, E)$ with certain properties. Subsequently in Theorem 2.36, we will use the partition classes V and S to represent the set of *active* vertices V^* in a network graph and the set of components connected by *active* edges in E^* (the weakly connected components of the graph G^*, see Theorem 2.36), respectively. A selection of edges with the properties specified in the lemma will allow us to derive a function α to prove that (E^*, V^*) is indeed an α-tree (see Fig. 2.14).

Lemma 2.35
Let $H := (V \dot\cup S, E)$ be a bipartite graph with $|S| = |V| + 1$ and $E = R \dot\cup U$ a partition of E satisfying the following conditions:

a) $\delta(v) \cap R = 1$ for all $v \in V$

b) for every $V' \subset V$, it holds that $|N(V')| \geq |V'| + 1$.

Then, there exists a selection of edges $U^* \subset U$ such that

i) $\delta(v) \cap U^* = 1$ for all $v \in V$ and

ii) the graph $(V \dot\cup S, R \dot\cup U^*)$ is connected (and therefore a tree).

Proof. We first observe that b), together with a), implies in particular that $\delta(v) \cap U \geq 1$ for all $v \in V$. Thus, since H is bipartite, a selection $U^* \subset U$ of edges with $\delta(v) \cap U^* = 1$ always exists. Furthermore, $|U| \geq |V|$.

We prove the statement by induction over $|U|$. If $|U| = |V|$, then every vertex $v \in V$ is incident with exactly one edge from U. The selection $U^* := U$ thus satisfies i). If we suppose that ii) does not hold, i. e. $(V \dot\cup S, R \dot\cup U)$ is not connected, then we can choose a connected component C and denote by V_C the subset of vertices from V that is covered by C. Since C is a connected component, $N(V_C) \cap N(V \setminus V_C) = \emptyset$ and from b) we obtain $|N(V_C)| \geq |V_C| + 1$. Hence,

$$|N(V \setminus V_C)| \leq |S \setminus N(V_C)| = |S| - |N(V_C)| \leq |V| + 1 - |V_C| - 1 = |V \setminus V_C|,$$

a contradiction to b).

Now, let us assume that $|U| > |V|$, which means that there exists $v \in V$ with $\delta(v) \cap U \geq 2$. Choose two distinct edges $e_1, e_2 \in \delta(v) \cap U$ and denote the corresponding neighbors of v by s_1 and s_2, respectively.

We claim that at least one of the two graphs $H_1 := (V \dot\cup S, E \setminus \{e_1\})$ and $H_2 := (V \dot\cup S, E \setminus \{e_2\})$ must satisfy the condition b). By induction, we then obtain a selection U^* of edges which satisfies i) and ii) with respect to the graph H_1 (or H_2) and thus also for H, proving the statement.

In the following, for any subset $V' \subset V$, we denote by $N(V')$ (without index) the neighborhood of the set V' in the graph H and by $N_{H_1}(V')$ and $N_{H_2}(V')$ the neighborhoods of the set V' in the graphs H_1 and H_2, respectively.

For contradiction, suppose that our claim is false. Then, there exists a set $V_1 \subset V$ which violates b) in the graph H_1 and a set $V_2 \subset V$ which violates b) within H_2. Since neither V_1 nor V_2 violate b) within H, it must hold that $v \in V_1 \cap V_2$.

Now, let $V_1' := V_1 \setminus V_2$, $V_2' := V_2 \setminus V_1$, and $V_{12} := V_1 \cap V_2 \setminus \{v\}$. We use the following statements, which we successively derive from each other below:

I) $|N_{H_1}(V_1)| = |V_1| = |N(V_1)| - 1$ and
$|N_{H_2}(V_2)| = |V_2| = |N(V_2)| - 1$

II) $|N(V_1') \cap N(V_{12})| \geq |N(V_1')| + |N(V_{12})| - |V_1'| - |V_{12}| - 1$ and
$|N(V_2') \cap N(V_{12})| \geq |N(V_2')| + |N(V_{12})| - |V_2'| - |V_{12}| - 1$

III) $|N(V_1' \dot\cup V_{12} \dot\cup V_2')| \leq |V_1'| + |V_{12}| + |V_2'| + 1$

IV) $N(v) \setminus \{s_1\} \subset N(V_1 \setminus \{v\})$ and
$N(v) \setminus \{s_2\} \subset N(V_2 \setminus \{v\})$

Statements III) and IV) together will then prove that the original graph H would already violate b), a contradiction.

To prove I) and II), let $i \in \{1, 2\}$. From the assumption that V_i violates b) in the graph H_i, but satisfies b) in the graph H, we can immediately follow that $|N_{H_i}(V_i)| \leq |V_i| \leq |N(V_i)| - 1$. On the other hand, H_i lacks only a single edge compared with H and thus $|N_{H_i}(V_i)| \geq |N(V_i)| - 1$. This proves I).

We can now conclude that

$$\begin{aligned}|V_i'| + |V_{12}| + 1 = |V_i| &\overset{I)}{\geq} |N_{H_i}(V_i)| \geq |N_{H_i}(V_i' \dot\cup V_{12})| = |N(V_i' \dot\cup V_{12})| \\ &= |N(V_i')| + |N(V_{12})| - |N(V_i') \cap N(V_{12})|,\end{aligned}$$

which proves II). Using the inclusion/exclusion principle, it now follows that

$$\begin{aligned}|N(V_1' \dot\cup V_{12} \dot\cup V_2')| &= |N(V_1')| + |N(V_{12})| + |N(V_2')| \\ &\quad - |N(V_1') \cap N(V_{12})| - |N(V_2') \cap N(V_{12})| \\ &\quad \underbrace{-|N(V_1') \cap N(V_2')| + |N(V_1') \cap N(V_2') \cap N(V_{12})|}_{\leq 0} \\ &\overset{II)}{\leq} |N(V_1')| + |N(V_{12})| + |N(V_2')| \\ &\quad - (|N(V_1')| + |N(V_{12})| - |V_1'| - |V_{12}| - 1) \\ &\quad - (|N(V_2')| + |N(V_{12})| - |V_2'| - |V_{12}| - 1) \\ &= -|N(V_{12})| + |V_1'| + |V_{12}| + |V_2'| + |V_{12}| + 2 \\ &\overset{b)}{\leq} |V_1'| + |V_{12}| + |V_2'| + 1,\end{aligned}$$

which proves III). On the other hand, to prove IV) let again $i \in \{1, 2\}$ and observe that $s_i \notin N(V_i \setminus \{v\})$, since otherwise $|N_{H_i}(V_i)| = |N(V_i)|$, a contradiction with I). Furthermore,

$$\begin{aligned}|N(V_i) \setminus N(V_i \setminus \{v\})| &= |N(V_i)| - |N(V_i \setminus \{v\})| \\ &\overset{I),b)}{\leq} (|N_{H_i}(V_i)| + 1) - (|V_i \setminus \{v\}| + 1) \\ &\overset{I)}{\leq} |V_i| + 1 - (|V_i| - 1 + 1) = 1,\end{aligned}$$

and hence $N(v) \setminus \{s_i\} \subset N(V_i \setminus \{v\})$, which proves IV).

In particular, since $s_1, s_2 \in N(v)$, this means that $s_2 \in N(V_1 \setminus \{v\})$ and $s_1 \in N(V_2 \setminus \{v\})$. But this implies that

$$N(v) \overset{IV)}{\subset} N(V_1 \setminus \{v\}) \cup N(V_2 \setminus \{v\}) = N((V_1 \cup V_2) \setminus \{v\}) = N(V_1' \dot\cup V_{12} \dot\cup V_2'),$$

which leads to

$$\begin{aligned}|N(V_1' \dot\cup V_{12} \dot\cup V_2' \dot\cup \{v\})| &= |N(V_1' \dot\cup V_{12} \dot\cup V_2')| \\ &\overset{III)}{=} |V_1'| + |V_{12}| + |V_2'| + 1 = |V_1' \dot\cup V_{12} \dot\cup V_2' \dot\cup \{v\}|,\end{aligned}$$

a contradiction with b).

It follows that our assumption was false and that indeed, at least one of the graphs H_1 or H_2 satisfies b). As that graph obviously satisfies a) (the set R remains unchanged), as well, the statement follows by induction. □

Observe that the core of the above proof is that we can remove edges from H in a way that the condition b) remains satisfied. This argument could thus be used to prove other statements similar to Lemma 2.35 about the existence of edge selections with certain properties in cases where

- a selection remains valid if we add additional edges, and
- if the number of edges is sufficiently small, then the existence of a corresponding edge selection is obvious from b).

Using the two previous lemmas, we can now prove the following theorem:

Theorem 2.36
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph and $f \in Q_f^{\mathrm{DC}}(G)$. If f is extremal, then there exists an α-tree in G that conforms with f.

Proof. Let $f \in Q_f^{\mathrm{DC}}(G)$ be extremal. From Proposition 2.20, we obtain that there exists a vector $\varphi \in \mathbb{R}^V$ with $f = B^\top \varphi$ and $|V| - 1$ linearly independent inequalities of (2.47) that are active in φ (uniquely determining φ up to the linearlity space $\mathbb{R} \cdot \mathbb{1}$ contained in Q_φ^{DC}). Let E^* and $V^* := \{v_1, v_2, \ldots, v_k\}$ be the set of edges and vertices, respectively, corresponding to these constraints and let

$$M := \begin{pmatrix} (B^\top)_{E^*} \\ (AB^\top)_{V^*} \end{pmatrix}. \tag{2.60}$$

Then M is $(|V| - 1) \times |V|$-dimensional and has $\operatorname{rank}(M) = |V| - 1$, thus satisfying the conditions of Lemma 2.34. Hence, the graph (V, E^*) consists of $k + 1$ (weakly) connected components $S_1, S_2, \ldots, S_{k+1} \subset V$ and there exists a $k \times (k+1)$-dimensional matrix $C = (\gamma_{ij})_{\substack{i \in [k] \\ j \in [p]}}$ with $\operatorname{rank}(C) = k$ such that $C \cdot \mathbb{1} = 0$ and

$$\gamma_{ij} \begin{cases} > 0 & \text{if } v_i \in S_j \\ < 0 & \text{if } v_i \notin S_j \text{ and } N(v_i) \cap S_j \neq \emptyset \\ 0 & \text{else.} \end{cases} \tag{2.61}$$

Now, let $\mathcal{S} := \{S_1, S_2, \ldots, S_{k+1}\}$, $R := \{\{v_i, s_j\} \mid \gamma_{ij} > 0\}$, and $U := \{\{v_i, s_j\} \mid \gamma_{ij} < 0\}$. For the graph $H := (V^* \dot\cup \mathcal{S}, R \dot\cup U)$, the following holds:

- H is bipartite with $|\mathcal{S}| = |V^*| + 1$ and
- $\delta(v) \cap R = 1$ for all $v \in V^*$, since there exists exactly one connected component that contains v.

Furthermore, we claim that for every $V' \subset V^*$, it holds that $|N(V')| \geq |V'| + 1$. To see this, suppose otherwise. Let $\bar{V} \subset V^*$ denote a set such that $|N(\bar{V})| \leq |\bar{V}|$. Reorder the rows and columns of C in such a way that the first rows are those corresponding to vertices in $\bar{V}$ and the first columns are those corresponding to vertices in $N(\bar{V})$. Let V^{**} be a subset of $\bar{V}$ of size $|N(\bar{V})|$ and let C' be the $|V^{**}| \times |V^{**}|$-dimensional submatrix composed of the rows corresponding to vertices in V^{**} and the columns corresponding to vertices in $N(\bar{V})$. Then, C can be written as

$$C = \begin{pmatrix} C' & 0 \\ * & * \end{pmatrix}. \tag{2.62}$$

Since $C \cdot \mathbb{1} = 0$, it follows that $C' \cdot \mathbb{1} = 0$ and hence $\operatorname{rank}(C') < |V^{**}|$ which implies that $\operatorname{rank}(C) < k$, a contradiction. This proves that, indeed, for every $V' \subset V$, it holds that $|N(V')| \geq |V'| + 1$.

The graph H hence satisfies all the requirements of Lemma 2.35 and we obtain a selection of edges $U^* \subset U$ with $\delta(v) \cap U^* = 1$ for every $v \in V^*$ such that the graph $H^* := (V^* \dot\cup \mathcal{S}, R \dot\cup U^*)$ is connected. Since $|\mathcal{S}| = |V| + 1$ and $|R| = |U^*| = |V|$, this implies that H^* is acyclic. For each $v_i \in V^*$, choose j_i such that $\{v_i, s_{j_i}\} \in U^*$ and choose $e_i \in E$ incident with v_i such that $S_{j_i} \cap e_i \neq \emptyset$. Since $v_i \notin S_{j_i}$, it holds that $e_i \notin E^*$. Such an edge exists, since $\{v_i, s_{j_i}\} \in U$ implies that $\gamma_{ij_i} < 0$ which in turn implies that $N(v_i) \cap S_{j_i} \neq \emptyset$.

Now, for all $v_i \in V^*$, let $\alpha(v_i) := e_i$. Then α is injective (otherwise H^* would contain a cycle) and $E^* \cup \alpha(V^*)$ does not contain an undirected cycle (otherwise, again, H^* would contain a cycle). Hence, $F := (E^*, V^*)$ is an α-forest, which can be seen using the (injective) vertex map $\alpha_F := \alpha$. Furthermore, F is of size $|E^*| + |V^*| = |V| - 1$, which means that it is in fact an α-tree. □

Note that we see again that the function α_F corresponding to an α-tree F is not necessarily unique: First, the selection U^* of edges, whose existence is guaranteed by Lemma 2.35, need not be unique. Furthermore, for an edge $\{v_i, s_{j_i}\} \in U^*$ it might be the case that there are more than one edge $e_i \in E$ incident with v_i such that $S_{j_i} \cap e_i \neq \emptyset$ (see the case of v_1 in Fig. 2.14). Each particular choice in both cases leads to a different function α_F for the same selection (E^*, V^*) of active edges and vertices.

2.5.5 Sufficient Conditions

In Example 2.32, we saw that a solution might not be extremal, even though there exists an α-tree that conforms with it. This motivates our interest in criteria under which we can guarantee a correspondence between extremal points in $Q_f^{\mathrm{DC}}(G)$ and α-trees. In this section, we will present some conditions for a given α-tree under which a solution $f \in Q_f^{\mathrm{DC}}$ that conforms with it is guaranteed to be extremal. We start by observing the following:

Remark 2.37
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph and $f \in Q_f^{\mathrm{DC}}(G)$. Then f is *not* extremal if and only if there exists a differential flow $g \neq 0$ such that $f \pm g \in Q_f^{\mathrm{DC}}(G)$. In particular, for every vertex v with $A_v f \in \{p^-, p^+\}$, we must have $A_v g = 0$ and for every edge $(v, w) \in E$ with $f_{vw} \in \{f_{vw}^-, f_{vw}^+\}$, we must have $g_{vw} = 0$.

In light ob the remark above, it makes sense to investigate the restrictions which an α-forest $F = (E_F, V_F)$ that conforms with f imposes on a differential flow g with $f \pm g \in Q_f^{\mathrm{DC}}(G)$. For a potential φ that induces g, the above observation implies that $\varphi_v = \varphi_w$ for every weakly connected component S of the graph (V, E_F) and $v, w \in S$, i. e. φ is constant within any such component.

In order to represent restrictions imposed by F on φ between different connected components of (V, E_F), we need the following definition of a *generalized differential flow* which is induced by a potential φ via a *generalized elasticity vector* b. In contrast to the elasticity vector in a network graph, a generalized elasticity vector is defined for all pairs $(v, w) \in V \times V$ (even if $(v, w) \notin E$). In particular, there may be pairs (v, w) with $b_{vw} \neq b_{wv}$ and one of the two values (or both) might be zero.

Definition 2.38 (Generalized Differential Flow)
Let V be a finite set and $b \in \mathbb{R}_{\geq 0}^{V \times V}$. A potential $\varphi \in \mathbb{R}^V$ induces a *generalized differential flow* $f \in \mathbb{R}^{V \times V}$ on V with respect to b by $f_{vw} = b_{vw}(\varphi_w - \varphi_v)$. We say that f is feasible if

$$\sum_{w \in V \setminus \{v\}} f_{vw} = 0 \tag{2.63}$$

for all vertices $v \in V$.

Note that, in contrast to a differential flow, a generalized differential flow is defined on all pairs of vertices from V. In particular, depending on b, it might be the case that both f_{vw} and f_{wv} are non-zero (or only one or none of them). If both are non-zero, however, then they must have opposite sign (since $b \geq 0$), but need not be each others negative (since we do not require $b_{vw} = b_{wv}$). Furthermore, equation (2.63) is similar to a common flow conservation constraint, but not identical: For every $v \in V$, we only sum over the entries f_{vw} for $w \in V \setminus \{v\}$ (which may be positive or negative), but not over the entries f_{wv}. These only appear in the corresponding constraint (2.63) for vertex w.

Note further that for any $\varphi_0 \in \mathbb{R}$, the trivial potential $\varphi \equiv \varphi_0$ induces the generalized differential flow $f \equiv 0$ (independently of b), which is always feasible. Analogously to Definition 2.30, the following tree structure can be used to characterize the cases where a generalized differential flow is uniquely determined (as we will see in Lemma 2.40):

Definition 2.39
Let $T = (V, E)$ be a weakly connected directed graph that does not contain a directed cycle. If furthermore $|\delta^{\mathrm{out}}(v)| \leq 1$ for all $v \in V$, then T is called an *anti-arborescence*.

In particular, if T is an anti-arborescence, then there is a unique vertex $v_0 \in T$ such that $|\delta^{\text{out}}(v)| = 0$ (the *root* of the anti-arborescence). Furthermore, (V, F) may not contain any *un*directed cycle, either (since any cycle is either directed or $|\delta^{\text{out}}(v)| > 1$ for some v on the cycle). We now show that, as noted above, the existence of an anti-arborescence on the vertices in V such that b is non-zero on its edges means that $f \equiv 0$ is the unique feasible generalized differential flow.

Lemma 2.40
Let V be a finite set, $b \in \mathbb{R}_{\geq 0}^{V \times V}$ and $E := \{(v, w) \in V \times V \mid v \neq w \text{ and } b_{vw} > 0\}$. Let T be a subgraph of (V, E) that is an anti-arborescence. If φ induces a feasible generalized differential flow f on V with respect to b, then $\varphi \equiv \varphi_0$ for some $\varphi_0 \in \mathbb{R}$ (and as a consequence $f \equiv 0$).

Proof. Let $\varphi \in \mathbb{R}^V$ be a potential that induces a feasible generalized differential flow. Let $v_0 \in V$ be the root of the anti-arborescence F. We prove that $\max_{v \in V} \varphi_v = \varphi_{v_0}$, the argument to see that $\min_{v \in V} \varphi_v = \varphi_{v_0}$ is identical. Let $v_1 \in \operatorname{argmax}_{v \in V} \varphi_v$, i. e. $\varphi_{v_1} \geq \varphi_w$ for all $w \in V$. If $\varphi_{v_1} = \varphi_{v_0}$, then we are done. Otherwise, in particular $v_1 \neq v_0$ and there exists an edge $(v_1, v_2) \in T$. Then, by the definition of a feasible generalized differential flow and by maximality of φ_{v_1},

$$0 = \sum_{w \in V \setminus \{v_1\}} \underbrace{\underbrace{b_{v_1 w}}_{\geq 0} \underbrace{(\varphi_w - \varphi_{v_1})}_{\leq 0}}_{\leq 0} \leq b_{v_1 v_2}(\varphi_{v_2} - \varphi_{v_1}).$$

Since furthermore $(v_1, v_2) \in E$, we have that $b_{v_1 v_2} > 0$ which implies that $\varphi_{v_2} - \varphi_{v_1} \geq 0$. By maximality of φ_{v_1} this implies that $\varphi_{v_2} = \varphi_{v_1} > \varphi_{v_0}$ and thus in particular $v_2 \neq v_0$. Hence, $v_2 \in \operatorname{argmax}_{v \in V} \varphi_v$ and we can apply the same argument to find a vertex v_3 with $\varphi_{v_2} = \varphi_{v_3}$. Since T is an anti-arborescence and we only traverse edges of T in the direction towards the root, it holds that $v_3 \neq v_1$. Iterating this argument, we eventually obtain that $\varphi_{v_1} = \varphi_{v_2} = \varphi_{v_3} = \cdots = \varphi_{v_0}$, a contradiction. □

We can now derive a sufficient condition for extremality of a differential flow $f \in Q_f^{\text{DC}}(G)$ based on the existence of an α-tree F with one additional property.

Theorem 2.41
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph, $f \in Q_f^{\text{DC}}(G)$ and $F = (E_F, V_F)$ an α-tree in G that conforms with f. Let F be such that every weakly connected component of the graph (V, E_F) contains at most one vertex which is active in F. Then, f is extremal.

Proof. In order to show that f is extremal, we need to prove that no non-zero differential flow g exists such that $f \pm g \in Q_f^{\text{DC}}(G)$ (see Remark 2.37). Suppose otherwise and let $\varphi \in \mathbb{R}^V$ be a potential that induces g. Denote by $\mathcal{S}$ the set of weakly connected components of the graph (V, E_F).

By our assumption, every component S contains at most one vertex which is active in F. Denote this vertex (if it exists) by v_S. For two components $S, T \in \mathcal{S}$, we define the value b'_{ST} as follows (we write $\delta(S) := \bigcup_{v \in S} \delta(v)$ for any $S \subset V$):

$$b'_{ST} := \begin{cases} \sum_{e \in \delta(v_S) \cap \delta(T)} b_e & \text{if } S \text{ contains a vertex } v_S \text{ active in } F \\ 0 & \text{else.} \end{cases} \tag{2.64}$$

Note that it may be the case that $b'_{ST} \neq b'_{TS}$, in particular one of the two may be 0 (or both), but in any case $b' \geq 0$ since $b > 0$ by the definition of a network graph.

Let α_F be a vertex map for F and denote by F' the set of pairs of components for which there exists an active vertex in the first component which is mapped by α_F to an edge connecting it with the second component:

$$F' := \{(S,T) \in \mathcal{S} \times \mathcal{S} \mid \exists (v,w) \in ((S \times T) \cup (T \times S)) \cap \alpha_F(V_F \cap S)\} \tag{2.65}$$

Then, by the definition of α-trees, the set $E_F \cup \alpha_F(V_F)$ does not contain an undirected cycle and since F is maximal, it spans V. This implies that the graph $(\mathcal{S}, F')$ is weakly connected and does not contain a cycle, either. Furthermore, since every component $S \in \mathcal{S}$ contains at most one active vertex, there is at most one $T \in \mathcal{S}$ such that $(S,T) \in F'$. The graph $(\mathcal{S}, F')$ is thus an anti-arborescence.

Returning to the differential flow g, Remark 2.37 implies for any $v, w \in V$ with $(v,w) \in E_F$ that $g_{vw} = 0$ and hence $\varphi_v = \varphi_w$. Let $\varphi' \in \mathbb{R}^{\mathcal{S}}$ be defined by $\varphi'_S := \varphi_v$ for any $S \in \mathcal{S}$ and $v \in S$ (note that this is well-defined, since $\varphi_v = \varphi_w$ for any $S \in \mathcal{S}$ and $v, w \in S$). Now, observe that φ' induces a generalized differential flow on the set $\mathcal{S}$ with respect to b'. Now, for all $S \in \mathcal{S}$, one of the following two statements holds:

- Either S contains no active vertex and then $b'_{ST} = 0$ for all $T \in \mathcal{S}$ and hence $\sum_{T \in \mathcal{S} \setminus \{S\}} b'_{ST}(\varphi'_T - \varphi'_S) = 0$, or
- S contains exactly one active vertex v_S and

$$\begin{aligned} \sum_{T \in \mathcal{S} \setminus \{S\}} b'_{ST}(\varphi'_T - \varphi'_S) &= \sum_{T \in \mathcal{S} \setminus \{S\}} \sum_{e \in \delta(v_S) \cap \delta(T)} b_e(\varphi'_T - \varphi'_S) \\ &= \sum_{T \in \mathcal{S} \setminus \{S\}} \sum_{w \in N^{\text{out}}(v_S) \cap T} b_{v_S w}(\varphi_w - \varphi_{v_S}) \\ &\quad - \sum_{T \in \mathcal{S} \setminus \{S\}} \sum_{w \in N^{\text{in}}(v_S) \cap T} b_{w v_S}(\varphi_{v_S} - \varphi_w) \\ &= \sum_{w \in N^{\text{out}}(v_S)} b_{v_S w}(\varphi_w - \varphi_{v_S}) - \sum_{w \in N^{\text{in}}(v_S)} b_{w v_S}(\varphi_{v_S} - \varphi_w) \\ &= \sum_{w \in N^{\text{out}}(v_S)} g_{v_S w} - \sum_{w \in N^{\text{in}}(v_S)} g_{w v_S} = 0 \end{aligned}$$

  where the last equality follows from Remark 2.37.

Hence, the generalized differential flow induced by φ' is feasible. Let $E' := \{(S,T) \in \mathcal{S} \times \mathcal{S} \mid S \neq T \text{ and } b'_{ST} > 0\}$ and observe that $F' \subset E'$, since the definition of F' captures exactly those pairs (S,T) for which the first case in the definition of b' applies. Hence, the anti-arborescence $(\mathcal{S}, F')$ is a subgraph of $(\mathcal{S}, E')$ and we now obtain from Lemma 2.40 that there exists $\varphi_0 \in \mathbb{R}$ such that $\varphi' \equiv \varphi_0$ and thus $\varphi \equiv \varphi_0$ which implies that $g \equiv 0$. □

Note that Theorem 2.41 in particular excludes the structure encountered in Example 2.32. Using Theorem 2.41, we can prove another sufficient condition, which excludes a different aspect of Example 2.32.

Theorem 2.42
Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph, $f \in Q_f^{\mathrm{DC}}(G)$ and $F = (E_F, V_F)$ an α-tree in G that conforms with f. Furthermore, let F be such that there is at most one $v \in V_F$ with $\deg(v) \geq 3$. Then, f is extremal.

Proof. We prove the statement by induction over the number $|V_F|$ of active vertices in F. If $|V_F| = 0$, i. e. there is no vertex active in F, then (V, E_F) is in fact a spanning tree and hence f is extremal (as we noted after Remark 2.37, any potential φ that induces a differential flow g with $f \pm g \in Q_f^{\mathrm{DC}}(G)$ has to be constant within connected components of (V, E_F), which in this case already implies that $g \equiv 0$).

Now, let $|V_F| = k$. We distinguish two cases: First, assume that V_F contains no vertex of degree at most 2 that is incident with an active edge. Then, for every weakly connected component S of (V, E_F), one of the following holds:

a) S is a singleton

b) every active vertex v in S is incident with an active edge, which by our assumption implies that $\deg(v) \geq 3$. Hence, there can only be one such active vertex and v is thus the only active vertex in S.

In both cases, S contains at most one active vertex, which implies that F satisfies the conditions of Theorem 2.41 and f is extremal.

Now, assume that V_F does contain a vertex of degree at most 2 that is incident with an active edge and let $\alpha_F : V_F \to E \setminus E_F$ be a vertex map for F.

Let $v \in V_F$ be an active vertex of degree at most 2 that is incident with an active edge $e^* \in E_F$. By assumption, $\deg(v) \leq 2$, but since v is incident with both $e^* \in E_F$ and $\alpha_F(v) \in E \setminus E_F$, these two edges are the only edges incident with v. Assume w. l. o. g. that both edges e^* and $\alpha_F(v)$ are oriented away from v and let $s, t \in V$ such that $\alpha_F(v) = (v,t)$ and $e^* = (v,s)$. Let $p_v^* := \sum_{e \in \delta^{\mathrm{out}}(v)} f_e - \sum_{e \in \delta^{\mathrm{in}}(v)} f_e = f_{vt} + f_{vs}$.

Since v and (v,s) are active, it holds that $p_v^* \in \{p_v^+, p_v^-\}$ and $f_{vs} \in \{f_{vs}^+, f_{vs}^-\}$. Thus, for every differential flow g with $f \pm g \in Q_f^{\mathrm{DC}}(G)$, it must hold by Remark 2.37 that

$g_{vs} = 0$ and hence $g_{vt} = 0$. Let $\bar{f}'$ be given by

$$\bar{f}'_e = \begin{cases} (p^*_v - f_{vs}, p^*_v - f_{vs}) & \text{if } e = (v,t) \\ \bar{f}_e & \text{else} \end{cases} \tag{2.66}$$

and define $H := (V, E, b, \bar{p}, \bar{f}')$. Then, we can conclude that f is extremal in $Q^{\mathrm{DC}}_f(G)$ if and only if f is extremal in $Q^{\mathrm{DC}}_f(H)$.

But now, in the network graph H, we obtain a new α-tree $F' := (E_F \cup \{(v,t)\}, V_F \setminus \{v\})$ that conforms with f. Since $|V_{F'}| < |V_F|$, we can conclude by induction that f is extremal in $Q^{\mathrm{DC}}_f(H)$ and hence it is also extremal in $Q^{\mathrm{DC}}_f(G)$. □

It can easily be seen that neither of the two sufficient conditions above is necessary: In Example 2.32 for the case of $\frac{b_{wt}}{b_{ws}} \neq \frac{b_{vt}}{b_{vs}}$, the α-forest violates both conditions, but f is extremal nonetheless.

2.5.6 The Family of α-tree Non-degenerate Graphs

The criteria developed in the previous section depend on the particular solution and corresponding α-forest in question. To complement these somewhat *a-posteriori* criteria, we will present in this section a characterization of all graphs that are non-degenerate irrespective of the given capacity and elasticity vectors, i. e. graphs for which in particular the antecedent of Lemma 2.33 holds. These naturally include, but are not limited to, all graphs for which the sufficient condition from Theorem 2.42 automatically holds (i. e. graphs that have at most one vertex of degree ≥ 3). As an example, we will see that any network graph on the variant of the graph from Example 2.32 which is shown in Fig. 2.15 is non-degenerate. This is true although the α-tree shown in Fig. 2.15 violates both of our sufficient conditions from Theorems 2.41 and 2.42. Note that all results in this section rely solely on the graph (V, E) that underlies a network graph. In particular, this means that they hold independently of whether or not we assume that the elasticity vector changes in response to a modification of edge capacities.

The idea for the assumptions used in the main Theorem 2.49 are inspired by [ZT13]. However the results in this subsection, as well as the proof techniques used, are entirely different. On the other hand, a similar proof technique was independently used by Leibfried et al. [Lei+15] to answer a question about the effect of a partial relaxation of the DC equations in power flow models (which corresponds to relaxing the differential constraints in Q^{DC}_f).

Lemma 2.43
Let $H_1 := (V_1 \cup \{v^*\}, E_1)$ and $H_2 := (V_2 \cup \{v^*\}, E_2)$ be two directed, weakly connected graphs with $V_1 \cap V_2 = \emptyset$ such that every network graph on these graphs is non-degenerate. Let (V, E) be a graph that results from connecting the two graphs in v^*, i. e., $V = V_1 \cup V_2 \cup \{v^*\}$ and $E = E_1 \cup E_2$. Then, every network graph on (V, E) is non-degenerate.

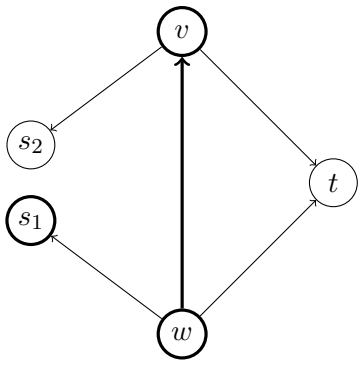


Figure 2.15: The graph shown is α-tree non-degenerate, irrespective of the elasticity vector used (Theorem 2.49). In particular, a solution conforming with the α-tree indicated by the active vertices and edges marked in bold is extremal, although the α-tree violates both of our sufficient conditions from Theorems 2.41 and 2.42 in the previous section.

Proof. Let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph on (V, E). Due to Theorem 2.36, it suffices to show that if there exists an α-tree that conforms with $f \in Q_f^{\mathrm{DC}}(G)$, then f is extremal.

Let therefore $f \in Q_f^{\mathrm{DC}}(G)$ and let F be an α-tree with vertex map $\alpha_F : V_F \to E \setminus E_F$ that conforms with f. If v^* is active in F, then we assume w. l. o. g. that $\alpha_F(v^*) \in E_2$.

Suppose that f is not extremal, i. e. there are flows $\bar{f}, \tilde{f} \in Q_f^{\mathrm{DC}}(G)$ and $\lambda \in (0, 1)$ such that $\tilde{f} \neq \bar{f}$ and $f = \lambda\bar{f} + (1 - \lambda)\tilde{f}$. We will prove that, in fact, $\tilde{f} = \bar{f} = f$, a contradiction.

We write $f = (f^1, f^2)$ for the restrictions of f to E_1 and E_2 and analogously for $\bar{f}$ and $\tilde{f}$. Consider the network graph G_1 obtained from restricting G to the graph H_1 and changing the bounds for the vertex v^* in such a way that $\bar{f}_1$ and $\tilde{f}_1$ (and thus also f_1) are feasible for G_1. Specifically, let

$$p_{v^*}^- = -\max\left\{\left|\sum_{e\in E_1} A_{v^*e}\bar{f}_e\right|, \left|\sum_{e\in E_1} A_{v^*e}\tilde{f}_e\right|\right\}$$
$$p_{v^*}^+ = \max\left\{\left|\sum_{e\in E_1} A_{v^*e}\bar{f}_e\right|, \left|\sum_{e\in E_1} A_{v^*e}\tilde{f}_e\right|\right\}.$$

Note that, indeed, by this definition, f^1, $\bar{f}^1$ and $\tilde{f}^1$ are all feasible for G_1. Furthermore, let $E_{F_1} := E_F \cap E_1$ and $V_{F_1} := V_F \cap V_1$. Then $F_1 := (E_{F_1}, V_{F_1})$ is an α-forest (with vertex map $\alpha_{F_1} := \alpha_F|_{V_1}$) which conforms with f^1. Since F was an α-tree in G and either $v^* \notin V_F$ or, by our assumption, $\alpha_F(v^*) \in E_2$, it holds that $E_{F_1} \cup \alpha_{F_1}(V_{F_1}) = (E_F \cup \alpha_F(V_F)) \cap E_1$, which means that F_1 is actually an α-tree in G_1.

Furthermore, since G_1 is a network graph on H_1 it is non-degenerate and hence f^1 is an extremal flow in G_1. Finally, since $\bar{f}^1$ and $\tilde{f}^1$ are also feasible for G_1 and $f = \lambda\bar{f} + (1-\lambda)\tilde{f}$, we have in particular that $\bar{f}^1 = \tilde{f}^1 = f^1$.

If v^* is not active in F, then we can apply the same argument to conclude that $\bar{f}^2 = \tilde{f}^2 = f^2$ and thus $\bar{f} = \tilde{f} = f$, a contradiction. Therefore, let us now assume that v^* is indeed active in F, which by our assumption implies that $\alpha_F(v^*) \in E_2$.

Since F conforms with f, the flow f satisfies either the vertex constraint in v^* with equality in the network graph G, w. l. o. g. let the upper bound $p^+_{v^*}$ be binding (the argument is the same if the lower bound $p^-_{v^*}$ is binding instead). We now have

$$\sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} \bar{f}_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} \bar{f}_e \leq p^+_{v^*} - \left(\sum_{e\in\delta^{\text{out}}(v^*)\cap E_1} \bar{f}_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_1} \bar{f}_e\right) \tag{2.67}$$

$$= p^+_{v^*} - \left(\sum_{e\in\delta^{\text{out}}(v^*)\cap E_1} f_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_1} f_e\right) \tag{2.68}$$

$$= \sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} f_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} f_e, \tag{2.69}$$

where the inequality (2.67) follows from the fact that $\bar{f}$ is feasible for the network graph G, the equality (2.68) follows from $f^1 = \bar{f}^1$ (which we have proved above) and the equality (2.69) follows from the fact that f satisfies the vertex constraint in v^* with equality. By the same argument, $\sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} \tilde{f}_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} \tilde{f}_e \leq \sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} f_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} f_e$ which, since $f = \lambda\bar{f} + (1-\lambda)\tilde{f}$, implies that

$$\begin{aligned}\sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} f_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} f_e &= \sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} \bar{f}_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} \bar{f}_e \\ &= \sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} \tilde{f}_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} \tilde{f}_e.\end{aligned}$$

We now consider the network graph G_2 defined as follows: We restrict G to the graph H_2 and fix both upper and lower bound of v^* by

$$p^-_{v^*} = p^+_{v^*} = \sum_{e\in\delta^{\text{out}}(v^*)\cap E_2} f_e - \sum_{e\in\delta^{\text{in}}(v^*)\cap E_2} f_e. \tag{2.70}$$

By our observations above, all of $f, \bar{f}, \tilde{f}$ are indeed feasible for G_2. Furthermore, the α-forest $(E_F \cap E_2, \{v \in V_F \mid \alpha_F(v) \in E_2\}$ is an α-tree in G_2 and by (2.70) it conforms with f. Since G_2 as a network graph on H_2 is non-degenerate, we have that $\bar{f}^2 = \tilde{f}^2 = f^2$. In summary, we obtain $\bar{f} = \tilde{f} = f$, a contradiction. □

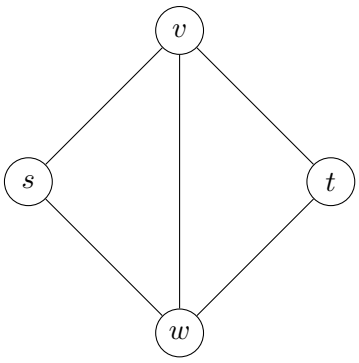


Figure 2.16: The diamond graph.

In order to make use of the above lemma, we now prove that network graphs on certain basic families of graphs are always non-degenerate. We start by considering network graphs of the extremely basic graph consisting of a single edge.

Lemma 2.44
If $G = (V, E, b, \bar{p}, \bar{f})$ is a network graph with $(V, E) = (\{v, w\}, \{(v, w)\})$, then G is non-degenerate.

Proof. Let $f \in Q_f^{\mathrm{DC}}(G)$ be extremal. Then, by Theorem 2.36, there exists an α-tree in G that conforms with f.

Conversely, let $f \in Q_f^{\mathrm{DC}}(G)$ and let F be an α-tree in G that conforms with f. If f is not extremal then there exist $f', f'' \in Q_f^{\mathrm{DC}}(G)$ with $f'_{vw} < f_{vw} < f''_{vw}$. This means that no vertex or edge constraint can be tight for f and hence F must be empty and cannot be an α-tree. □

The next corollary follows immediately from repeated application of Lemma 2.43 and Lemma 2.44, keeping in mind that every tree has a leaf.

Corollary 2.45
If $G = (V, E, b, \bar{p}, \bar{f})$ is a network graph such that the graph (V, E) is a tree, then G is non-degenerate.

Next, we consider network graphs for which the underlying graph is an undirected cycle.

Definition 2.46
Let K_4 be the complete undirected graph on 4 vertices. The graph that results from deleting one edge from K_4 is called the *diamond graph* (see Fig. 2.16). Remember that a graph G is a *subdivision* of a graph H, if G can be obtained by adding an arbitrary number of additional vertices along the edges of H. Furthermore, H is a *topological minor* of G if G contains a subdivision of H as a subgraph (see, e. g., [Die17]).

An undirected graph G is called *cactus* if it does not contain the diamond graph as a topological minor. We say that a directed graph (V, E) is a cactus if the graph that results from replacing all edges in E by undirected edges (and removing multi-edges if necessary) is a cactus.

Lemma 2.47
If $G = (V, E, b, \bar{p}, \bar{f})$ is a network graph such that E is an undirected cycle, then G is non-degenerate.

Proof. Let $f \in Q_f^{\mathrm{DC}}(G)$ be extremal. Then, by Theorem 2.36, there exists an α-tree in G that conforms with f.

Conversely, let $f \in Q_f^{\mathrm{DC}}(G)$ and F an α-tree which conforms with f. As all vertices in G have degree 2, F cannot contain an active vertex with degree ≥ 3. Therefore, by Theorem 2.42, f is extremal. □

The above results can be combined to characterize a set of graphs on which network graphs are guaranteed to be non-degenerate.

The diamond graph (Fig. 2.16) is known in the context of electrical circuits as the *Wheatstone bridge*, a device to measure the electrical parameters of a circuit [Whe43]. The structure is characterized by the fact that if a voltage is applied between the vertices s and t, then the direction of the current between v and w changes in response to the resistance of the individual branches.

We say that a directed graph is weakly 2-connected if the graph contains at least 3 vertices and remains weakly connected whenever any one if its vertices is removed (see, e. g., [Die17]). The following equivalent characterization of cacti can be found, e. g., in [EC88]:

Proposition 2.48
A directed graph $G = (V, E)$ is a cactus if and only if the edges of every induced subgraph which is weakly 2-connected form a simple undirected cycle. Equivalently, G is a cactus if and only if no edge $e \in E$ appears in more than one simple undirected cycle.

As suggested above, we will now prove that every network graph defined on the graph (V, E) is guaranteed to be non-degenerate if and only if the graph (V, E) is a cactus.

Theorem 2.49
A network graph $G = (V, E, b, \bar{p}, \bar{f})$ is non-degenerate for all choices of $b, \bar{p}, \bar{f}$ if and only if (V, E) is a cactus.

Proof. We first prove the "if"-part of the statement by induction over the number $|V|$ of vertices in V. If $|V| = 2$, then the statement is true by Lemma 2.44. Suppose that the statement is true for all graphs with $|V| < n$ and let $G = (V, E, b, \bar{p}, \bar{f})$ be a network graph such that the graph (V, E) is a cactus graph with $|V| = n$.

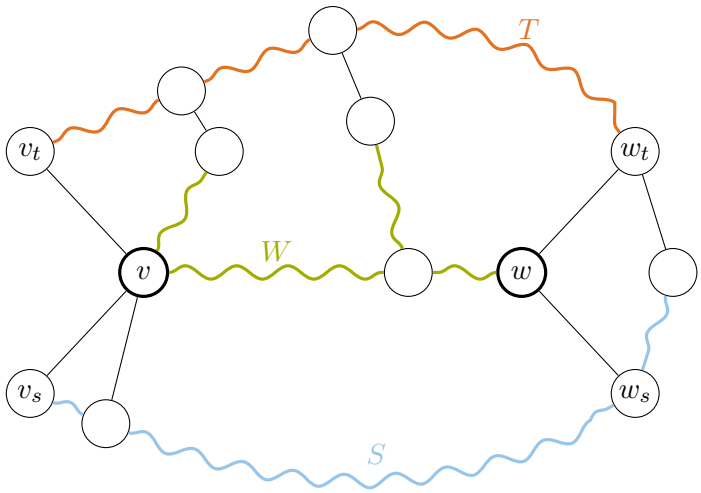


Figure 2.17: Given a graph that contains the diamond graph as a topological minor, we select three sets of edges S, T and W that cover all vertices as shown in the figure. We can now choose capacities and edge weights in such a way that we obtain a degenerate network graph: In particular, a solution f conforming with the α-tree $(\{v, w\}, S \cup T \cup W)$ (shown in bold) need not be extremal.

If (V, E) contains a cut-vertex v^* such that the subgraph induced by $V \setminus \{v^*\}$ consists of two sets of vertices V_1 and V_2 which are disconnected, then we know by definition that the subgraphs induced by $V_1 \cup \{v^*\}$ and $V_2 \cup \{v^*\}$ are cacti, as well (they cannot contain a topological minor that the original graph does not contain). Thus, the statement is true by induction using Lemma 2.43. If on the other hand (V, E) does not contain such a cut-vertex, then (V, E) is weakly 2-connected. As (V, E) is a cactus, this means by Proposition 2.48 that E is a simple undirected cycle and the statement follows immediately from Lemma 2.47.

For the "only-if"-part, assume that (V, E) is not a cactus graph and thus contains the diamond graph as a topological minor. Let $H = (V, E')$ denote a minimal subgraph of (V, E) that still contains the diamond graph as a topological minor. Then, H is a subdivision of the diamond graph.

Let $v, w \in V$ denote the two vertices of degree 3 in H. As H is a subdivision of the diamond graph, it consists of three vertex-disjoint undirected paths connecting v and w, at least two of which have length at least 2 (see Fig. 2.16). Choose two such paths and on each of these, denote the intermediate vertex directly adjacent to v by v_s and v_t, respectively. Analogously, denote the intermediate vertex directly adjacent to w by w_s and w_t (see Fig. 2.17). Note that it may be the case that $v_s = w_s$ and/or $v_t = w_t$.

Denote by P_{vw} the undirected path between v and w in H which does not pass through any of v_s, w_s, v_t, w_t. Furthermore, denote by P_s the undirected path between

v_s and w_s which does not contain v or w, and by P_t the undirected path between v_t and w_t which does not contain v or w. As H is a subgraph of (V, E), the graphs (V, P_{vw}), (V, P_s) and (V, P_t) are also subgraphs of (V, E).

We now build three sets of edges $W \supset P_{vw}$, $S \supset P_s$ and $T \supset P_t$ such that every vertex in V is covered by exactly one of the three sets as follows: Let $W \subset E$ be the set of edges of a maximal tree which contains P_{vw} and does not cover any vertex that is already covered by P_s or P_t. Analogously, let S be the set of edges of a maximal tree which contains P_s and does not cover any vertex already covered by W or P_t. Finally, let T be the set of edges of a maximal tree which contains P_t and does not cover any vertex already covered by W or S.

To see that every vertex is covered by either W, S or T, assume that a vertex v^* is not covered. Since (V, E) is weakly connected, there exists an undirected path P connecting v^* to v. Let w_1 be the first vertex from W, S or T that is encountered on this path (such a vertex exists since v is covered by W) and denote its predecessor on the path by w_2 (it might be the case that $w_2 = v^*$). Assume, w. l. o. g., that w_1 is covered by W. Then, as w_2 is not covered by $W \cup S \cup T$, the edge connecting w_1 and w_2 can be added to W, a contradiction to maximality of W.

We obtain three sets of edges S, T and W that do not contain an undirected cycle. Furthermore, v_s, w_s are covered by S, v_t, w_t are covered by T and v, w are covered by W. Finally, $W \cup S \cup T$ covers all vertices in G. Now, let E_{vs}, E_{ws}, E_{vt}, E_{wt} denote the sets of edges that connect v and w with nodes covered by S and T, respectively. We set $f_e^- = f_e^+ = 0$ for all edges e in $W \cup S \cup T$ and $-f_e^- = f_e^+ = \infty$ for all other edges. Similarly, we set $p_v^- = p_v^+ = p_w^- = p_w^+ = 0$ and $-p_{v'}^- = p_{v'}^+ = \infty$ for all $v' \in V \setminus \{v, w\}$. Let $f \equiv 0 \in \mathbb{R}^E$. Then, f is feasible for the resulting network graph and the α-tree $\{T \cup W \cup S, \{v, w\}\}$ conforms with f.

At the same time, we can choose b such that

$$\frac{\sum_{e \in E_{wt}} b_e}{\sum_{e \in E_{ws}} b_e} = \frac{\sum_{e \in E_{vt}} b_e}{\sum_{e \in E_{vs}} b_e} = 1.$$

Then, with

$$\varphi_{v'} = \begin{cases} 0 & \text{if } v' \text{ is covered by } W \\ 1 & \text{if } v' \text{ is covered by } S \\ -1 & \text{if } v' \text{ is covered by } T \end{cases}$$

we claim that $\pm B^\top \varphi \in Q_f^{\mathrm{DC}}(G)$: For every edge $e \in S \cup T \cup W$, it holds that $(B^\top \varphi)_e = 0$. The only other bounds that are not $\pm\infty$ (and hence can possibly be active) are the

vertex constraints in v and w. But there,

$$\begin{aligned}(AB^\top\varphi)_v &= \sum_{v'\in N^{\text{in}}(v)\cap W} b_{v'v}(\varphi_v-\varphi_{v'}) + \sum_{v'\in N^{\text{in}}(v)\cap T} b_{v'v}(\varphi_v-\varphi_{v'}) + \sum_{v'\in N^{\text{in}}(v)\cap S} b_{v'v}(\varphi_v-\varphi_{v'}) \\ &\quad - \sum_{v'\in N^{\text{out}}(v)\cap W} b_{vv'}(\varphi_{v'}-\varphi_v) - \sum_{v'\in N^{\text{out}}(v)\cap T} b_{vv'}(\varphi_{v'}-\varphi_v) - \sum_{v'\in N^{\text{out}}(v)\cap S} b_{vv'}(\varphi_{v'}-\varphi_v) \\ &= \sum_{v'\in N^{\text{in}}(v)\cap T} b_{v'v}\cdot 1 + \sum_{v'\in N^{\text{in}}(v)\cap S} b_{v'v}\cdot(-1) \\ &\quad - \sum_{v'\in N^{\text{out}}(v)\cap T} b_{vv'}\cdot(-1) - \sum_{v'\in N^{\text{out}}(v)\cap S} b_{vv'}\cdot 1 \\ &= \sum_{e\in E_vt} b_e - \sum_{e\in E_vs} b_e = 0\end{aligned}$$

and analogously for $(AB^\top\varphi)_w$. Thus, by Remark 2.37, $f = \frac{1}{2}(B^\top\varphi + (-B^\top\varphi))$ is not extremal, which proves that the network graph $(V, E, b, \bar{p}, \bar{f})$ is degenerate. □

The following fact is generally well-known (see, e. g., [BLS99]) and can be seen by adapting the depth-first search algorithm in a suitable way. It provides us with a simple algorithm to determine whether non-degenerateness can be guaranteed for a given graph.

Theorem 2.50
Let $G = (V, E)$ be a graph. It can be decided in linear time whether G is a cactus.

We have already encountered the diamond graph, the forbidden minor which characterizes cacti, in Example 2.32: It was our first example of a network graph that is degenerate. Interestingly, Theorem 2.49 now shows that this is in fact the characterizing structure for non-degenerateness, at least from a topological point of view.

Cactus graphs are closely related to several results from the context of electrical circuits and related areas: They represent exactly those graphs that, if we add a single edge, remain *confluent* (or *of series parallel type*) in the sense of [Duf65] (remember that a graph is *of series parallel type* if it can be constructed by iteratively replacing an edge by two sequential or two parallel edges, starting from a loop). Such networks are those, for which the equivalent resistance can be computed by iteratively applying Ohm's laws for resistors in series and parallel. In this context, it is maybe not surprising that these networks also encompass most network topologies on which tight convex relaxations for AC Optimal Power Flow have been found (see Section 2.4), including so-called radial networks. This provides an alternative perspective also on previous results about these convex relaxations: The networks in which the AC model makes the Optimal Power Flow problem (asymptotically) computationally no more complex

than the DC-OPF are also those in which the structure of the DC-OPF is closest to the structure of the TR-OPF.

The family of cactus graphs is thus a limited, but well-known family of networks on which electrical power flows are "well-behaved". However, network topology is only one aspect of non-degenerateness, in fact the exact values of the vector b of edge weights highly matter, as well: The network graph from Example 2.32 is almost always non-degenerate unless it holds that $\frac{b_{14}}{b_{13}} = \frac{b_{24}}{b_{23}}$. In the theory of electrical circuits, this case is known as the one where the Wheatstone bridge is *balanced*.

This implies that we can in fact expect non-degenerateness to hold not only for the (quite limited) family of network graphs covered by Theorem 2.49. Indeed, all network graphs are non-degenerate unless the weight vector b is chosen from the union of finitely many lower-dimensional subspaces of the parameter space (one subspace for every diamond structure in the network graph). In other words, every network graph can be made non-degenerate by slightly perturbing the edge weights.

We conclude this section noting that it is $\mathbb{NP}$-complete (see [GJ79]) to determine whether a given network graph is non-degenerate.

Theorem 2.51
Given a network graph $G = (V, E, b, \bar{p}, \bar{f})$, it is $\mathbb{NP}$-complete to decide whether G is α-tree non-degenerate.

Proof. We first settle membership in $\mathbb{NP}$: Suppose that there exists a non-extremal solution $f \in Q_f^{\mathrm{DC}}$ that conforms with the α-tree $F = (E_F, V_F)$. Let P denote the set of vertices of Q_f^{DC} that are contained in the minimal face of Q_f^{DC} which contains f (since f is not extremal, $|P| \geq 2$) and let $f' := {}^1/_{|P|} \sum_{p \in P} p$. Then, f' conforms with F, as well and is not extremal (since $|P| \geq 2$). Given a certificate in the form of P, F and a corresponding vertex map $\alpha : V_F \to E \setminus E_F$ (which are all polynomial in size), we can easily check that F is indeed an α-tree, f' conforms with F and, as a convex combination of the points in $|P|$, is not extremal (if $|P| \geq 2$).

To prove the hardness, we provide a reduction from the $\mathbb{NP}$-complete problem SubsetSum [GJ79]:[4] Given a number $n \in \mathbb{N}$, sizes $\alpha_i \in \mathbb{N}$ for each $i \in [n]$ and $\beta \in \mathbb{N}$, decide whether there is a subset $I \subset [n]$ with $\sum_{i \in I} \alpha_i = \beta$.

Let $(n, \alpha_1, \alpha_2, \ldots, \alpha_n, \beta)$ be an instance of SubsetSum. Consider the network graph depicted in Fig. 2.18 where the shown edge weights denote the edge elasticity b, thickly dashed edges have a upper capacity bound of 1, $\bar{p}_v = (-\beta, \beta)$ and $\bar{p}_w = \bar{p}_{v_1} = \cdots = \bar{p}_{v_n} = (0, \infty)$. All other edges and vertices have infinite upper and lower bounds. We prove that the SubsetSum instance is a *yes* instance if and only if the network graph is degenerate.

[4] Garey and Johnson [GJ79] attribute this to Karp [Kar72], where the problem is not explicitly mentioned. However, it can easily be reduced from Partition, the hardness of which is indeed proven in [Kar72].

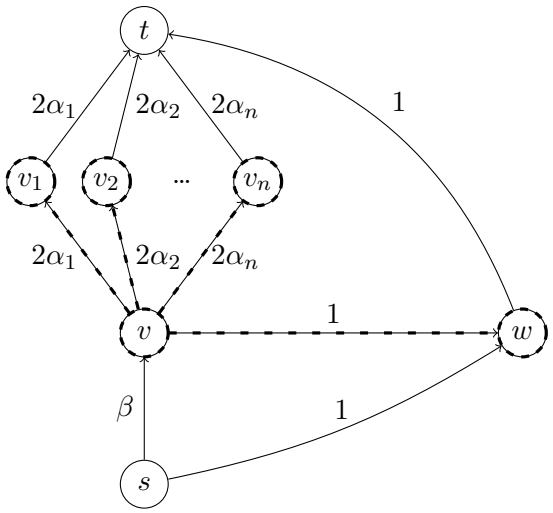


Figure 2.18: Solving the well-known ℕℙ-complete problem SubsetSum is equivalent to determining whether the shown network graph (the edge weights represent the edge elasticity b) is non-degenerate.

Let f be a flow and $F = (E_F, V_F)$ an α-tree which conforms with f. Let $\alpha_F : V_F \to E \setminus E_F$ be a vertex map for F. We start by observing that by Theorem 2.42 and since $\bar{p}_t = (-\infty, \infty)$, f is extremal unless both vertices v, w are active in F. Now, let φ be a potential such that $f \pm B^\top \varphi$ is feasible. By Remark 2.37, this means in particular that $B^\top \varphi$ disappears on all inequalities active in F. Without loss of generality, let $\varphi_v := 0$. Since $f = 1/2(f + B^\top \varphi + f - B^\top \varphi)$, it holds that f is extremal if and only if $\varphi \equiv 0$ is the only possible solution.

We now show that if for one of the vertices v_i, both the vertex and the edge connecting it to v are active, then f is extremal. As the edge (v, v_i) is active, we have that $\varphi_{v_i} = 0$. Similarly, as v_i is active, $\varphi_t = 0$. We now distinguish two cases:

a) The edge (v, w) is active. In this case, it follows that $\varphi_w = \varphi_v = 0$ and hence $\varphi_s = 0$ by activity of w. As F is an α-tree, for every vertex v_j at least one of the two edges incident with v_j must be contained in $E_F \cup \alpha_F(V_F)$. Since F conforms with f, we have that for *all but one* of these, either v_j is active or the edge (v, v_j) (the remaining vertex can be connected by v). In both cases, $\varphi_v = \varphi_t = 0$ implies that $\varphi_{v_j} = 0$. For the final vertex v_{j^*}, the same now follows by activity of v.

b) The edge (v, w) is not active. Note that, as F is an α-tree, $|(E_F \cup \alpha_F(V_F)) \cap \{(s, v), (s, w), (v, w), (w, t)\}| = 2$ and hence, since $s, t \notin V_F$, we must have $\alpha_F(v) \in \{(s, v), (v, w)\}$. As above, for every vertex v_j at least one of the two edges incident with v_j must be contained in $E_F \cup \alpha_F(V_F)$. Since F conforms with f and this time $\alpha_F(v) \in \{(s, v), (v, w)\}$, we have that for *all* of these, either v_j or the edge (v, v_j)

is active. In both cases, $\varphi_v = \varphi_t = 0$ implies that $\varphi_{v_j} = 0$. Now, suppose that $\varphi_w > \varphi_v$. Then, activity of v implies that $\varphi_s < \varphi_v$. At the same time, activity of w implies (since $\varphi_t = \varphi_v < \varphi_w$) that $\varphi_s > \varphi_w$, a contradiction. Therefore, $\varphi = 0$.

We have concluded that $\varphi \equiv 0$ if for one of the vertices v_i, both the vertex and the edge connecting it to v are active.

If f is *not* extremal, it must therefore hold that for all vertices v_i, only the vertex itself or the edge connecting it to v can be active. At the same time, as F is an α-tree it has to hold that $|F| = |A| + 3$. This implies that the edge (v, w) is active and for all vertices v_i, *exactly* one of the two (the vertex itself or the edge connecting it to v) needs to be active. In this case, $\varphi_v = 0$ implies $\varphi_w = 0$ (by activity of (v, w)) and furthermore $\varphi_{v_j} = 0$ for all v_j for which the edge (v, v_j) is active.

Now, suppose that $\varphi_t \neq 0$ and assume w. l. o. g. that $\varphi_t > 0$. By activity of w, it follows that $\varphi_s = -\varphi_t$ and in particular $\varphi_s \neq 0$. At the same time, for all vertices v_j that are active, it follows that $\varphi_{v_j} = 1/2 \cdot \varphi_t$ (since $b_{vv_j} = b_{v_j t}$). The vector φ now satisfies all constraints that are active in F if and only if the flow balance at v is 0, i. e.,

$$-\varphi_s \cdot \beta = (\varphi_v - \varphi_s) \cdot \beta = \sum_{v_a \text{ active in } F} \left(\frac{\varphi_t}{2} - \varphi_v\right) \cdot 2\alpha_a = -\varphi_s \cdot \sum_{v_a \text{ active in } F} \alpha_a.$$

Such a selection of active vertices v_j hence exists if and only if $(n, \alpha_1, \alpha_2, \ldots, \alpha_n, \beta)$ is a *yes* instance of SubsetSum. □

2.6 Conclusion

We conclude this chapter by a brief summary of the results and an outlook on interesting research questions that arise in the context of our results.

In Section 2.1, we gave a brief overview of network models used in energy system optimization. In particular, we presented three loss functions η^{AC}, $\eta^{\log}$ and η^2 that can be seen as good approximations of Ohmic losses incurred on a transmission line under a given load. We then used one of these loss functions, $\eta^{\log}$, to derive a global equivalence result for optimal flows under the Transport model and the DC model (Theorem 2.14). To support the usage of (non-linear) loss functions in the context of LP based optimization models, we derived a special type of piecewise-linear approximation that keeps the approximation error constant if we change the capacity of a transmission line, e. g. in the context of a TCEP (Theorem 2.17).

We then turned our attention to the caveats implied by the rather strong assumptions on the underlying network structure imposed by Theorem 2.14: We evaluated their practical effect in three reference networks (Section 2.2.2) and finally focussed on the structural differences between the sets of TR- and DC-feasible flows. These sets are almost independent of the loss function used, which implies that the difference between

them remains, even in a higher-accuracy Transport model which uses the loss function $\eta^{\log}$.

We derived a number of statements on the relation between the set of TR-feasible and DC-feasible power injections (projections of the feasible regions of TR-OPF and DC-OPF where the latter set is contained in the former). Under reasonable assumptions, both sets share several points on their boundary (Theorem 2.26), despite the fact that the set of DC-feasible power injections is generally "much smaller" (for a more precise statement, see Theorem 2.24). On the reverse, we provide some bounds on how much smaller it can be, depending on the parameters of the network (Theorem 2.27).

Finally, we investigated the differential flow polytope, another projection of the feasible regions of TR-OPF and DC-OPF (Section 2.5). This polytope bears some resemblance to well-known network flow polyhedra and we derived a characterization of their extremal points in terms of so-called α-trees, similar to the characterization of extremal points in the network flow polyhedron by spanning trees consisting of edges with non-zero flow.

We prove that our condition is necessary for extremality in all networks (Theorem 2.36). While it is generally sufficient for almost all choices of the network parameters, it is sufficient regardless of the choice of network parameters for networks on a family of graphs known as cacti (Theorem 2.49). Finally, for a given network, it is generally $\mathbb{NP}$-hard to decide whether our condition is indeed sufficient (Theorem 2.51).

From the perspective of power networks, this theoretical work contributes in two ways to a better understanding of DC power flows: First, it allows us to translate the set of active inequalities in a DC-feasible point into a structure on the underlying network graph that can be interpreted, e. g. in terms of identifying bottlenecks in the network. Second, it provides us with an alternative perspective on points (obtained, e. g., as solutions to the Transport model) that are *infeasible* for the DC model:

The flow values of such a solution typically violate the DC equations (2.44). Starting from there, it is very difficult to obtain any information about how a *similar* DC-feasible flow might look or how production and transmission capacities would have to be adapted in order to admit a DC-feasible flow at all. On the other hand, starting from an α-tree, it is easy to obtain a corresponding potential flow that automatically satisfies the DC equations (2.44), but instead might violate some edge or vertex capacities. This gives a much better indication as to how capacities can be adapted to achieve feasibility.

Our results raise a number of questions that we would deem worthy of further attention in the future:

- Can the relative size of the sets of TR- and DC-feasible power injections be specified beyond Theorem 2.24 and Theorem 2.27? In particular, can we derive bounds that take into account the size of the network and/or the values of the network parameter b?

- In light of the experimental results from Section 2.2.2, can properties of an optimal solution to the modified Transport model be used to obtain a bound on the error value of that solution (without having to compute a solution to the DC model to compare it against)?
- Combining the two questions above, can we evaluate a solution from the Transport model in a network that is infeasible for the DC model to provide us with an useful indication of how much capacities would have to be increased to make the DC model feasible?
- In light of the hardness result in Theorem 2.51 and our positive results for networks on cacti, is there a larger family of networks for which we can guarantee the sufficiency of our characterization of extremal points if we take into account the edge parameters b?
- As argued above, every α-tree provides us with a potential flow that violates certain edge and vertex constraints. Starting from a solution to the TR-TCAP, how can a useful α-tree be extracted such that this violation is in some sense "small"? Can this be used to obtain a useful upper bound on the cost of an optimal solution to the DC-TCAP from a solution to the TR-TCAP?
- Can we interpret/reformulate the simplex algorithm for DC-OPF in terms of α-trees (analogously to the network simplex algorithm)? This might provide us with a (faster) combinatorial algorithm for DC-OPF as well as more valuable information if no feasible solution is found (again, with respect to edge and vertex constraints).

Chapter 3

Benders Decomposition for Energy System Optimization

Contents

Large-scale optimization models that try to capture energy systems at a high temporal and geographical resolution easily reach the limits of what can be handled on a given computing platform, both with respect to the available memory and computation time. In this situation, decomposition approaches are a common tool, not just in the

context of energy system optimization, to improve computational performance. One such approach, Benders decomposition, has received particularly wide adoption in the context of energy system optimization (see, e. g., [Lat+03; SF05]) due to the particular problem structure often encountered in those settings.

In this chapter, we describe the results from our analysis of the Benders decomposition algorithm. Since these are mostly not restricted to the context of energy system optimization, we will keep the first part (Sections 3.1 to 3.3) entirely general. In the second part (Sections 3.4 and 3.5) we apply our results to the context of energy system optimization and present some computational experiments.

Major parts of Sections 3.1 and 3.2 in this chapter are currently being prepared for publication as [BS19b].

3.1 Benders Decomposition

Consider a generic optimization problem with two subsets of variables x and y where x is restricted to lie in some set $S \subset \mathbb{R}^n$ and x and y are jointly constrained by a set of m linear constraints. Such a problem can be written in the following form:

$$\begin{aligned} \min\ & c^\top x + d^\top y \\ \text{s.t. } & Hx + Ay \leq b \\ & x \in S \subset \mathbb{R}^n \\ & y \in \mathbb{R}^k \end{aligned} \tag{3.1}$$

The matrix $H \in \mathbb{R}^{m \times n}$, sometimes called *interaction matrix*, captures the influence of the x-variables on the y-subproblem: For fixed x^*, (3.1) reduces to an ordinary linear program with constraints $Ay \leq b - Hx^*$.

We are interested in cases where the size of the complete problem (3.1) leads to infeasibly high computation times (or memory demands), but both the problem over S and the problem resulting from fixing x can separately be solved much more efficiently due to their special structures. We will later give some examples where this is the case.

To deal with such problems, Benders [Ben62] introduced a method that works by iterating between these two "easier" problems. In this section, we will present three alternative perspectives on this approach: We begin by viewing it as a sampling algorithm before we proceed to the standard view employed, for instance, in the original paper by Benders [Ben62]. The standard perspective can be viewed as a more algebraic one, but we subsequently derive the same results in a purely geometric way, which provides us with new insights which we then use in the remainder of the chapter to improve state-of-the-art implementations of Benders decomposition.

For a problem of the form (3.1), let the function $z : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ represent the

value of the optimal y-part of the objective function for a given vector x:

$$z(x) := \min_{y \in \mathbb{R}^k} \left\{ d^\top y \;\middle|\; Ay \leq b - Hx \right\} \tag{3.2}$$

Note that we follow the common convention that the minimum over an empty set is ∞. Note further that z is convex (see Lemma 3.1) and let us assume for now that the set $\{y \in \mathbb{R}^k \mid Ay \leq b - Hx\}$ is non-empty and bounded for all $x \in \mathbb{R}^n$, i. e., $z(x) \in \mathbb{R}$ for all $x \in \mathbb{R}^n$.

As an explicit description of z may be computationally prohibitive to obtain, we can use a sampling approach to approximate z by a function $\bar{z}$ that is composed of subgradients of z (see Definition A.13) in different support points $x^1, \ldots, x^N$:

$$\bar{z}(x) := \max_{i \in [N]} \{ z(x^i) + (a^i)^\top (x - x^i) \}$$

where $a^i \in \partial z(x^i)$ is a subgradient of z in x^i. Note that $\bar{z}$, as a maximum over linear functions, is a convex piecewise-linear function. Furthermore, since all a^i are subgradients of z, it follows immediately from the definition that $\bar{z}(x) \leq z(x)$ for all $x \in S$.

We now assume that an algorithm is available to solve convex piecewise-linear optimization problems over the set S without any further constraints. The Benders decomposition algorithm then iteratively refines the above approximation by adding new support points to $\bar{z}$. It proceeds by repeatedly solving the *master problem*

$$\begin{aligned} \min\ & c^\top x + \bar{z}(x) \\ & x \in S \end{aligned} \tag{3.3}$$

to obtain a tentative solution x^* for the x-variables. Based on this tentative solution, it then solves the dual of the *subproblem*

$$\begin{aligned} \min\ & d^\top y \\ \text{s.t. } & Ay \leq b - Hx^* \\ & y \in \mathbb{R}^k \end{aligned} \tag{3.4}$$

to obtain from an optimal dual solution a new subgradient $a^{N+1} \in \partial z(x^*)$. Then the function $\bar{z}(x)$ is refined by adding the new support point $x^{N+1} := x^*$.

There are several situations where particularly the Linear Program (3.4), which contains only y as a variable, and its dual are far easier to solve than the original problem (3.1) and a separation of the problem into a small "complicated" part (the master problem) and a large "easy" part (the subproblem) may thus present computational advantages. Two common examples are the following: If the set S is non-convex then the y-subproblem is "easy" due to its linearity. For instance, Benders had in

mind the application to mixed-integer problems where the set S is non-convex due to integrality constraints [Ben62]. Alternatively (or in addition), the matrix A may have a special structure (e. g. block-diagonal) that makes the y-subproblem "easy" because it reduces further to a set of smaller problems which can be solved independently once the x-variables are fixed.

The sampling approach outlined above has two major drawbacks: First, it is not yet clear that the algorithm terminates after a finite number of iterations. Second, it leaves open the question of what to do when the subproblem is infeasible in x^* and we thus cannot obtain a new subgradient from $\partial z(x^*)$. We will see that linear programming duality theory can be used to resolve both of these issues.

To this end, we will take a geometric approach using the epigraph of z (see Definition A.12):

$$\operatorname{epi}(z) = \left\{ (x, \eta) \in \mathbb{R}^n \times \mathbb{R} \;\middle|\; \exists y \in \mathbb{R}^k : \begin{matrix} Ay \leq b - Hx \\ d^\top y \leq \eta \end{matrix} \right\} \tag{3.5}$$

Before we move on, we state some simple observations about $\operatorname{epi}(z)$:

Lemma 3.1
Given an optimization problem of the form (3.1)*, the epigraph* $\operatorname{epi}(z)$ *is closed and convex.*

Proof. First, observe that the set $S_0 = \{x \in \mathbb{R}^n \mid z(x) < \infty\}$, being a projection of the polyhedron that defines the feasible region of (3.1) without the constraint $x \in S$, is closed and convex. Furthermore, if $z(x^*) = -\infty$ for some $x^* \in \mathbb{R}^n$, then $z(x) \in \{\pm\infty\}$ for all $x \in \mathbb{R}^n$ (since x only affects the right-hand side in (3.2)). In this case, $\operatorname{epi}(z) = S_0 \times \mathbb{R}$, which is closed and convex.

Otherwise, $z(x) > -\infty$ for all $x \in \mathbb{R}^n$. Then, for convexity, let $x, x' \in \mathbb{R}^n$ and $x'' := \lambda x + (1-\lambda x')$. If $z(x) = \infty$ (or $z(x') = \infty$), then $\lambda z(x) + (1-\lambda) z(x') = \infty = z(x'')$. Otherwise, let y, y' be corresponding minimizers in (3.2). Then $\lambda y + (1-\lambda) y'$ is feasible for (3.2) with $x := x''$ and hence

$$z(x'') \leq d^\top (\lambda y + (1-\lambda) y') = \lambda z(x) + (1-\lambda) z(x').$$

For closedness, let $x_1, x_2, \ldots \in \mathbb{R}^n$ be a series converging to a point x^*. Then either $\lim_{i\to\infty} z(x_i) = z(x^*) < \infty$ (since S_0 is closed) or $\lim_{i\to\infty} z(x_i) = \infty$. The function z is thus lower semi-continuous and hence $\operatorname{epi}(z)$ is closed (see [Roc70, Theorem 7.1]).□

Note that, in addition, we will generally assume that $\operatorname{epi}(z) \neq \emptyset$. This is a very weak technical assumption that always holds if (3.1) is feasible but is actually even weaker than that, since we do not require that $x \in S$.

The epigraph in (3.5) provides us with an alternative representation of the problem (3.1), which is captured by the following lemma, originally proved by Benders [Ben62], which we present in an alternative form using the notation that we have introduced above and writing $\operatorname{epi}_S(z) := \operatorname{epi}(z) \cap (S \times \mathbb{R})$ (see Definition A.12).

Figure 3.1: Different epigraphs for the function z. The set S consists of the three distinct points on the horizontal axis. The entire shaded region makes up $\text{epi}(z)$ (which is always unbounded in the direction of η). On the other hand, $\text{epi}_S(z)$ consists of the three rays shown in dark green. Finally, $\text{epi}_{\text{conv}(S)}(z)$ corresponds to the darkly shaded region.

Proposition 3.2
Given a problem of the form (3.1)*, the point* (x^*, η^*) *minimizes* $c^\top x + \eta$ *over the set* $\text{epi}_S(z)$ *if and only if there is an optimal solution* $(\bar{x}, \bar{y})$ *to* (3.1) *with objective value* $c^\top x + d^\top y = c^\top x^* + \eta^*$.

Proof. If $(\bar{x}, \bar{y})$ is a feasible solution to (3.1) and $\bar{\eta} := d^\top \bar{y}$ then obviously $(\bar{x}, \bar{\eta}) \in \text{epi}_S(z)$ and $c^\top \bar{x} + \bar{\eta} = c^\top \bar{x} + d^\top \bar{y}$. On the other hand, let (x^*, η^*) be a solution minimizing $c^\top x + \eta$ over $\text{epi}_S(z)$, then in particular $x^* \in S$. Now, by the definition of $\text{epi}(z)$, there exists $y^* \in \mathbb{R}^k$ with $Ay^* \leq b - Hx^*$ and $d^\top y^* = \eta^*$ (as $d^\top y^* < \eta^*$ would contradict the optimality of (x^*, η^*)). Hence, (x^*, y^*) is feasible for (3.1) and again $c^\top x^* + \eta^* = c^\top x^* + d^\top y^*$. □

This suggests the following iterative algorithm to solve the optimization problem (3.1) to optimality: Start by finding a solution $(x, \eta) \in S \times \mathbb{R}$ that minimizes $c^\top x + \eta$ without any additional constraints (add a generous lower bound for η to make the problem bounded). If $(x, \eta) \in \text{epi}(z)$, then $(x, \eta) \in \text{epi}_S(z)$ (since $x \in S$) and the solution is optimal (we can find $y \in \mathbb{R}^k$ such that (x, y) is feasible and $d^\top y = \eta$ using problem (3.4),

Input: An instance of the generic optimization problem (3.1), a lower bound $\bar{\eta}$ for $z(x)$
Output: optimal solution (x^*, y^*) to (3.1)

1: set $i := 1$ and initialize the set $\text{epi}_S(z)^{(1)} := \{(x, \eta) \in S \times \mathbb{R} \mid \eta \geq \bar{\eta}\}$
2: solve the problem $\min\{c^\top x + \eta \mid (x, \eta) \in \text{epi}_S(z)^{(1)}\}$ to obtain $(x^{(1)}, \eta^{(1)})$
3: **while** $(x^{(i)}, \eta^{(i)}) \notin \text{epi}(z)$ **do**
4: find inequality $\pi^\top x + \pi_0 \eta \leq \alpha$ satisfied by all $(x, \eta) \in \text{epi}(z)$ but not by $(x^{(i)}, \eta^{(i)})$
5: $\text{epi}_S(z)^{(i+1)} := \text{epi}_S(z)^{(i)} \cap \{(x, \eta) \mid \pi^\top x + \pi_0 \eta \leq \alpha\}$
6: $i := i + 1$
7: solve the problem $\min\{c^\top x + \eta \mid (x, \eta) \in \text{epi}_S(z)^{(i)}\}$ to obtain $(x^{(i)}, \eta^{(i)})$
8: **end while**
9: set $x^* := x^{(i)}$
10: solve problem (3.4) to compute y^*
11: **return** (x^*, y^*)

Algorithm 1: The Benders decomposition algorithm.

see Proposition 3.2). Otherwise, we add constraints violated by (x, η) but satisfied by all $(x', \eta') \in \text{epi}(z)$ and iterate. This approach is captured in Algorithm 1, the classical Benders decomposition algorithm.

From this perspective, Algorithm 1 is of course just an ordinary cutting plane algorithm, where the crucial step, the selection of a separating cut, happens in line 4. In the following, we will address this *separation subproblem* from two different perspectives: In Section 3.1.1, we present the classical, more algebraic approach to solving the problem, before employing a more geometric perspective in Section 3.1.2. Prior to that, we note that Algorithm 1 actually applies to a slightly more general setting than the one described above, as can easily be seen:

Remark 3.3
Consider the following version of problem (3.1) with a more general objective function:

$$\begin{aligned} \min\ & f(x) + d^\top y \\ \text{s.t. } & Hx + Ay \leq b \\ & x \in S \subset \mathbb{R}^n \\ & y \in \mathbb{R}^k \end{aligned} \tag{3.6}$$

Given an algorithm that can solve optimization problems over the set S for objective functions of the form $f + \bar{z}$ where $\bar{z}$ is a convex piecewise-linear function, Algorithm 1 can be used to solve problems of the above form. In particular, given an algorithm to solve convex piecewise-linear optimization problems over S, we can solve problems of the form (3.6) where f is convex piecewise-linear.

With the objective function no longer constrained to being linear, the question arises whether something similar can be accomplished regarding the constraints $Hx + Ay \leq b$. Indeed, Geoffrion [Geo72] has generalized the method to problems where the constraints linking x- and y-variables are convex, but not necessarily linear. He shows that under certain conditions, the generalized method converges after finitely many iterations. At least theoretically, the approach can further be generalized to arbitrary mathematical programming problems [Wol81], albeit with substantial computational challenges. More recently, Hooker and Ottosson [HO03] have extended the concept of Benders decomposition to a framework that may be applied to other problem domains such as satisfiability problems (see also [Rah+17] for an overview of related literature).

In a separate article, Geoffrion [Geo70] established a classification of different methods for solving large-scale mathematical programming problems by the methods they use to represent the problem and then solve it. Benders decomposition is described as an example for the class of methods that use *projection* to represent the problem and *outer linearization/relaxation* to solve it.

Since the linear setting (3.1) is sufficient for our work, we will focus on this simpler case. Nevertheless, the perspective that we employ as well as many of the results will be equally applicable to the more general non-linear setting outlined in Remark 3.3.

3.1.1 Benders Cuts

Algorithm 1 leaves open the procedure to solve the separation subproblem in line 4, i. e., how to select an inequality which satisfies the specified conditions. Indeed, several different methods have been proposed in the literature. We begin by reviewing the approach used in most of the existing work on Benders decomposition.

This approach, which is also used in the original paper by Benders [Ben62], relies on the following observations about the linear program

$$\begin{aligned} \max\ & \gamma^\top (Hx^* - b) \\ \text{s.t.}\ & -\gamma^\top A = d^\top \\ & \gamma \geq 0, \end{aligned} \tag{3.7}$$

which is the dual of (3.4). Some version of the following lemma lies at the core of any textbook description of Benders decomposition:

Lemma 3.4
Let $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$.

a) If the LP (3.4) *is unbounded, then the original optimization problem* (3.1) *is unbounded.*

b) If the LP (3.4) *is infeasible and* (3.7) *is also infeasible, then the original problem* (3.1) *is infeasible or unbounded.*

c) If the LP (3.4) *is infeasible and* (3.7) *has an unbounded ray* γ *(i. e.,* $\gamma^\top(Hx^* - b) > 0$, $\gamma \geq 0$ *and* $\gamma^\top A = 0$*), then the inequality* $\gamma^\top(Hx - b) \leq 0$ *is satisfied by all* $(x, \eta) \in \operatorname{epi}(z)$*, but violated by* (x^*, η^*)*.*

d) If the LP (3.4) *is feasible and* (3.7) *has an optimal solution* γ *with objective value* η'*, then the inequality* $\gamma^\top(Hx - b) \leq \eta$ *is satisfied by all* $(x, \eta) \in \operatorname{epi}(z)$*. In particular, in this case* $(x^*, \eta^*) \in \operatorname{epi}(z)$ *if and only if* $\eta^* \geq \eta'$*.*

Proof. Part a) follows immediately from the fact that the LP (3.4) is just (3.1) with x fixed to x^*. For part b), observe that if (3.7) is infeasible, then it is infeasible for all values of x^*. The statement then follows from linear duality theory.

In part c), for γ to be a dual unbounded ray, we need to have $\gamma^\top(Hx^* - b) > 0$ which proves that the inequality $\gamma^\top(Hx - b) \leq 0$ is indeed violated by (x^*, η^*). This is true, independently of the value of η^*. On the other hand, for every $(x, \eta) \in \operatorname{epi}(z)$, the problem (3.4) is feasible for $x^* := x$ and hence the dual is bounded. But this means that γ cannot be an unbounded ray, which implies that $\gamma^\top(Hx' - b) \leq 0$, hence the inequality is satisfied by any $(x, \eta) \in \operatorname{epi}(z)$.

Regarding part d), note that varying x^* in (3.7) does not modify the feasible region. Hence, γ is a feasible solution to (3.7) for every x^* and $\gamma^\top(Hx^* - b)$ is a valid lower bound for (3.4). In other words, the cut $\gamma^\top(Hx - b) \leq \eta$ is valid for all $(x, \eta) \in \operatorname{epi}(z)$.

For the second part of the statement, (3.7) and thus (3.4) has an optimal solution with objective value η', which proves that $(x^*, \eta') \in \operatorname{epi}(z)$. By definition of $\operatorname{epi}(z)$, if $\eta^* \geq \eta'$, then $(x^*, \eta^*) \in \operatorname{epi}(z)$. On the other hand, since $\gamma^\top(Hx^* - b) = \eta'$ is a valid lower bound for (3.4), $\eta^* < \eta'$ implies that $(x^*, \eta^*) \notin \operatorname{epi}(z)$. □

The inequality obtained from case c) is commonly referred to as a *feasibility cut*, that obtained from case d) is called an *optimality cut* (see, e. g., [VW10]).

Regarding the two major limitations of the sampling approach mentioned above, note that Lemma 3.4 covers both the case where (3.4) is feasible and where it is infeasible. Note further that, as the dual feasible region of (3.4) does not depend on x^*, only a finite number of extremal dual unbounded rays and dual optimal solutions can ever appear over the course of the algorithm which implies that the algorithm terminates after a finite number of iterations.

While Lemma 3.4 guarantees that we can always solve the problem from line 4 of Algorithm 1 by solving the Linear Program (3.4), this is obviously not the only way and the resulting cut is by no means guaranteed to be the best choice with respect e. g. to the running time of the algorithm. In particular in the case where the LP (3.4) is infeasible, an arbitrary dual unbounded ray is chosen, the selection of which typically depends only on the specific implementation of the LP solver used. In addition, the normal vector of the cut in this case will always lie in the x-space, providing no information about the feasible values of η.

Against this background, it makes sense to take a unified perspective on feasibility cuts and optimality cuts. One way to do this (see, e. g., [FSZ10]) is by viewing the subproblem (3.4) as a pure feasibility problem, represented by the set

$$\left\{ y \in \mathbb{R}^k \;\middle|\; \begin{array}{c} Ay \leq b - Hx^* \\ d^\top y \leq \eta^* \end{array} \right\}. \tag{3.8}$$

This polyhedron will be empty if and only if $(x^*, \eta^*) \notin \operatorname{epi}(z)$ (see (3.5)). Similar to Lemma 3.4 c), we can use a Farkas certificate for emptiness of (3.8) to derive a valid inequality: The polyhedron (3.8) is empty if and only if there exists a vector $(\gamma, \gamma_0) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}_{\geq 0}$ such that

$$\gamma^\top (b - Hx^*) + \gamma_0 \eta^* < 0 \tag{3.9}$$

$$\gamma^\top A + \gamma_0 d^\top = 0 \tag{3.10}$$

If such a (γ, γ_0) exists, then the inequality $\gamma^\top (b - Hx) + \gamma_0 \eta \geq 0$ is obviously violated by the point (x^*, η^*). At the same time, since (3.10) holds independently from (x, η), the inequality $\gamma^\top (b - Hx) + \gamma_0 \eta \geq 0$ must be satisfied by any $(x, \eta) \in \operatorname{epi}(z)$ since the corresponding set (3.8) is non-empty by the definition of $\operatorname{epi}(z)$ (see (3.5)).

Furthermore, since both (3.9) and (3.10) are homogenous, every valid (γ, γ_0) can be scaled to obtain a valid certificate with $\gamma^\top (b - Hx^*) + \gamma_0 \eta^* = -1$. We can thus use the so-called *alternative polyhedron* (see, e. g., [FSZ10]) to find a suitable inequality:

Definition 3.5 (Alternative Polyhedron)
Let z be defined as in (3.2) and $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. The *alternative polyhedron* $P(x^*, \eta^*)$ is given by

$$P(x^*, \eta^*) := \left\{ \gamma, \gamma_0 \geq 0 \;\middle|\; \gamma^\top A + \gamma_0 d^\top = 0, \gamma^\top (b - Hx^*) + \gamma_0 \eta^* = -1 \right\}. \tag{3.11}$$

The set $P(x^*, \eta^*)$ contains (up to scaling) all valid Farkas certificates satisfying (3.9) and (3.10). Therefore, $P(x^*, \eta^*) = \emptyset$ if and only if $(x^*, \eta^*) \in \operatorname{epi}(z)$. Furthermore, every point $(\gamma, \gamma_0) \in P(x^*, \eta^*)$ induces an inequality $\gamma^\top (Hx - b) - \gamma_0 \eta \leq 0$ that is valid for $\operatorname{epi}(z)$ but violated by (x^*, η^*). This will be proved in detail in Corollary 3.8.

3.1.2 Benders Decomposition as a Cutting Plane Algorithm

In this section, we want to present an alternative perspective on the problem of finding a violated inequality, starting from the question of characterizing the normal vectors of such inequalities. Naturally, we will arrive sooner or later at the same set of violated inequalities that is also induced by (3.11). However, we believe that this perspective provides a better understanding for possible criteria to select a cut from that set. The

approach that we employ is similar in spirit to that used by Cornuéjols and Lemaréchal [CL06] in their work on the selection of disjunctive cuts.

It should be noted at this point that we are not the the first ones to notice the similarity between the approaches of Fischetti, Salvagnin, and Zanette [FSZ10] and Cornuéjols and Lemaréchal [CL06]. Indeed, the authors of [FSZ10] cite explicitly the work of Cornuéjols and Lemaréchal [CL06], albeit only in a remark about the possibility to exchange normalization and objective function in optimization problems over the alternative polyhedron (see Corollary 3.21).

Fischetti, Salvagnin, and Zanette [FSZ10] focus on properties of the alternative polyhedron, as well as practical considerations about the implementation of Benders decomposition. They list a number of interesting conclusions from practical experience, which to our knowledge have not been written down clearly anywhere else and make the paper a very interesting read. This is even more true for the more extensive unpublished draft [FSZ09].

However, neither [FSZ09] nor [FSZ10] goes into much detail regarding the exact relation between the the alternative polyhedron and the reverse polar, which underlies the approach of Cornuéjols and Lemaréchal [CL06]. This is a main motivation for our work in this chapter: We will formulate the relation very precisely, which will allow us to gain some new insights about the selection of Benders cuts which help us to improve the performance of practical implementations of the algorithm.

Using the notions from convex geometry defined in Appendix A.3, we are interested in the following: Given an optimization problem of the form (3.1), let the function z be defined as in (3.2). Now, for a tentative solution $(x^*, \eta^*) \in S \times \mathbb{R}$, determine whether there exists a hyperplane that strongly separates (x^*, η^*) from $\text{epi}(z)$. If not, then $(x^*, \eta^*) \in \text{epi}_S(z)$ (since $\text{epi}(z)$ is a polyhedron) and we have found an optimal solution. If, on the other hand, a strongly separating hyperplane exists, we want to find a halfspace that contains $\text{epi}(z)$ but does not contain (x^*, η^*). This halfspace is defined by an inequality valid for $\text{epi}(z)$ which is violated by (x^*, η^*).

Indeed, our strongly separating hyperplane yields one such halfspace which does, however, not support $\text{epi}(z)$. Ideally, though, we would prefer a halfspace that supports $\text{epi}(z)$ to avoid the situation where we obtain a halfspace with the same normal vector again at a later iteration of the algorithm.

Since we are always concerned in this chapter with separating a point x from a convex set C, we use a special definition of an *x-separating halfspace for C* (see Definition A.9) to capture those halfspaces that contain the set C, but do not contain x. On the one hand, this notion is (intentionally!) weaker than strong separation, since the halfspace may support C. On the other hand, since x is a singleton, any x-separating halfspace for C yields a strongly separating hyperplane by shifting its boundary slightly in the direction of x.

We start by applying the above idea to problems of the form (3.1), characterizing the set of normal vectors of (x^*, η^*)-separating halfspaces for $\text{epi}(z)$:

Theorem 3.6
Let z be defined as in (3.2) with $\operatorname{epi}(z) \neq \emptyset$ and let $(x^*, \eta^*), (\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R}$. The vector (π, π_0) is the normal vector of an (x^*, η^*)-separating halfspace for $\operatorname{epi}(z)$ if and only if $\pi_0 \leq 0$ and there exists a vector $\gamma \in \mathbb{R}^m_{\geq 0}$ satisfying

$$(\pi^\top, \pi_0) \begin{pmatrix} x^* \\ \eta^* \end{pmatrix} - \gamma^\top b > 0 \tag{3.12}$$

$$\gamma^\top A - \pi_0 d^\top = 0 \tag{3.13}$$

$$\gamma^\top H = \pi^\top. \tag{3.14}$$

Proof. Let $h_{\operatorname{epi}(z)}$ be the support function of $\operatorname{epi}(z)$ (see Definition A.9). The vector (π, π_0) is the normal vector of an (x^*, η^*)-separating halfspace for $\operatorname{epi}(z)$ if and only if

$$0 < \inf_{(x,\eta)\in\operatorname{epi}(z)} \left\{ (\pi^\top, \pi_0) \left(\begin{pmatrix} x^* \\ \eta^* \end{pmatrix} - \begin{pmatrix} x \\ \eta \end{pmatrix} \right) \right\} = (\pi^\top, \pi_0) \begin{pmatrix} x^* \\ \eta^* \end{pmatrix} - h_{\operatorname{epi}(z)}(\pi, \pi_0). \tag{3.15}$$

By the definition of $\operatorname{epi}(z)$ (which is closed and polyhedral) and then by strong LP duality, we obtain

$$h_{\operatorname{epi}(z)}(\pi, \pi_0) = \max_{\substack{x\in\mathbb{R}^n, y\in\mathbb{R}^k \\ \eta\in\mathbb{R}}} \left\{ (\pi^\top, \pi_0) \begin{pmatrix} x \\ \eta \end{pmatrix} \;\middle|\; \begin{array}{r} Ay \leq b - Hx \\ d^\top y \leq \eta \end{array} \right\} \tag{3.16}$$

$$= \min_{\substack{\gamma_0\in\mathbb{R}_{\geq 0} \\ \gamma\in\mathbb{R}^m_{\geq 0}}} \left\{ \gamma^\top b \;\middle|\; \begin{array}{r} \gamma^\top A + \gamma_0 d^\top = 0 \\ \gamma^\top H = \pi^\top \\ -\gamma_0 = \pi_0 \end{array} \right\}. \tag{3.17}$$

Note that in order for the equality $-\gamma_0 = \pi_0$ to hold and (3.17) to be feasible (and hence (3.16) to be bounded), we need that $\pi_0 \leq 0$. For any $\gamma \geq 0$ satisfying the conditions (3.12) to (3.14), we thus have $\gamma^\top b \geq h_{\operatorname{epi}(z)}(\pi, \pi_0)$. Inequality (3.12) then implies that (3.15) is satisfied, which proves the claim. □

Now that Theorem 3.6 provides us with a complete characterization of possible cut normals, we can try to choose among them cuts that are *good* in a certain sense. We saw in the proof that for any γ satisfying (3.13) and (3.14), $\gamma^\top b$ is an upper bound for the support function $h_{\operatorname{epi}(z)}$ of $\operatorname{epi}(z)$. This means that once we obtain a certificate γ to prove that a vector (π, π_0) belongs to an (x^*, η^*)-separating halfspace $H^{\leq}_{((\pi,\pi_0),\alpha)}$, we immediately obtain a corresponding right hand side $\alpha := \gamma^\top b$.

Furthermore, the definition of the support function $h_{\operatorname{epi}(z)}$ immediately tells us when this right-hand side is actually optimal and the resulting halfspace supports $\operatorname{epi}(z)$:

Remark 3.7
Let $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$ and let (π, π_0) be the normal vector of an (x^*, η^*)-separating halfspace for epi(z). If γ minimizes $\gamma^\top b$ among all possible certificates in Theorem 3.6, then the halfspace $H^{\leq}_{((\pi,\pi_0),\gamma^\top b)}$ supports the set epi(z).

This means that, given a suitable normal vector (π, π_0) for which an (x^*, η^*)-separating halfspace exists, we can compute the corresponding supporting inequality by solving the problem (3.17) or (3.16). We further note that the set defined by (3.12) to (3.14) is homogenous: If γ is a valid certificate for (π, π_0), then $\lambda\gamma$ is a valid certificate for $\lambda \cdot (\pi, \pi_0)$.

3.1.3 Alternative Polyhedron and Reverse Polar Set

Using Theorem 3.6 and the above observations, we can immediately confirm the validity of the approach from [FSZ10], which uses the alternative polyhedron from Definition 3.5. In fact, we can prove even more: While it would be sufficient to be able to obtain an *arbitrary* (x^*, η^*)-separating halfspace from the alternative polyhedron whenever (3.8) is empty, it turns out that the alternative polyhedron $P(x^*, \eta^*)$ actually completely characterizes the set of *all* possible normal vectors of such halfspaces:

Corollary 3.8
The alternative polyhedron (3.11) *completely characterizes all normal vectors of* (x^*, η^*)*-separating halfspaces for* epi(z)*. In particular:*

a) Let $(\gamma, \gamma_0) \in P(x^*, \eta^*)$*. Then* $\gamma^\top Hx - \gamma_0\eta \leq \gamma^\top b$ *is violated by* (x^*, η^*)*, but satisfied by all* $(x, \eta) \in$ epi(z)*.*

b) Let (π, π_0) *be the normal vector of an* (x^*, η^*)*-separating halfspace for* epi(z)*. Then there exist* $(\gamma, \gamma_0) \in P(x^*, \eta^*)$ *and* $\lambda \geq 0$ *such that* $(\gamma^\top H, -\gamma_0) = \lambda \cdot (\pi, \pi_0)$*.*

Observe, however, that in contrast to Remark 3.7, Corollary 3.8 does not guarantee that the cut generated from a point in the alternative polyhedron is supporting: A given vector $(\gamma, \gamma_0) \in P(x^*, \eta^*)$ does not necessarily minimize $\gamma^\top b$ among all points in $P(x^*, \eta^*)$ which lead to the same cut normal. Indeed, as the following example shows, there are cases where this actually occurs in practice and even a cut generated from an optimal vertex of the alternative polyhedron may not be supporting.

Example 3.9
Consider the following optimization problem:

$$\begin{aligned} \min\ & x + y \\ & y + 2x \geq 5 \\ & y + \frac{x}{2} \geq 3 \\ & 4y + 4x \geq 14 \end{aligned} \tag{3.18}$$

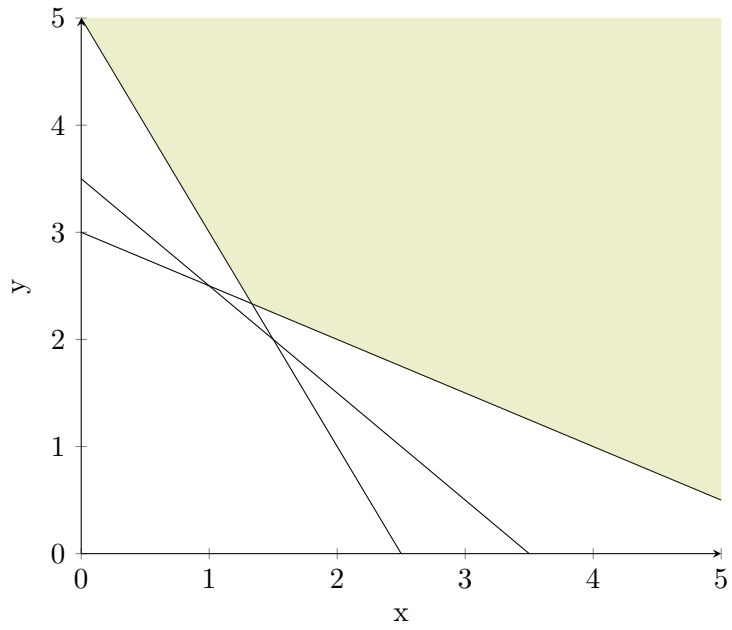


Figure 3.2: Constraints and feasible region for the optimization problem from Example 3.9.

The constraints as well as the feasible region are shown in Fig. 3.2. Note that the third constraint $4y + 4x \geq 14$ is redundant and does not support the feasible region. Suppose that we want to decompose the problem into its x-part and its y-part. To obtain the alternative polyhedron for a tentative master solution (x^*, η^*), we rewrite the subproblem in the way of (3.8) as

$$\left\{ y \in \mathbb{R} \;\middle|\; \begin{array}{c} -y \leq -5 - (-2x^*) \\ -y \leq -3 - (-\frac{1}{2}x^*) \\ -4y \leq -14 - (-4x^*) \\ y \leq \eta^* \end{array} \right\}. \tag{3.19}$$

The alternative polyhedron $P(x^*, \eta^*)$ is then given by

$$P(x^*, \eta^*) := \left\{ \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_0 \end{pmatrix} \geq 0 \;\middle|\; \begin{array}{c} \gamma_0 - \gamma_1 - \gamma_2 - 4\gamma_3 = 0 \\ \gamma_0\eta^* + \gamma_1(-5 + 2x^*) + \gamma_2\left(-3 + \frac{1}{2}x^*\right) + \gamma_3(-14 + 4x^*) = -1 \end{array} \right\}.$$

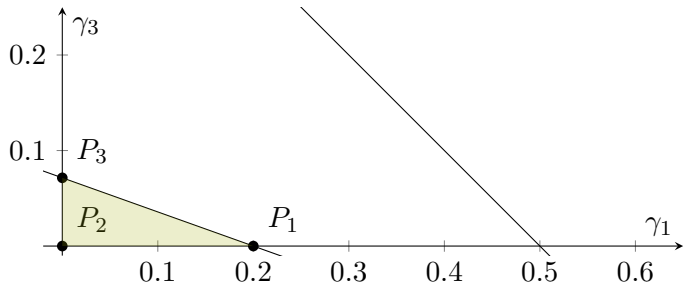
Assuming that $x^* + 2\eta^* \neq 6$, we can reformulate the equality constraints to

$$\gamma_0 = \gamma_1 + \gamma_2 + 4\gamma_3 = \frac{1 + \gamma_1(-2 + \frac{3}{2}x^*) + \gamma_3(-2 + 2x^*)}{3 - \frac{1}{2}x^* - \eta^*}$$
$$\gamma_2 = \frac{1 + \gamma_1(-5 + 2x^* + \eta^*) + \gamma_3(-14 + 4x^* + 4\eta^*)}{3 - \frac{1}{2}x^* - \eta^*}$$

As an example, let $(x^*, \eta^*) := (0, 0)$. We obtain

$$\gamma_0 = \frac{1 - 2\gamma_1 - 2\gamma_3}{3}$$
$$\gamma_2 = \frac{1 - 5\gamma_1 - 14\gamma_3}{3}$$

We can visualize the two-dimensional alternative polyhedron via its projection into the γ_1-γ_3-plane:



Note that, in order to be consistent with the notation in Definition 3.5, we will write the components for points in $P(x^*, \eta^*)$ in the order $(\gamma_1, \gamma_2, \gamma_3, \gamma_0)$. In this notation, the three extremal points of $P(0, 0)$ are

$$P_1 = \left(\frac{1}{5}, 0, 0, \frac{1}{5}\right)^\top$$
$$P_2 = \left(0, \frac{1}{3}, 0, \frac{1}{3}\right)^\top$$
$$P_3 = \left(0, 0, \frac{1}{14}, \frac{2}{7}\right)^\top.$$

We can see that for each of these points, as shown by Gleeson and Ryan [GR90], the set of inequalities for which the corresponding dual variable is positive represents a minimal infeasible subsystem of (3.19). Consequently, each extremal point yields one of the original inequalities as a cut. This notably includes the redundant inequality $4y + 4x \geq 14$, which does not support the feasible region but is derived from the extremal point P_3 in the alternative polyhedron.

One major advantage of the alternative polyhedron as opposed to working directly with the set (3.12) to (3.14) is that the alternative polyhedron is closed. On the other hand, it is a set of dual vectors which makes it more difficult to interpret and relate to properties of the set $\text{epi}_S(z)$ that we are interested in.

Alternatively, as argued by Cornuéjols and Lemaréchal [CL06], we can obtain a different closed set from (3.12) to (3.14) by replacing the strict inequality (3.12) by $(\pi^\top, \pi_0)(x^*, \eta^*)^\top - \gamma^\top b \geq 1$. The resulting set is called the *reverse polar set* of $\text{epi}(z) - (x^*, \eta^*)$ as introduced by Balas [Bal79], which is defined as follows:

Definition 3.10
Let $C \subset \mathbb{R}^n$ be a convex set. The *reverse polar set* $C^- \subset \mathbb{R}^n$ is defined as

$$C^- := \left\{ c \in \mathbb{R}^n \;\middle|\; c^\top x \leq -1 \;\forall\; x \in C \right\}.$$

Remark 3.11
The terminology explains itself if we remember that the *polar set* $C^\circ \subset \mathbb{R}^n$ of a convex set $C \subset \mathbb{R}^n$ is defined as

$$C^\circ := \left\{ c \in \mathbb{R}^n \;\middle|\; c^\top x \leq 1 \;\forall\; x \in C \right\}.$$

Observe furthermore that it holds that $C^- \subset \text{cl}(\text{pos}(C^-)) = \text{cl}(\text{pos}(C))^\circ \subset C^\circ$ for all $C \subset \mathbb{R}^n$.

Note that by Definition 3.10, the reverse polar set is given by an intersection of halfspaces (albeit infinitely many), an *$\mathcal{H}$-representation*. If C is a polyhedron, it is actually sufficient to consider those halfspaces that correspond to vertices of C, but nonetheless an $\mathcal{H}$-representation of C^- cannot be efficiently computed in general for a polyhedron C given in $\mathcal{H}$-representation (since computing all the vertices of C is $\mathbb{NP}$-hard in general [Kha+08]).

Even without an explicit $\mathcal{H}$-representation of the set $\text{epi}(z)$ (which is itself known to us only by its extended formulation (3.5)), we can use Theorem 3.6 and the fact that the set defined by (3.12) to (3.14) is homogenous, to easily obtain the following description of the reverse polar set $(\text{epi}(z) - (x^*, \eta^*))^-$:

Let a problem of the form (3.1) and a point $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$ be given. The reverse polar set of $\text{epi}(z) - (x^*, \eta^*)$ is given by

$$(\text{epi}(z) - (x^*, \eta^*))^- := \left\{ (\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R}_{\leq 0} \;\middle|\; \exists \gamma \in \mathbb{R}^m_{\geq 0} : \begin{array}{r} (\pi^\top, \pi_0) \begin{pmatrix} x^* \\ \eta^* \end{pmatrix} - \gamma^\top b \geq 1 \\ \gamma^\top A - \pi_0 d^\top = 0 \\ \gamma^\top H = \pi^\top \end{array} \right\}. \quad (3.20)$$

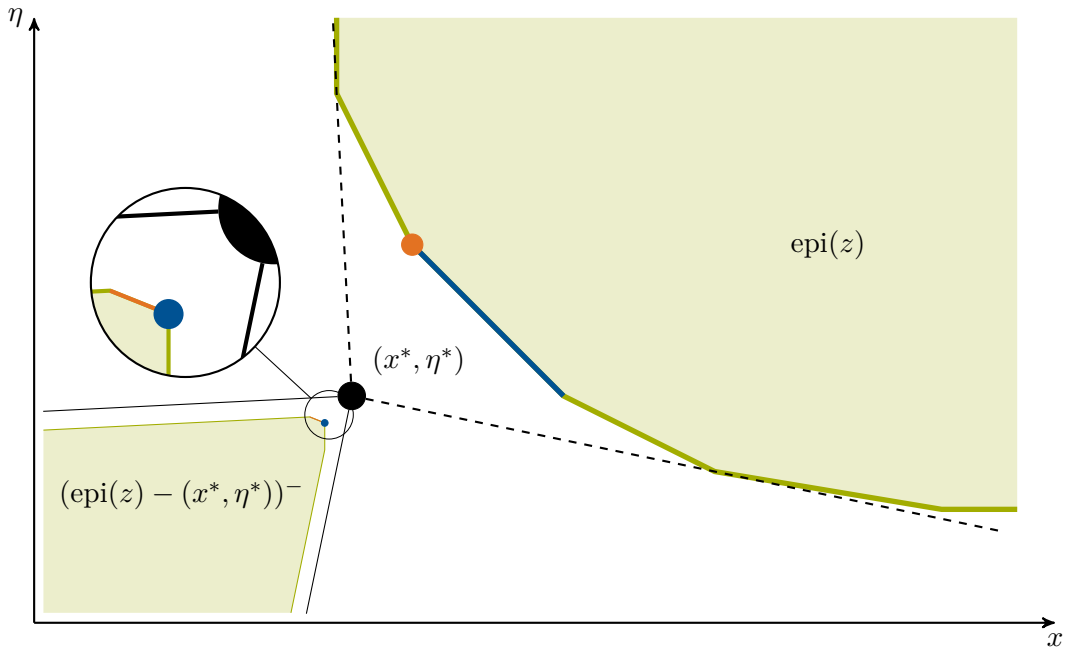


Figure 3.3: The reverse polar set $(\text{epi}(z) - (x^*, \eta^*))^-$ and the corresponding polar cone (drawn in a coordinate system with (x^*, η^*) as the origin). It can be seen that $(\text{epi}(z) - (x^*, \eta^*))^-$ is contained in the polar cone $\text{pos}(\text{epi}(z) - (x^*, \eta^*))^\circ$ (indicated by the black solid lines) but offers a "richer" boundary from which we can choose cut normals. Specifically, facets and vertices of $(\text{epi}(z) - (x^*, \eta^*))^-$ correspond to vertices and facets, respectively, of epi(z) that a cut with the corresponding normal vector can support (see Theorem 3.30).

The reverse polar set thus contains exactly those directions that we obtained as possible cut normals from Theorem 3.6 (see Fig. 3.3): For every point in $(\text{epi}(z) - (x^*, \eta^*))^-$, there obviously exists $\gamma \in \mathbb{R}^m_{\geq 0}$ such that (3.12) to (3.14) is satisfied. Conversely, any pair (π, π_0) satisfying (3.12) to (3.14) can be scaled by an appropriate positive factor to obtain a point in $(\text{epi}(z) - (x^*, \eta^*))^-$, scaling the corresponding certificate γ by the same factor.

We thus have at our disposal two alternative characterizations of the set of possible normal vectors of (x^*, η^*)-separating halfspaces: The alternative polyhedron and the reverse polar set. Despite their similarity, subtle differences exist between both representations that affect their usefulness for the generation of Benders cuts.

Before we proceed, we introduce a variant of the alternative polyhedron, the *relaxed alternative polyhedron*, which also appears in [GR90]. We will see that it is equivalent to the original alternative polyhedron for almost all purposes, but can more easily be connected to the reverse polar set:

Definition 3.12
Let a problem of the form (3.1) and a point $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$ be given. The *relaxed alternative polyhedron* $P^{\leq}(x^*, \eta^*)$ is defined as

$$P^{\leq}(x^*, \eta^*) := \left\{ \gamma, \gamma_0 \geq 0 \;\middle|\; \begin{array}{r} \gamma^\top A + \gamma_0 d^\top = 0 \\ \gamma^\top (b - Hx^*) + \gamma_0 \eta^* \leq -1 \end{array} \right\}. \tag{3.21}$$

With this definition, the following theorem characterizes the relation between alternative polyhedron and reverse polar set.

Theorem 3.13
Let z be defined as in (3.2) and $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. Then

$$(\mathrm{epi}(z) - (x^*, \eta^*))^- = \begin{pmatrix} H^\top & 0 \\ 0 & -1 \end{pmatrix} \cdot P^{\leq}(x^*, \eta^*).$$

Proof.

$$\begin{aligned} \begin{pmatrix} H^\top & 0 \\ 0 & -1 \end{pmatrix} P^{\leq}(x^*, \eta^*) &= \left\{ (H^\top \gamma, -\gamma_0) \;\middle|\; \begin{array}{r} \gamma, \gamma_0 \geq 0 \\ \gamma^\top A + \gamma_0 d^\top = 0 \\ \gamma^\top (b - Hx^*) + \gamma_0 \eta^* \leq -1 \end{array} \right\} \\ &= \left\{ (\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0} \;\middle|\; \exists \gamma \geq 0 : \begin{array}{r} H^\top \gamma = \pi \\ \gamma^\top A - \pi_0 d^\top = 0 \\ \gamma^\top b - \pi^\top x^* - \pi_0 \eta^* \leq -1 \end{array} \right\} \\ &= (\mathrm{epi}(z) - (x^*, \eta^*))^- \end{aligned}$$

□

We revisit Example 3.9 to illustrate this observation.

Example 3.9 (continued)
In the situation of the optimization problem (3.18), observe that the relaxed alternative polyhedron $P^{\leq}(x^*, \eta^*)$ is

$$P^{\leq}(x^*, \eta^*) := \left\{ \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_0 \end{pmatrix} \geq 0 \;\middle|\; \begin{array}{c} \gamma_0 - \gamma_1 - \gamma_2 - 4\gamma_3 = 0 \\ \gamma_0 \eta^* + \gamma_1(-5 + 2x^*) + \gamma_2 \left(-3 + \frac{1}{2} x^* \right) + \gamma_3(-14 + 4x^*) \leq -1 \end{array} \right\}$$

Assuming that $x^* + 2\eta^* < 6$, we can again reformulate the constraints to

$$\gamma_0 = \gamma_1 + \gamma_2 + 4\gamma_3 \tag{3.22}$$

$$\gamma_2 \geq \frac{1 + \gamma_1(-5 + 2x^* + \eta^*) + \gamma_3(-14 + 4x^* + 4\eta^*)}{3 - \frac{1}{2}x^* - \eta^*} \tag{3.23}$$

which, again, for $(x^*, \eta^*) := (0, 0)$ simplifies to

$$\gamma_2 \geq \frac{1 - 5\gamma_1 - 14\gamma_3}{3} \tag{3.24}$$

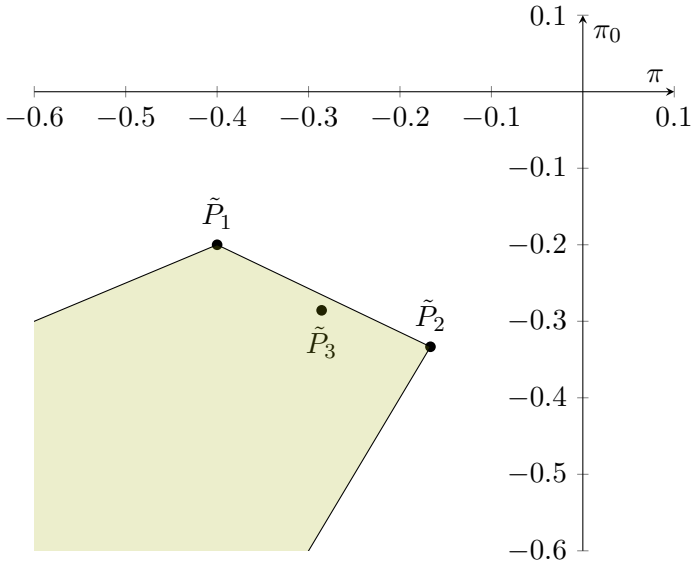
The relaxed alternative polyhedron $P^{\leq}(0,0)$ is hence a three-dimensional unbounded polyhedron with extremal rays $\text{pos}(0,1,0,1), \text{pos}(1,0,0,1), \text{pos}(0,0,1/4,1)$ and with the original alternative polyhedron $P(0,0)$ as the only bounded facet. The latter can be seen as follows: If in any extremal point the inequality (3.24) does not hold with equality, then at least three of the non-negativity constraints must be tight. By (3.22), this means that the fourth must be tight, as well, but the resulting point violates (3.24). Hence, in all vertices of $P^{\leq}(0,0)$, the inequality (3.24) must be tight, which implies that the relaxed alternative polyhedron $P^{\leq}(0,0)$ has the same extremal points as $P(0,0)$. Writing

$$T := \begin{pmatrix} H^\top & 0 \\ 0 & -1 \end{pmatrix},$$

we can use Theorem 3.13 to derive the reverse polar set:

$$\begin{aligned}
\text{epi}(z)^- &= (\text{epi}(z) - (0,0))^- = TP^{\leq}(0,0) \\
&= \text{conv}(TP_1, TP_2, TP_3) + \text{pos}(T(0,1,0,1), T(1,0,0,1), T(0,0,1/4,1)) \\
&= \text{conv}\left(\left(-\frac{2}{5}, -\frac{1}{5}\right), \left(-\frac{1}{6}, -\frac{1}{3}\right), \left(-\frac{2}{7}, -\frac{2}{7}\right)\right) \\
&\qquad + \text{pos}\left(\left(-\frac{1}{2}, -1\right), \left(-2, -1\right), \left(-1, -1\right)\right)
\end{aligned}$$

The set $\text{epi}(z)^-$ is visualized below:



We can see that the point P_3, which lead to the non-supporting cut above, is mapped to the interior of the reverse polar set and will hence not appear as an extremal solution.

The fact that in our example every vertex of the reverse polar set generates a supporting cut is no coincidence: Indeed, for every point in the reverse polar set that is *visible from the origin*, the right hand side obtained from a valid γ in the definition (3.20) always yields a supporting cut. For points that are not visible from the origin, this is not necessarily true, but even there a vector γ that leads to a supporting cut always exists.

The following theorem provides us with a sufficient criterion for this property. This criterion will be particularly useful in the context of cut-generating linear programs, which we will consider in Section 3.1.4 below.

Theorem 3.14
Let $(\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$ and let $(\pi, \pi_0) \in (\text{epi}(z) - (x^*, \eta^*))^-$ be maximal with respect to the objective (ω, ω_0) with $\omega^\top \pi + \omega_0 \pi_0 < 0$. Furthermore, let $\gamma \in \mathbb{R}^m_{\geq 0}$ be a valid certificate for (π, π_0) in (3.20). Then the halfspace $H^{\leq}_{((\pi,\pi_0),\gamma^\top b)}$ supports epi(z).

Proof. By Remark 3.7, the statement is true if γ minimizes $\gamma^\top b$ among all possible certificates for the vector (π, π_0) in Theorem 3.6. It is easy to verify that γ is indeed a valid certificate for (π, π_0) in Theorem 3.6. For contradiction, we hence assume that it does not minimize $\gamma^\top b$. Let $\gamma' \geq 0$ be an alternative certificate for (π, π_0) with $\gamma'^\top b < \gamma^\top b$. Then from (3.12) to (3.14) we obtain that $\gamma'^\top A - \pi_0 d^\top = 0$ and $\gamma'^\top H = \pi^\top$.

Furthermore,

$$\pi^\top x^* + \pi_0 \eta^* - \gamma'^\top b > \pi^\top x^* + \pi_0 \eta^* - \gamma^\top b \geq 1.$$

We can thus scale both (π, π_0) and γ' by an appropriate factor $\lambda \in (0, 1)$ to obtain that $\lambda \cdot (\pi, \pi_0) \in (\text{epi}(z) - (x^*, \eta^*))^-$ (as certified by the vector $\lambda \cdot \gamma'$ in (3.20)) and

$$(\omega^\top, \omega_0)(\lambda \cdot (\pi, \pi_0)) = \lambda \cdot \underbrace{\left(\omega^\top \pi + \omega_0 \pi_0\right)}_{<0} > \omega^\top \pi + \omega_0 \pi_0,$$

a contradiction to optimality of (π, π_0). □

In the next section we provide some answers as to when the conditions of the above theorem can be assumed to hold. Furthermore, we derive a method of using the original alternative polyhedron while guaranteeing that every cut that is generated corresponds to an optimal point in the reverse polar set in the sense of Theorem 3.14. In particular, this means that every cut that is generated will support the set epi(z).

3.1.4 Cut-Generating Optimization Problems

One way to select a particular cut normal from the reverse polar set or the alternative polyhedron is by maximizing a linear objective function over these sets. As both the reverse polar set and the relaxed alternative polyhedron are unbounded, however, there are objective directions for which no (finite) optimal solution exists. Cornuéjols and Lemaréchal [CL06, Theorem 2.3] establish the criteria on the objective function for which optimization problems over the reverse polar set are bounded. We have rephrased the relevant parts of the theorem according to our terminology below.

Theorem 3.15 (Cornuéjols and Lemaréchal [CL06, Theorem 2.3])
Let $(x^*, \eta^*) \notin \operatorname{epi}(z)$, $(\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$, *and*

$$z^* := \max\Big\{\omega^\top \pi + \omega_0 \pi_0 \;\Big|\; (\pi, \pi_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))^-\Big\}.$$

Then

$$z^* \begin{cases} \leq 0 & \text{if } (\omega, \omega_0) \in \operatorname{cl}(\operatorname{pos}(\operatorname{epi}(z) - (x^*, \eta^*))) \\ = +\infty & \text{otherwise.} \end{cases}$$

Furthermore, if $(\omega, \omega_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))$, *then* $z^* \leq -1$.

Note in particular that the last part of the above statement implies that $z^* < 0$ whenever $(\omega, \omega_0) \in \operatorname{pos}(\operatorname{epi}(z) - (x^*, \eta^*)) \setminus \{0\}$, which provides us with a large variety of objective functions for which $\omega^\top \gamma + \omega_0 \gamma_0 < 0$ in the optimal solution. By Theorem 3.14, this means that the cut which results from maximizing these objectives over the reverse polar set is guaranteed to be supporting. A similar statement holds with respect to the relaxed alternative polyhedron:

Lemma 3.16
Let z *be defined as in* (3.2) *and* $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. *Then*

a) $P^{\leq}(x^*, \eta^*) = \emptyset$ *if and only if* $(x^*, \eta^*) \in \operatorname{epi}(z)$

b) For any $\bar{c} \in \mathbb{R}^m \times \mathbb{R}$, $\max\{\bar{c}^\top(\gamma, \gamma_0) \mid (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\} \in \mathbb{R}_{\leq 0} \cup \{\infty\}$.

Proof. The first statement follows immediately from the respective property of the original alternative polyhedron. For the second statement, observe that $P^{\leq}(x^*, \eta^*)$ is contained in its recession cone, which is given by

$$P^{\leq}(x^*, \eta^*) \subset \left\{\gamma, \gamma_0 \geq 0 \;\middle|\; \begin{array}{r} \gamma^\top A + \gamma_0 d^\top = 0 \\ \gamma^\top (b - Hx^*) + \gamma_0 \eta^* \leq 0 \end{array}\right\} =: \operatorname{rec}(P^{\leq}(x^*, \eta^*)).$$

But then

$$\max\Big\{\bar{c}^\top(\gamma, \gamma_0) \;\Big|\; (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\Big\} \leq \max\Big\{\bar{c}^\top(\gamma, \gamma_0) \;\Big|\; (\gamma, \gamma_0) \in \operatorname{rec}(P^{\leq}(x^*, \eta^*))\Big\}.$$

Now, let $z_{\bar{c}} := \max\{\bar{c}^\top(\gamma, \gamma_0) \mid (\gamma, \gamma_0) \in \operatorname{rec}(P^{\leq}(x^*, \eta^*))\}$. Since $\operatorname{rec}(P^{\leq}(x^*, \eta^*))$ is a cone, it holds that $z_{\bar{c}} \in \{0, \infty\}$. In the case of $z_{\bar{c}} = 0$, this proves the statement. In the case of $z_{\bar{c}} = \infty$, it follows by definition of the recession cone that $\max\{\bar{c}^\top(\gamma, \gamma_0) \mid (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\} = \infty$, as well, which concludes the proof. □

Theorem 3.15 and Lemma 3.16 already suggest that optimization problems over the relaxed alternative polyhedron and the reverse polar set behave very similarly with respect to existence of finite optima. Indeed, Theorem 3.13 implies a much more precise relation between optimization problems over both sets.

Theorem 3.17
Let z be defined as in (3.2), $(x^*, \eta^*), (\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$ and

$$(\tilde{\omega}, \tilde{\omega}_0)^\top := (H\omega, -\omega_0)^\top. \tag{3.25}$$

Then (π, π_0) is an optimal solution to the problem

$$\max\Big\{\omega^\top \pi + \omega_0 \pi_0 \;\Big|\; (\pi, \pi_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))^-\Big\} \tag{3.26}$$

if and only if there exists γ^* such that $H^\top \gamma^* = \pi$ and $(\gamma^*, -\pi_0)$ is an optimal solution to the problem

$$\max\Big\{\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 \;\Big|\; (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\Big\} \tag{3.27}$$

Furthermore, the objective values of both optimization problems are identical.

Proof. Let (π, π_0) be an optimal solution to (3.26). By Theorem 3.13, there exists a vector γ with $H^\top \gamma = \pi$ such that $(\gamma, -\pi_0) \in P^{\leq}(x^*, \eta^*)$. Suppose that (γ, γ_0) is not optimal for (3.27). Then there exists $(\gamma', \gamma_0') \in P^{\leq}(x^*, \eta^*)$ with $\tilde{\omega}^\top \gamma' + \tilde{\omega}_0 \gamma_0' > \tilde{\omega}^\top \gamma + \tilde{\omega}_0(-\pi_0)$. But again, by Theorem 3.13, $(H^\top \gamma', -\gamma_0') \in (\operatorname{epi}(z) - (x^*, \eta^*))^-$ and

$$\begin{aligned} \omega^\top (H^\top \gamma') + \omega_0(-\gamma_0') &= (H\omega)^\top \gamma' - \omega_0 \gamma_0' = \tilde{\omega}^\top \gamma' + \tilde{\omega}_0 \gamma_0' \\ &> \tilde{\omega}^\top \gamma + \tilde{\omega}_0(-\pi_0) = \omega^\top H^\top \gamma - \omega_0(-\pi_0) = \omega^\top \pi + \omega_0 \pi_0, \end{aligned} \tag{3.28}$$

a contradiction to optimality of (π, π_0).

Similarly, let (γ, γ_0) be an optimal solution to (3.27). Let $\pi := H^\top \gamma$ and $\pi_0 := -\gamma_0$, then by Theorem 3.13, $(\pi, \pi_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))^-$. Now, suppose that (π, π_0) is not optimal for (3.26). Then there exists $(\pi', \pi_0') \in (\operatorname{epi}(z) - (x^*, \eta^*))^-$ with $\omega^\top \pi' + \omega_0 \pi_0' > \omega^\top \pi + \omega_0 \pi_0$. By Theorem 3.13, there exists γ' with $H^\top \gamma' = \pi'$ such that $(\gamma', -\pi_0') \in P^{\leq}(x^*, \eta^*)$. Furthermore,

$$\begin{aligned} \tilde{\omega}^\top \gamma' + \tilde{\omega}_0(-\pi_0') &= (H\omega)^\top \gamma' - \omega_0(-\pi_0') = \omega^\top \pi' + \omega_0 \pi_0' \\ &> \omega^\top \pi + \omega_0 \pi_0 = \omega^\top H^\top \gamma - \omega_0 \gamma_0 = \tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0, \end{aligned} \tag{3.29}$$

a contradiction to optimality of (γ, γ_0). □

Depending on the particular application, the structure of the matrix H can vary in many ways, but in line with our assumption that the *master* problem should be significantly smaller than the *subproblem*, it is reasonable to assume that H has more rows than columns (as will be the case in our application to energy systems).

In this sense, the relaxed alternative polyhedron can be understood as an *extended formulation* for the reverse polar set, which in particular is always polynomial in size. It allows us to generate Benders cuts from points in the reverse polar set while algorithmically relying on the relaxed alternative polyhedron, an explicit description of which is generally trivial to obtain.

This observation establishes an interesting link with [CW18], which deals with separation problems for which either the set from which we want to separate or the set of normal vectors and right-hand-sides of separating inequalities is described in an extended space. Benders decomposition is mentioned as an example of the former and we will revisit the connection with [CW18] later in the context of our discussion of facet-defining Benders cuts.

Note that the problem (3.27) is technically more general than (3.26), since there is no reason *a priori* to limit ourselves to objective functions of the form (3.25). If we choose a different objective function, we still obtain a valid cut. However, since there may be no objective function (ω, ω_0) such that the resulting cut normal is optimal for (3.26), we lose some of the properties associated with optimal solutions from the reverse polar set.

Indeed, this is the approach that Fischetti, Salvagnin, and Zanette [FSZ10] take: They use the problem (3.27) with $\tilde{\omega}_m = 0$ for all m that correspond to rows of zeros in the interaction matrix H, $\tilde{\omega}_m = 1$ for all other m and $\tilde{\omega}_0 = 1$ (or some other manual scaling factor). In general, there exists no vector (ω, ω_0) such that this choice can be obtained by (3.25).

Before we return to Example 3.9 to illuminate this issue further, we note that optimization problems over the original and the relaxed alternative polyhedron are equivalent, provided that the optimization problem over the relaxed alternative polyhedron has a finite non-zero optimum:

Remark 3.18
Let z be defined as in (3.2) and let $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. Let $(\tilde{\omega}, \tilde{\omega}_0) \in \mathbb{R}^m \times \mathbb{R}$ be such that $\max\{\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 \mid \gamma, \gamma_0 \in P^{\leq}(x^*, \eta^*)\} < 0$. Then the sets of optimal solutions for $\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0$ over $P^{\leq}(x^*, \eta^*)$ and $P(x^*, \eta^*)$ are identical. Furthermore, every vertex of $P^{\leq}(x^*, \eta^*)$ is also a vertex of $P(x^*, \eta^*)$.

We now take a closer look at the role of objective functions in the context of Example 3.9:

Example 3.9 (continued)
In the situation of the optimization problem (3.18), remember that the alternative

polyhedron $P(0,0)$ is given by

$$P(0,0) := \left\{ \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_0 \end{pmatrix} \geq 0 \;\middle|\; \begin{array}{c} \gamma_0 - \gamma_1 - \gamma_2 - 4\gamma_3 = 0 \\ \gamma_1 \cdot (-5) + \gamma_2 \cdot (-3) + \gamma_3 \cdot (-14) = -1 \end{array} \right\}$$

with the three extremal points

$$\begin{aligned} P_1 &= \left(\frac{1}{5}, 0, 0, \frac{1}{5}\right) \\ P_2 &= \left(0, \frac{1}{3}, 0, \frac{1}{3}\right) \\ P_3 &= \left(0, 0, \frac{1}{14}, \frac{2}{7}\right). \end{aligned}$$

Note that P_3 actually minimizes the 1-norm over $P(0,0)$ and is hence the unique result of the selection procedure by Fischetti, Salvagnin, and Zanette [FSZ10]. On the other hand, remember that the point P_3, which lead to the non-supporting cut above, is mapped to the interior of the reverse polar set and will hence never appear as an extremal solution.

Instead of dealing directly with the reverse polar set, we can, as observed in Theorem 3.17, almost always achieve the same result by optimizing over the alternative polyhedron. We only have to make sure that the objective function that we use can be written in the form $(H\omega, -\omega_0)^\top$. In our example, we obtain the following set of possible objective functions:

$$\left\{ (H\omega, -\omega_0)^\top \;\middle|\; (\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R} \right\} = \left\{ \left(\begin{pmatrix} -2 \\ -1/2 \\ -4 \end{pmatrix} \cdot \omega, -\omega_0 \right)^\top \;\middle|\; \omega, \omega_0 \in \mathbb{R} \right\}$$

It can be verified using the description of the alternative polyhedron above that if (ω, ω_0) is chosen such that $\operatorname{epi}(z)^-$ is bounded in its direction, then the point $P_3 \in P(0,0)$ is never optimal with respect to an objective of the form $(H\omega, -\omega_0)$, just as $\tilde{P}_3 = T \cdot P_3$, being an internal point of $\operatorname{epi}(z)^-$, is never optimal for any linear objective over the reverse polar set.

One interesting difference between the alternative polyhedron and the reverse polar set, which can be verified using the above example, is their different behavior with respect to algebraic operations on the set of inequalities: If, for instance, we scale one of the inequalities by a positive factor, the reverse polar set remains the same (just as the feasible region defined by the set of inequalities does not change). The alternative

polyhedron, on the other hand, is distorted in response to the scaling of the system of inequalities. If an objective function is used which does not take into account that scaling, such as the vector of zeros and ones that Fischetti, Salvagnin, and Zanette [FSZ10] use, then the selected cut might change depending on the scaling factor. This difference emphasizes the intuition that the reverse polar set captures the geometric properties of the feasible set as a polytope, whereas the alternative polyhedron, as a set of Farkas certificates, is based on the algebraic properties of the feasible region as the solution set of a system of inequalities.

From a practical perspective, the above example shows that if we do not select the objective function for a problem of the type (3.27) carefully, then the cut generated according to Corollary 3.8 from a resulting optimal solution might not even be supporting. This is avoided by selecting the objective function according to (3.25).

On the other hand, Theorem 3.17 shows that with an appropriate choice of the objective function, we can solve any cut-generating optimization problem over the reverse polar set by solving a corresponding problem over the relaxed alternative polyhedron. In particular this means that we never need to obtain an explicit representation of the reverse polar set, which might not be readily available. These results are summarized in the following theorem:

Theorem 3.19
Let z be defined as in (3.2) and let $(x^*, \eta^*), (\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$. Let $(\tilde{\omega}, \tilde{\omega}_0)^\top := (H\omega, -\omega_0)^\top$ and let $(\gamma, \gamma_0) \in P(x^*, \eta^*)$ be maximal with respect to the objective $(\tilde{\omega}, \tilde{\omega}_0)$ with $\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 < 0$. Then the inequality $\gamma^\top Hx - \gamma_0 \eta \leq \gamma^\top b$ supports epi(z).

Proof. By Remark 3.18, (γ, γ_0) maximizes the objective $(\tilde{\omega}, \tilde{\omega}_0)$ over the set $P^{\leq}(x^*, \eta^*)$, as well. Using Theorem 3.17, this implies that $(\pi, \pi_0) := (H^\top \gamma, -\gamma_0)$ is an optimal solution with respect to the objective (ω, ω_0) over the set $(\text{epi}(z) - (x^*, \eta^*))^-$. Furthermore, since both problems have the same objective value, we have that $\omega^\top \pi + \omega_0 \pi_0 < 0$.

Since $(\gamma, \gamma_0) \in P(x^*, \eta^*)$, we have that γ is a valid certificate for the vector (π, π_0) in (3.20). By Theorem 3.14, this implies that the inequality $\gamma^\top Hx - \gamma_0 \eta \leq \gamma^\top b$ does indeed support epi(z). □

As a consequence of Theorem 3.19, it is sufficient in this context to focus on the selection of cut *normals*, since we automatically obtain the corresponding optimal right-hand side at no additional computational cost.

Finally, to conclude our dictionary of cut-generating optimization problems, we derive an alternative representation of the optimization problem (3.27) which will turn out to be much more useful in practice. For instance, the structure of the resulting problem will be very similar to the original subproblem, which makes it easy to use existing solution algorithms for the subproblem in a cut-generating program.

Cornuéjols and Lemaréchal [CL06, Theorem 4.2] prove that linear optimization problems over the reverse polar set can be evaluated in terms of the support function

of the original set (in our case $\text{epi}(z) - (x^*, \eta^*)$). This can also be applied to the alternative polyhedron, as mentioned (without proof) by Fischetti, Salvagnin, and Zanette [FSZ10]. The following theorem generalizes Cornuéjols and Lemaréchal [CL06, Theorem 4.2] and makes a similar statement, which is applicable to a wider range of settings.

Theorem 3.20
Let $K \subset \mathbb{R}^n$ be a cone and $c_1, c_2 \in \mathbb{R}^n$. Consider the optimization problems

$$\max\left\{c_1^\top x \;\middle|\; x \in K, c_2^\top x = -1\right\} \tag{3.30}$$

and

$$\max\left\{c_2^\top x \;\middle|\; x \in K, c_1^\top x \geq 1\right\}. \tag{3.31}$$

Then the following hold:

a) If x^* is an optimal solution for (3.30) with objective value $\xi > 0$, then $\frac{1}{\xi} \cdot x^*$ is an optimal solution for (3.31) with objective value $-\frac{1}{\xi} < 0$.

b) Conversely, if x^* is an optimal solution for (3.31) with objective value $\xi < 0$, then $-\frac{1}{\xi} \cdot x^*$ is an optimal solution for (3.30) with objective value $-\frac{1}{\xi} > 0$.

Proof. For a), let x^* be an optimal solution of (3.30) with objective value $\xi > 0$. Then $\frac{1}{\xi} \cdot x^* \in K$ and $c_1^\top \left(\frac{1}{\xi} \cdot x^*\right) = 1$. Hence the point $\frac{1}{\xi} \cdot x^*$ is feasible for (3.31). Furthermore, its objective value is

$$c_2^\top \left(\frac{1}{\xi} \cdot x^*\right) = \frac{1}{\xi} \cdot c_2^\top x^* = -\frac{1}{\xi}.$$

To see that $\frac{1}{\xi} \cdot x^*$ is indeed optimal, let x' be feasible for (3.31). We first claim that $c_2^\top x' < 0$: Suppose for contradiction that $c_2^\top x' \geq 0$. Choose $\varepsilon > 0$ such that $\varepsilon \cdot c_2^\top x' < 1$. Then, $c_2^\top (x^* + \varepsilon x') = c_2^\top x^* + \varepsilon \cdot c_2^\top x' =: \lambda \in [-1, 0)$. But this means that $c_2^\top \frac{-1}{\lambda}(x^* + \varepsilon x') = -1$ and since $x' \in K$ we furthermore have that $\frac{-1}{\lambda}(x^* + \varepsilon x') \in K$. Together, this implies that $\frac{-1}{\lambda}(x^* + \varepsilon x')$ is feasible for (3.30). But

$$c_1^\top \frac{-1}{\lambda}(x^* + \varepsilon x') \geq \frac{-1}{\lambda}(c_1^\top x^* + \varepsilon) \geq (c_1^\top x^* + \varepsilon) > c_1^\top x^*,$$

a contradiction with the optimality of x^*. This proves that, indeed, $c_2^\top x' < 0$.

Now, suppose that $\mu := c_2^\top x' > -\frac{1}{\xi}$. As we have seen above, $\mu < 0$ and hence $-\frac{1}{\mu} > \xi > 0$. Since $-\frac{1}{\mu} \cdot x' \in K$ and $c_2^\top \left(-\frac{1}{\mu} \cdot x'\right) = -\frac{1}{c_2^\top x'} \cdot c_2^\top x' = -1$, the point $-\frac{1}{\mu} \cdot x'$ is feasible for (3.30) with objective value

$$c_1^\top \left(-\frac{1}{\mu} \cdot x'\right) = -\frac{1}{\mu} c_1^\top x' \geq -\frac{1}{\mu} > \xi = c_1^\top x^*,$$

again contradicting the optimality of x^*, which proves a).

For b), let x^* be an optimal solution for (3.31) with objective value $\xi < 0$. First, note that $c_1^\top x^* = 1$: Otherwise, let $\varepsilon \in (0,1)$ such that $(1-\varepsilon) \cdot c_1^\top x^* \geq 1$. Then, $(1-\varepsilon)x^*$ is feasible for (3.31) and $c_2^\top (1-\varepsilon)x^* = (1-\varepsilon)\xi > \xi$, a contradiction with optimality of x^*.

Now, $-\frac{1}{\xi} \cdot x^* \in K$ and $c_2^\top(-\frac{1}{\xi} \cdot x^*) = -\frac{1}{\xi} \cdot c_2^\top x^* = -1$. Hence, $-\frac{1}{\xi} \cdot x^*$ is feasible for (3.30) and its objective value is

$$c_1^\top(-\frac{1}{\xi} \cdot x^*) = -\frac{1}{\xi} \cdot c_1^\top x^* = -\frac{1}{\xi}.$$

To see that $-\frac{1}{\xi} \cdot x^*$ is indeed optimal, suppose that there exists x' feasible for (3.30) with $\mu := c_1^\top x' > c_1^\top(-\frac{1}{\xi} \cdot x^*) = -\frac{1}{\xi} > 0$. Since $\frac{1}{\mu} \cdot x' \in K$ and $c_1^\top(\frac{1}{\mu} \cdot x') = \frac{1}{\mu} c_1^\top x' = 1$, the point $\frac{1}{\mu} \cdot x'$ is feasible for (3.31) with objective value $c_2^\top(\frac{1}{\mu} \cdot x') = \frac{1}{\mu} c_2^\top x' = -\frac{1}{\mu} > \xi$, a contradiction with the optimality of x^*. □

Choosing $c_1 := (Hx^* - b, -\eta^*)^\top$, $c_2 := (\tilde{\omega}, \tilde{\omega}_0)^\top$ and $K := \{(\gamma, \gamma_0) \geq 0 | \gamma^\top A + \gamma_0 d^\top = 0\}$, the following corollary immediately follows from part a) of the above theorem:

Corollary 3.21
Let $(\tilde{\omega}, \tilde{\omega}_0) \in \mathbb{R}^m \times \mathbb{R}$ and let (γ^*, γ_0^*) denote an optimal solution with value $\xi > 0$ for the problem

$$\max_{\gamma, \gamma_0 \geq 0} \ \gamma^\top (Hx^* - b) - \gamma_0 \eta^* \tag{3.32}$$

$$\gamma^\top A + \gamma_0 d^\top = 0 \tag{3.33}$$

$$\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 = -1. \tag{3.34}$$

Then $\frac{1}{\xi} \cdot (\gamma^*, \gamma_0^*)$ is an optimal solution with value $-\frac{1}{\xi}$ for the problem

$$\max\Big\{\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 \ \Big| \ (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\Big\}. \tag{3.35}$$

The structural similarity of (3.32) to (3.34) to the original problem becomes more apparent when we consider the dual problem:

Corollary 3.22
Let $(\tilde{\omega}, \tilde{\omega}_0) \in \mathbb{R}^m \times \mathbb{R}$, let (λ, x, y) be an optimal solution for the problem

$$\min \ \lambda \tag{3.36}$$

$$Ay \leq b - Hx^* - \lambda \cdot \tilde{\omega} \tag{3.37}$$

$$d^\top y \leq \eta^* - \tilde{\omega}_0 \lambda \tag{3.38}$$

with $\lambda > 0$ and denote a corresponding dual solution by (γ, γ_0). Then $\frac{1}{\lambda}(\gamma, \gamma_0)$ is an optimal solution with objective value $-\frac{1}{\lambda}$ for

$$\max\Big\{\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 \ \Big| \ (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\Big\}.$$

Note that, together with our observations in the context of Definition 3.5, this means in particular that

a) whenever (3.36) to (3.38) has objective value 0, then the alternative polyhedron is empty and $(x^*, \eta^*) \in \text{epi}(z)$, and

b) whenever (3.36) to (3.38) is feasible with (finite) objective value greater than 0, then (3.26) and (3.27) have objective values strictly less than 0, which means that the requirements e. g. for Theorem 3.14 or Remark 3.18 are satisfied.

Finally, Corollary 3.22 exposes another interesting perspective on the restriction $(\tilde{\omega}, \tilde{\omega}_0)^\top := (H\omega, -\omega_0)^\top$ on the objective function (and hence the relation between optimization problems over the alternative polyhedron and the reverse polar set):

Remark 3.23
If $(\tilde{\omega}, \tilde{\omega}_0)^\top := (H\omega, -\omega_0)^\top$, then the optimization problem (3.36) to (3.38) becomes

$$\min \lambda \tag{3.39}$$
$$Ay \leq b - H(x^* + \lambda \cdot \omega) \tag{3.40}$$
$$d^\top y \leq \eta^* + \omega_0 \lambda \tag{3.41}$$

Comparing the two optimization problems, both can be seen as a relaxation of the feasibility version of the original subproblem (3.8): They allow a solution to violate certain constraints, possibly (depending on the signs of entries in H and ω) at the cost of strengthening others. In any case, a feasible solution for (3.8) is feasible for both problems with objective value 0.

The only difference between the two is how exactly this relaxation is handled: In (3.36) to (3.38), it works on the level of individual inequalities by relaxing their right-hand sides, whereas in (3.39) to (3.41) it works on the level of the master solution (x^*, η^*), allowing us to choose a possibly more advantageous value for the variable x itself.

Another convenient consequence of the bijection between optimal solutions for (3.32) to (3.34) and (3.35) as implied by Theorem 3.20 (Corollary 3.21 only states the more useful direction) is that extremality of optimal solutions is also maintained in both directions, which will later turn out to be very useful:

Corollary 3.24
Let $(\tilde{\omega}, \tilde{\omega}_0) \in \mathbb{R}^m \times \mathbb{R}$.

a) If (γ^*, γ_0) *is an* extremal *optimal solution with value* $\xi > 0$ *for the problem* (3.32) *to* (3.34)*, then* $\frac{1}{\xi} \cdot (\gamma^*, \gamma_0)$ *is an* extremal *optimal solution with value* $-\frac{1}{\xi}$ *for the problem* (3.35)*.*

Input: An instance of the generic optimization problem (3.1), a lower bound $\bar{\eta}$ for $z(x)$
Output: optimal solution (x, y) to (3.1)

1: set $i := 1$ and initialize the set $\text{epi}_S(z)^{(1)} := \{(x, \eta) \in S \times \mathbb{R} \mid \eta \geq \bar{\eta}\}$
2: solve the problem $\min\{c^\top x + \eta \mid (x, \eta) \in \text{epi}_S(z)^{(1)}\}$ to obtain $(x^{(1)}, \eta^{(1)})$
3: **while** $(x^{(i)}, \eta^{(i)}) \notin \text{epi}(z)$ **do**
4: choose a weight vector $(\tilde{\omega}, \tilde{\omega}_0)$
5: solve (3.36) to (3.38) with $(x^*, \eta^*) := (x^{(i)}, \eta^{(i)})$ to obtain a dual solution (γ, γ_0)
6: set $\text{epi}_S(z)^{(i+1)} := \text{epi}_S(z)^{(i)} \cap \{(x, \eta) \mid \gamma^\top Hx - \gamma_0 \eta \leq \gamma^\top b\}$
7: set $i := i + 1$
8: solve the problem $\min\{c^\top x + \eta \mid (x, \eta) \in \text{epi}_S(z)^{(i)}\}$ to obtain $(x^{(i)}, \eta^{(i)})$
9: **end while**
10: set $x^* := x^{(i)}$
11: solve problem (3.4) to compute y^*
12: **return** (x^*, y^*)

Algorithm 2: The improved Benders decomposition algorithm.

b) Conversely, if (γ^*, γ_0) *is an* extremal *optimal solution with value* $\xi < 0$ *for the problem* (3.35)*, then* $-\frac{1}{\xi} \cdot (\gamma^*, \gamma_0)$ *is an* extremal *optimal solution with value* $-\frac{1}{\xi}$ *for the problem* (3.32) *to* (3.34)*.*

Based on these observations, we can define the modified Benders decomposition algorithm using weight vectors $(\tilde{\omega}, \tilde{\omega}_0)$ as in Algorithm 2. If we choose $(\tilde{\omega}, \tilde{\omega}_0)^\top := (H\omega, -\omega_0)^\top$ for some $(\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$, then by Theorem 3.17 we can interpret step 5 as an optimization problem over the reverse polar set with objective function (ω, ω_0).

3.2 Cut Selection

In the following we will discuss different criteria for the selection of Benders cuts to see what conclusions can be drawn from our above analysis. Cut selection is one of four major areas of algorithmic improvements for Benders decomposition that recent work has focussed on (see the recent and extensive literature review in [Rah+17]).

As we have seen in the previous section, Benders' decomposition can be viewed as an instance of a classical cutting plane algorithm (Theorem 3.6). The Benders subproblem takes the role of the separation problem and the alternative polyhedron that is commonly used to select a Benders cut is a higher-dimensional representation (an *extended formulation*) of the reverse polar set, which characterizes all possible cut normals (Theorem 3.13). Focussing on cut normals is sufficient, since the corresponding optimal right-hand side is easy to obtain as shown in Theorem 3.19.

Finally, Corollary 3.22 and Remark 3.23 show that selecting a cut normal by a linear objective over the alternative polyhedron or the reverse polar set can be interpreted as two different relaxations (3.36) to (3.38) and (3.39) to (3.41) of the original Benders feasibility subproblem (3.8). The former relaxation is more general and coincides with the latter for a particular selection of the objective function.

A number of selection criteria for Benders cuts have previously been explicitly proposed in the literature, some of which also arise naturally from our discussion and analysis of the Benders decomposition algorithm above. We will first present these criteria in the way they typically appear in the literature and then link them to the reverse polar set and/or the alternative polyhedron.

3.2.1 Minimal Infeasible Subsystems

The approach to cut selection proposed by Fischetti, Salvagnin, and Zanette [FSZ10] is based on the premise that "one is interested in detecting a 'minimal source of infeasibility'" whenever the feasibility subproblem (3.8) is empty. They hence suggest to generate Benders cuts based on Farkas certificates that correspond to *minimal infeasible subsystems (MIS)* of (3.8). We define this criterion as follows:

Definition 3.25
Let z be defined as in (3.2) and let $(\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R}$. We say that (π, π_0) *satisfies the* MIS *criterion* if there exists $(\gamma, \gamma_0) \geq 0$ such that $\pi = H^\top \gamma, \pi_0 = -\gamma_0$ and the inequalities which correspond the non-zero components of (γ, γ_0) form a minimal infeasible subsystem of (3.8).

Note that we have defined the MIS criterion as a property of a normal vector, rather than a property of a cut. The reason for this is that, as we have argued above, the cut normal is the only relevant choice to make, given that the optimal right-hand side for each cut normal is obvious and can be computed using Theorem 3.19. Accordingly, we will call any cut with a normal vector that satisfies the MIS criterion a MIS-cut.

Gleeson and Ryan [GR90] show that the set of (γ, γ_0) that appear in the above definition is exactly (up to homogeneity) the set of vertices of the alternative polyhedron:

Theorem 3.26 (Gleeson and Ryan [GR90])
Let $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. A vector v is a vertex of the relaxed alternative polyhedron (3.21) *if and only if the set of constraints such that the corresponding entries of v are non-zero forms a minimal infeasible subsystem of* (3.8).

This immediately provides a characterization of cut normals which satisfy MIS in terms of the alternative polyhedron, which is also used in [FSZ10]:

Corollary 3.27
Let z be defined as in (3.2) *and let* $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. *The vector* (π, π_0) *satisfies the* MIS *criterion if and only if there is an extremal point* (γ, γ_0) *of* $P(x^*, \eta^*)$ *with* $(\pi, \pi_0) = (H^\top \gamma, -\gamma_0)$.

Theorem 3.13 allows us to transfer the *if*-part of this characterization to the reverse polar set. The *only-if*-part is generally not true for the reverse polar set, i. e., there might be minimal infeasible subsystems that do not correspond to vertices of the reverse polar set (see, e. g., Example 3.9).

Corollary 3.28
Let z be defined as in (3.2) *and* $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. *If* (π, π_0) *is a vertex of* $(\mathrm{epi}(z) - (x^*, \eta^*))^-$, *then it satisfies the criterion* MIS.

Fischetti, Salvagnin, and Zanette [FSZ10] empirically study the performance of MIS-cuts on a set of multi-commodity network design instances. Their results suggest that MIS-based cut selection outperforms the standard implementation of Benders decomposition by a factor of at least 2-3. Furthermore, this advantage increases substantially when focussing on harder instances (e. g. those which could not be solved by the standard implementation within 10 hours).

3.2.2 Facet-defining Cuts

In cutting plane algorithms for polyhedra, facet-defining cuts are generally considered a very useful family of cuts. They form the smallest set of inequalities which completely describe a (full-dimensional) target polyhedron. A cutting-plane algorithm that can separate distinct facet inequalities in this context is hence guaranteed to terminate after a finite number of iterations. Also in practical applications, facet cuts have turned out to be extremely useful, e. g. in the context of branch-and-cut algorithms for integer programs such as the Traveling Salesman Problem. This is why the description of facet-defining inequalities has been a large and very active area of research for decades (see, e. g., [Bal75; NW88; Coo+98; KV08] and, as mentioned before, [CW18]).

Remember that a halfspace $H^{\leq}_{((\pi,\pi_0),\alpha)}$ is facet-defining for a set C if $C \subset H^{\leq}_{((\pi,\pi_0),\alpha)}$ and $H^{=}_{((\pi,\pi_0),\alpha)} \cap C$ contains $\dim(C)$ many affinely independent points. Analogously to the MIS criterion above, we define the Facet criterion for a normal vector in the context of Benders decomposition as follows:

Definition 3.29
Let z be defined as in (3.2) and $(\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R} \setminus \{0\}$. We say that (π, π_0) *satisfies the* Facet *criterion* if there exists $\alpha \in \mathbb{R}$ such that $H^{\leq}_{((\pi,\pi_0),\alpha)}$ is either facet-defining for $\mathrm{epi}(z)$ or if the corresponding hyperplane $H^{=}_{((\pi,\pi_0),\alpha)}$ contains $\mathrm{epi}(z)$.

Note that, in deviation from the definition of a facet-defining cut, we require that the halfspace supports *at least* $\dim(C)$ affinely independent points. In other words, in the case where $\text{epi}(z)$ is not full-dimensional, we also allow that the halfspace $H^{\leq}_{((\pi,\pi_0),\alpha)}$ supports all of $\text{epi}(z)$. In this situation, the comparison of different cut normals is inherently difficult: If one cuts both support all of $\text{epi}(z)$ and another one supports a facet, which one should be preferred? In this sense, the criterion FACET captures arguably the strongest statement about a cut in relation to $\text{epi}(z)$ that we can make in a given situation: In no case would we want to select a cut that supports *neither* a facet *nor* the entire set $\text{epi}(z)$.

The following result was originally obtained by Balas [Bal98, Theorem 4.5] in his analysis of disjunctive cuts. It reappears in Cornuéjols and Lemaréchal [CL06, Theorem 6.2] using more familiar notation, but that version contains a minor error in the case where the set P is sub-dimensional. We therefore re-prove a corrected version of the important parts of [CL06, Theorem 6.2] below along the lines of their original proof.

Theorem 3.30
Let $P \subset \mathbb{R}^n$ be a polyhedron, $x^* \notin P$ and let

$$r := \begin{cases} \dim(P) - 1, & x^* \in \text{aff}(P) \\ \dim(P), & x^* \notin \text{aff}(P). \end{cases}$$

Then, there exists an x^*-separating halfspace with normal vector $d \neq 0$ supporting a r-dimensional face of P if and only if there exists a vertex d^* of $(P - x^*)^- \cap \text{lin}(P - x^*)$ and $\lambda > 0$ such that $\lambda d \in d^* + \text{lin}(P - x^*)^\perp$.

Proof. We begin by observing that $r + 1 = \dim(\text{lin}(P - x^*))$, regardless of whether $x \in \text{aff}(P)$ or not.

If d supports an r-dimensional face of P, then there exist $r + 1$ affinely independent points in P with $d^\top x = h_P(d) < d^\top x^*$ (since the halfspace with normal vector d is x^*-separating). Denote these points by $x^1, \ldots, x^{(r+1)}$ and let $d' := \frac{d}{d^\top x - h_P(d)}$. Then, $d' \in (P - x^*)^-$ and the inequalities $(d')^\top (x^i - x^*) \leq -1$ constitute a system of $r + 1$ linearly independent inequalities valid for $(P - x^*)^-$, which are all satisfied with equality by d'.

Let d^* denote the orthogonal projection of d' onto $\text{lin}(P - x^*)$. Since $(d^*)^\top x = (d')^\top x$ for all $x \in \text{lin}(P - x^*)$, it holds in particular that $d^* \in (P - x^*)^-$ and d^* satisfies with equality the same set of linearly independent inequalities as d' above. Furthermore, since $(x^i - x^*) \in \text{lin}(P - x^*)$ for all $i \in [r + 1]$, the point d^* is indeed a vertex of $(P - x^*)^- \cap \text{lin}(P - x^*)$. With $\lambda := \frac{1}{d^\top x - h_P(d)}$, this proves the *only-if* part of the statement.

For the *if* part, let d^* be a vertex of the polyhedron $(P - x^*)^- \cap \text{lin}(P - x^*)$ and $\lambda d \in d^* + \text{lin}(P - x^*)^\perp$. Then, by the definition of the reverse polar set there exist

$r+1$ linearly independent points in $(P - x^*)$ such that d^* satisfies the corresponding inequalities with equality. Denote these points by $x^1 - x^*, x^2 - x^*, \ldots, x^{(r+1)} - x^*$.

Then, $(\lambda d)^\top x = (d^*)^\top x$ for all $x \in \operatorname{lin}(P - x^*)$ and therefore $\lambda d \in (P - x)^-$ and λd exposes a face of P containing the affinely independent points $x^1, \ldots, x^{(r+1)}$. The face is thus r-dimensional and, since $\lambda > 0$, it is supported by an x^*-separating halfspace with normal vector d. □

Most notably, for the case where P is full-dimensional the above theorem implies the following:

Corollary 3.31
Let $P \subset \mathbb{R}^n$ be a polyhedron with $\dim(P) = n$ and $x^* \notin P$. Then there exists an x^*-separating halfspace with normal vector d supporting a facet of P if and only if there exists a vertex d^* of $(P - x^*)^-$ and $\lambda \geq 0$ such that $\lambda d = d^*$.

In this case, every cut generated from a vertex of the reverse polar set defines a facet of $\operatorname{epi}(z)$. If an explicit $\mathcal{H}$-representation of the reverse polar set is available, we can thus easily obtain a facet-defining cut, e. g. by linear programming.

Note that since $P^{\leq}(x^*, \eta^*)$ is line-free, Theorem 3.13 implies that for every vertex of the reverse polar set there exists a vertex of the relaxed alternative polyhedron (and hence of the original alternative polyhedron) that leads to the same cut normal. In other words, if the normal of an x^*-separating halfspace satisfies the Facet criterion, then it also satisfies the MIS criterion.

On the other hand, Theorem 3.13 is not sufficient to guarantee that selecting a vertex of the alternative polyhedron yields a facet-defining cut: As Example 3.9 shows, even if (γ, γ_0) is a vertex of $P^{\leq}(x^*, \eta^*)$, its image under the mapping of Theorem 3.13 need not be a vertex of the reverse polar set. This yields a useful hierarchy of subsets of the alternative polyhedron according to the properties of the cut normals which they induce: Any point in the alternative polyhedron can be used to obtain a valid cut. Any vertex of the alternative polyhedron guarantees that the resulting cut normal satisfies the criterion MIS. Furthermore, a subset of these vertices consists of exactly those points that lead to cut normals satisfying the criterion Facet. The latter are generally a good choice in the context of any cutting plane algorithm and the approach of selecting MIS-cuts can in this context be viewed as a heuristic method to find Facet-cuts.

Although cuts satisfying MIS do not in general satisfy Facet, we can obtain some information on when this is the case in the situation of Theorem 3.17, i. e., if the objective function $(\tilde{\omega}, \tilde{\omega}_0)$ used to select the cut via problem (3.27) satisfies $(\tilde{\omega}, \tilde{\omega}_0) = (H\omega, -\omega_0)$ for some valid objective (ω, ω_0) for problem (3.26).

In this case it turns out that we actually *almost always* obtain a Facet-cut. More precisely, we can prove the following characterization of the relationship between

extremal points of the alternative polyhedron and cut normals satisfying the criterion FACET:

Theorem 3.32
Let z be defined as in (3.2) *and $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$ with $(x^*, \eta^*) \notin \operatorname{epi}(z)$. Let (ω, ω_0) be such that* (3.26) *is bounded and $(\tilde{\omega}, \tilde{\omega}_0) := (H\omega, -\omega_0)$. Then, there exists an optimal extremal point $(\gamma^*, \gamma_0^*) \in P^{\leq}(x^*, \eta^*)$ with respect to the objective function $(\tilde{\omega}, \tilde{\omega}_0)$ such that the resulting cut normal $(H^\top \gamma^*, -\gamma_0^*)$ satisfies the criterion* FACET.

Proof. First, observe that for any $C \subset \mathbb{R}^n$, it holds that $\operatorname{lin}(C)^\perp$ is the lineality space of C^- (since $c^\top x^* = 0$ for all $c \in C$ if and only if $x + \lambda x^* \in C^-$ for all $x \in C^-$ and $\lambda \in \mathbb{R}$). In particular, let $L := \operatorname{lin}(\operatorname{epi}(z) - (x^*, \eta^*))$. Then $L^\perp$ is the lineality space of $(\operatorname{epi}(z) - (x^*, \eta^*))^-$.

Now, since the reverse polar set $(\operatorname{epi}(z) - (x^*, \eta^*))^-$ is bounded in the direction of (ω, ω_0), it holds that $(\omega, \omega_0)^\top (\pi, \pi_0) = 0$ for all $(\pi, \pi_0) \in L^\perp$. The intersection $(\operatorname{epi}(z) - (x^*, \eta^*))^- \cap L$ thus contains an optimal solution with respect to the objective (ω, ω_0) (take the orthogonal projection of any optimal solution onto L, it has the same objective value and lies in $(\operatorname{epi}(z) - (x^*, \eta^*))^-$ since $L^\perp$ is the lineality space). While the reverse polar need not be line-free, note that $(\operatorname{epi}(z) - (x^*, \eta^*))^- \cap L$ is indeed line-free and we can therefore choose (π, π_0) to be extremal in $(\operatorname{epi}(z) - (x^*, \eta^*))^- \cap L$. By Theorem 3.17, there exists γ' with $H^\top \gamma' = \pi$ such that $(\gamma', -\pi_0)$ is an optimal solution to the problem

$$\max\Big\{\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 \;\Big|\; (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\Big\}. \tag{3.42}$$

Denote by P^* the face of optimal solutions of (3.42) and observe that

$$\begin{aligned}(\gamma', -\pi_0) &\in P^* \cap \{(\gamma, \gamma_0) \mid (H^\top \gamma, -\gamma_0) - (\pi, \pi_0) = 0\} \\ &\subset P^* \cap \{(\gamma, \gamma_0) \mid (H^\top \gamma, -\gamma_0) - (\pi, \pi_0) \in L^\perp\}.\end{aligned}$$

Let (γ^*, γ_0^*) be an extremal point of $P^* \cap \{(\gamma, \gamma_0) \mid (H^\top \gamma, -\gamma_0) - (\pi, \pi_0) \in L^\perp\}$ (which exists, since $P^{\leq}(x^*, \eta^*)$ is line-free). Then (γ^*, γ_0^*) is obviously optimal for (3.42) and furthermore $(H^\top \gamma^*, -\gamma_0^*) = (\pi, \pi_0) + v$ with $v \in L^\perp$, which means by Theorem 3.30 that it satisfies the criterion FACET. It remains to show that (γ^*, γ_0^*) is a vertex of P^*, which would imply that it is also a vertex of $P^{\leq}(x^*, \eta^*)$.

To see this, let $(\gamma^1, \gamma_0^1), (\gamma^2, \gamma_0^2) \in P^*$ such that $(\gamma^*, \gamma_0^*) \in \operatorname{relint}([(\gamma^1, \gamma_0^1), (\gamma^2, \gamma_0^2)])$. However, $(\pi, \pi_0) + v = (H^\top \gamma^*, -\gamma_0^*) \in \operatorname{relint}([(H^\top \gamma^1, -\gamma_0^1), (H^\top \gamma^2, -\gamma_0^2)])$ and by Theorem 3.13, $[(H^\top \gamma^1, -\gamma_0^1), (H^\top \gamma^2, -\gamma_0^2)] \subset (\operatorname{epi}(z) - (x^*, \eta^*))^-$. As (π, π_0) is extremal in $(\operatorname{epi}(z) - (x^*, \eta^*))^- \cap L$, this implies that $(H^\top \gamma^1, -\gamma_0^1), (H^\top \gamma^2, -\gamma_0^2) \in (\pi, \pi_0) + L^\perp$ which means that $(\gamma^1, \gamma_0^1), (\gamma^2, \gamma_0^2) \in P^* \cap \{(\gamma, \gamma_0) \mid (H^\top \gamma, -\gamma_0) - (\pi, \pi_0) \in L^\perp\}$. As (γ^*, γ_0^*) is extremal in $P^* \cap \{(\gamma, \gamma_0) \mid (H^\top \gamma, -\gamma_0) - (\pi, \pi_0) \in L^\perp\}$, this implies that $(\gamma^1, \gamma_0^1) = (\gamma^2, \gamma_0^2) = (\gamma^*, \gamma_0^*)$, which proves extremality of (γ^*, γ_0^*) in P^*. □

In particular, the above theorem implies the following: If $(\gamma', \gamma'_0) \in P^{\leq}(x^*, \eta^*)$ is an optimal extremal point with respect to the objective function $(H\omega, -\omega_0)$ such that the resulting cut normal $(H^\top \gamma', -\gamma'_0)$ does not satisfy the criterion Facet, then the optimal solution for maximizing $(H\omega, -\omega_0)$ over $P^{\leq}(x^*, \eta^*)$ is not unique. Furthermore, by Theorem 3.17, this implies that the same is true for maximizing the objective (ω, ω_0) over $(\text{epi}(z) - (x^*, \eta^*))^-$.

The implications of the above theorem can be summarized as follows: While any Facet-cut is also an MIS-cut, the reverse is not always true. However, if we choose a vector (ω, ω_0) and optimize the objective $(\tilde{\omega}, \tilde{\omega}_0) := (H\omega, -\omega_0)$ over the alternative polyhedron, then there exists only a sub-dimensional set of choices for the vector (ω, ω_0) for which the resulting cut might not satisfy Facet (those, for which the optimum over the reverse polar set is non-unique).

This suggests that these cases should be "rare" in practice, especially if we choose (or perturb) (ω, ω_0) randomly from some full-dimensional set. This argument why a cut obtained for a generic vector (ω, ω_0) can be expected to be facet-defining is identical to the concept of "almost surely" finding facet-defining cuts proposed by Conforti and Wolsey [CW18] in a more general context.

Looking back at Remark 3.23, this similarity should not come as a surprise: With $(\omega, \omega_0) = (\bar{x} - x^*, \bar{\eta} - \eta^*)$ for a point $(\bar{x}, \bar{\eta}) \in \text{relint epi}(z)$, the resulting cut-generating LP is almost identical. In fact, the point $(\bar{x}, \bar{\eta})$ in this case takes the role of the point that the origin is relocated into in the approach from [CW18]. Observe, however, that while Conforti and Wolsey [CW18] require that point to lie in the relative interior of $\text{epi}(z)$, we can actually obtain a cut satisfying the Facet criterion from any (ω, ω_0) for which the optimal objective over the reverse polar is strictly negative. By Theorem 3.15, one sufficient (but not necessary) criterion for this is to choose (ω, ω_0) as above, even for an arbitrary point $(\bar{x}, \bar{\eta}) \in \text{epi}(z)$.

By this observation, our work in particular represent a (slightly more general) alternative proof for the main result from [CW18]: In fact, Cornuéjols and Lemaréchal [CL06] already provided a (theoretical) method to generate facet-defining cuts. Using the alternative polyhedron as a computationally efficient representation of the relevant geometric object (the reverse polar set), we show that this method can actually be used in practice. We thus provide an (arguably) simpler proof, avoiding some of the technicalities of [CW18]. Our approach furthermore connects the result more directly to previous work on cut selection in Benders decomposition, such as [FSZ10].

Finally, if we want to be sure that a cut computed from the alternative polyhedron satisfies Facet, then the following observation allows us to iteratively restrict the optimal set to a singleton, until we obtain a facet-defining cut for sure. Again, this is only true if $(\tilde{\omega}, \tilde{\omega}_0) := (H\omega, -\omega_0)$.

To see this, observe that if we choose the objective as described above, then the optimal face of the alternative polyhedron corresponds to a face of the reverse polar, as the following lemma shows.

Lemma 3.33
Let $P \subset \mathbb{R}^m$ and $T \in \mathbb{R}^{n\times m}$. Let $c \in \mathbb{R}^n$ and $\tilde{c} := T^\top c$ such that $\max_{x\in P} \tilde{c}^\top x =: \alpha < \infty$. Let $P' := \operatorname{argmax}\{\tilde{c}^\top x \mid x \in P\}$. Then $T \cdot P'$ is a face of $T \cdot P$.

Proof. We first observe that for all $y \in TP$ and $y' \in TP'$, it holds that $c^\top y \leq c^\top y'$: By the definition of TP and TP', there exist $x \in P$ and $x' \in P'$ such that $y = Tx$ and $y' = Tx'$. But then $c^\top y = c^\top Tx = (T^\top c)^\top x = \tilde{c}^\top x \leq \tilde{c}^\top x' = (T^\top c)^\top x' = c^\top y'$.

Clearly, $P' \subset P$ and hence $TP' \subset TP$. Let therefore F^* be the lowest-dimensional face of TP that contains TP' (in the worst case, $F^* = TP$). We show that $F^* = TP'$.

First, suppose that c is not orthogonal on F^*, then $c^\top y = \alpha$ for all $y \in T^\top P'$ but not for all $y \in F^*$. As, in addition, $c^\top y \leq \alpha$ for all $y \in T^\top P$, the set $F^* \cap \{c^\top y = \alpha\}$ constitutes a lower-dimensional face of $T^\top P$ that contains $T^\top P'$, a contradiction.

Hence, c is orthogonal on F^*. This implies that for every $y \in F^*$ and $x \in P'$, there is $x^* \in P$ with $y = Tx^*$ and hence $\tilde{c}^\top x^* = c^\top Tx^* = c^\top y = c^\top Tx = \tilde{c}^\top x$ where the second-to-last equality follows from the fact that $\tilde{c}$ is orthogonal on F^* and $TP' \subset F^*$. But this means that $x^* \in P'$ and we have that $T^\top P' = F^*$. □

In particular, the above lemma implies that every vertex of $T \cdot P'$ is a vertex of $T \cdot P$. This allows us in principle to determine whether a cut obtained from the alternative polyhedron is facet-defining and, if not, to lift it to a facet-defining cut: We can iteratively reduce the dimension of the optimal set while maintaining that its image under H contains a vertex of the reverse polar set. Once we reach dimension 0, the remaining point must be a vertex of the reverse polar set, as well.

More formally, let $(\omega^1, \omega_0^1), (\omega^2, \omega_0^2), \ldots, (\omega^{(n+1)}, \omega_0^{(n+1)}) \in \mathbb{R}^{n+1}$ be linearly independent and hence form a basis of $\mathbb{R}^{n+1}$. With

$$T := \begin{pmatrix} H^\top & 0 \\ 0 & -1 \end{pmatrix},$$

let $P_1 := \operatorname{argmax}\{(T(\omega^1, \omega_0^1))^\top(\gamma, \gamma_0) \mid (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\}$. Furthermore, let $P_i := \operatorname{argmax}\{(T(\omega^1, \omega_0^1))^\top(\gamma, \gamma_0) \mid (\gamma, \gamma_0) \in P_{i-1}\}$. As $(\omega^1, \omega_0^1), \ldots, (\omega^{(n+1)}, \omega_0^{(n+1)})$ are linearly independent, we have that $\dim(P_n) = 0$. By iteratively applying Lemma 3.33, the polyhedron TP_n contains a vertex of $TP^{\leq}(x^*, \eta^*) = (\operatorname{epi}(z) - (x^*, \eta^*))^-$, which we obtain as the unique (optimal) solution.

While theoretically feasible, the above approach will in most cases be very impractical, because of the large number of iterations necessary to compute a single cut. Furthermore, the problem that has to be solved becomes more and more cumbersome, the more objective values have to be fixed.

On the other hand, even a single iteration of the approach outlined above might be useful, as it dramatically decreases the likelihood of obtaining a cut that does not satisfy Facet (in the sense of reducing by one the dimension of the linear subspace in which the objective function has to lie). For this case where only one objective value

needs to be fixed, the resulting problem is not too different from problem (3.27) and we can derive a representation similar to (3.36) to (3.38) via theorem Theorem 3.20.

Analogously to Corollary 3.22, we obtain the following version of (3.36) to (3.38) where the objective value for the original version of (3.36) to (3.38) has been fixed to the optimal value D.

Theorem 3.34
Let z be defined as in (3.2) and $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$ with $(x^*, \eta^*) \notin \text{epi}(z)$ as well as $(\tilde{\omega}, \tilde{\omega}_0) \in \mathbb{R}^m \times \mathbb{R}$. Let $(\lambda^*, \sigma^*, y^*)$ be an optimal solution for the problem

$$\begin{aligned} \min\ & \lambda \\ & Ay \leq (b - Hx^*) \cdot (1 - D\sigma) - \sigma\tilde{\omega}' - \lambda \cdot \tilde{\omega} \\ & d^\top y \leq \eta^* \cdot (1 - D\sigma) - \sigma\tilde{\omega}_0' - \tilde{\omega}_0\lambda \end{aligned} \tag{3.43}$$

with $\lambda^* > 0$ and denote the corresponding dual solution by (γ^*, γ_0^*). Then $\frac{1}{\lambda^*}(\gamma^*, \gamma_0^*)$ is an optimal solution for the problem

$$\max\{\tilde{\omega}^\top\gamma + \tilde{\omega}_0\gamma_0 \mid (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*), \tilde{\omega}'^\top\gamma + \tilde{\omega}_0'\gamma_0 = D\}. \tag{3.44}$$

Proof. The dual LP for (3.43) is

$$\begin{aligned} \max\ & \gamma^\top(Hx^* - b) - \gamma_0\eta^* \\ & \gamma^\top A + \gamma_0 d^\top = 0 \\ & \tilde{\omega}'^\top\gamma + \tilde{\omega}_0'\gamma_0 = D \cdot (\gamma^\top(Hx^* - b) - \gamma_0\eta^*) \\ & \tilde{\omega}^\top\gamma + \tilde{\omega}_0\gamma_0 = -1 \\ & \gamma, \gamma_0 \leq 0 \end{aligned}$$

which, by introducing an additional variable ζ, can be rewritten as

$$\begin{aligned} \max\ & \gamma^\top(Hx^* - b) - \gamma_0\eta^* \\ & \gamma^\top A + \gamma_0 d^\top = 0 \\ & \tilde{\omega}'^\top\gamma + \tilde{\omega}_0'\gamma_0 - D\zeta = 0 \\ & \gamma^\top(Hx^* - b) - \gamma_0\eta^* - \zeta = 0 \\ & \tilde{\omega}^\top\gamma + \tilde{\omega}_0\gamma_0 = -1 \\ & \gamma, \gamma_0 \leq 0, \zeta \in \mathbb{R}. \end{aligned}$$

By strong duality, the optimal objective of the above problem is $\lambda^* > 0$ and we can hence use part a) of Theorem 3.20: If $(\gamma^*, \gamma_0^*, \zeta^*)$ is an optimal solution with objective value $\lambda^* > 0$ for the above LP, then $\frac{1}{\lambda^*}(\gamma^*, \gamma_0^*, \zeta^*)$ is an optimal solution for

$$\begin{aligned}
\max\ & \tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 \\
& \gamma^\top A + \gamma_0 d^\top = 0 \\
& \tilde{\omega}'^\top \gamma + \tilde{\omega}'_0 \gamma_0 - D\zeta = 0 \\
& \gamma^\top (Hx^* - b) - \gamma_0 \eta^* - \zeta = 0 \\
& \gamma^\top (Hx^* - b) - \gamma_0 \eta^* \geq 1.
\end{aligned} \tag{3.45}$$

For any feasible solution of (3.45), we have that $\zeta \geq 1$. But if $\zeta > 1$, then scaling $(\gamma, \gamma_0, \zeta)$ with a positive factor < 1 yields a better solution, thus for any optimal solution of (3.45) we have that $\zeta = 1$ and hence $\frac{1}{\lambda^*}\zeta^* = 1$, which proves the statement. □

Analogously to our discussion in the context of Remark 3.23, we can interpret (3.43) as a relaxation of the original feasibility subproblem (3.8): In addition to the relaxation by the variable λ (which we know from Corollary 3.22), we have added an additional relaxation that works in a slightly different way: By choosing $\sigma \neq 0$, we are allowed to scale all right hand sides uniformly by a common factor. In contrast to λ, this relaxation does not directly affect the objective, instead we incur an (additive) penalty tightening all inequalities in proportion to the values of $(\tilde{\omega}', \tilde{\omega}'_0)$.

3.2.3 Pareto Optimality

The first systematic work on the general selection of Benders cuts to our knowledge was undertaken by Magnanti and Wong [MW81]. The paper, which has proven very influential and is still referred to regularly, focusses on the property of Pareto optimality. It can intuitively be described as follows: A cut is Pareto-optimal if there is no other cut which is *clearly superior*, which *dominates* the first cut.

Magnanti and Wong focus on cuts which are valid for $\mathrm{epi}(z)$, i. e., they call a cut Pareto-optimal if it is not dominated by any other cut valid for $\mathrm{epi}(z)$. In this setting, a cut which does not support $\mathrm{epi}(z)$ is obviously dominated. Between supporting cuts, however there is no general mathematical criterion for domination. If the cut normal (π, π_0) satisfies $\pi_0 \neq 0$, however (this is also the case covered by [MW81]), things become somewhat easier:

Definition 3.35
For a problem of the form (3.1) with z as defined as in (3.2), let $S \subset \mathbb{R}^n$ and $\mathrm{epi}_S(z) := (S \times \mathbb{R}) \cap \mathrm{epi}(z)$. An inequality $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ with $\pi_0 < 0$ is *dominated* by another inequality $(\pi'^\top, \pi'_0)(x, \eta)^\top \leq \alpha'$ if $\pi'_0 < 0$ and

$$\frac{\pi'^\top x - \alpha'}{-\pi'_0} \geq \frac{\pi^\top x - \alpha}{-\pi_0} \quad \text{for all } x \in S$$

with strict inequality for at least one $x \in S$. If $\pi_0 < 0$ and $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ is not dominated by any inequality valid for $\mathrm{epi}(z)$, then we call it *Pareto-optimal*.

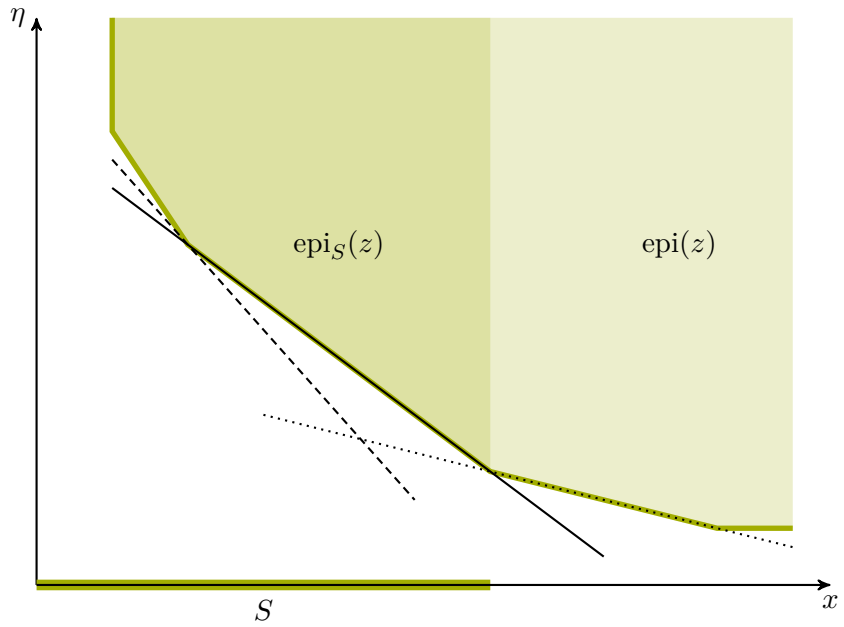


Figure 3.4: The dotted cut supports a facet of $\text{epi}(z)$ and it supports $\text{epi}_S(z)$, but it is still not Pareto-optimal. The solid cut supports a facet of $\text{epi}_S(z)$ and is hence Pareto-optimal. The dashed cut is Pareto-optimal even though it does not support a facet of $\text{epi}(z)$ (or $\text{epi}_S(z)$). Since the set S is convex, both the solid and the dashed cut are also strongly Pareto-optimal (Theorem 3.38).

By the above definition, a cut dominates another cut if the minimum value of η that it enforces is at least as good for all $x \in S$ and strictly better for at least one $x \in S$ (see Fig. 3.4).

In the context of Pareto-optimality, the reliance on the set $\text{epi}(z)$ outside of its intersection with $S \times \mathbb{R}$ seems a little unnatural: Remember that the set S contains all points $x \in \mathbb{R}^n$ that are feasible for an optimization problem of the form (3.1) if we ignore the linear constraints $Hx + Ay \leq b$. Since different functions z might therefore coincide when restricted to the relevant region S (but differ outside of S), we might intuitively prefer the following, slightly stronger definition of Pareto-optimality:

Definition 3.36
For a problem of the form (3.1) with z as defined as in (3.2), let $S \subset \mathbb{R}^n$ and $\text{epi}_S(z) := (S \times \mathbb{R}) \cap \text{epi}(z)$. If $\pi_0 < 0$ and $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ is not dominated by any inequality valid for $\text{epi}_S(z)$, then we call it *strongly Pareto-optimal*.

It is important to note at this point that the concept of (strong) Pareto-optimality refers to the set S explicitly. This is in contrast to the two previous criteria MIS and FACET, which were defined with respect to the set $\text{epi}(z)$. Depending on the structure

of the set S, this difference can be quite significant: For instance, a facet-defining cut for $\text{epi}(z)$ does not even need to support $\text{epi}_S(z)$ (see, e. g., the dotted cut in Fig. 3.5). On the other hand, it is easy to see that a cut cannot be strongly Pareto-optimal unless it supports $\text{epi}_S(z)$. The somewhat weaker condition of Pareto-optimality is less dependent on the set S, but nonetheless, even if a cut exposes a facet of $\text{epi}(z)$ *and* supports $\text{epi}_S(z)$ it might still be Pareto-dominated (cf. Fig. 3.4).

Clearly, since any inequality that is valid for $\text{epi}(z)$ is also valid for $\text{epi}_S(z)$, strong Pareto-optimality implies Pareto-optimality. The reverse is not true in general and it will furthermore turn out that while Pareto-optimal cuts are relatively easy to obtain, strongly Pareto-optimal cuts are not. This motivates the following theorem, which characterizes an important situation in which any Pareto-optimal cut is strongly Pareto-optimal.

We use the following separation lemma, which can be found in [Roc70, Theorem 20.2]:

Lemma 3.37 ([Roc70])
Let $C \subset \mathbb{R}^n$ be a non-empty convex set and $K \subset \mathbb{R}^n$ a non-empty polyhedron such that $\text{relint}(C) \cap K = \emptyset$. Then, there exists a hyperplane separating C and K which does not contain C.

Theorem 3.38
For a problem of the form (3.1) with z as defined as in (3.2), let $S \subset \mathbb{R}^n$ be convex and $\text{epi}_S(z) := (S \times \mathbb{R}) \cap \text{epi}(z)$. If the inequality $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ is Pareto-optimal, then it is also strongly Pareto-optimal.

Proof. We prove the contrapositive of the above statement: Let $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ be an inequality that is not strongly Pareto-optimal. We show that the inequality is not Pareto-optimal, either. By Definition 3.36, there exists an inequality $(\pi'^\top, \pi'_0)(x, \eta)^\top \leq \alpha'$ with $\pi'_0 < 0$ which is valid for $\text{epi}_S(z)$ and dominates $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$. Let

$$S^* := \text{conv}\left(\left\{(x, \eta) \;\middle|\; x \in S, (\pi'^\top, \pi'_0)(x, \eta)^\top \geq \alpha'\right\}\right).$$

Then, since $\pi'_0 < 0$, there is $\eta_x \in \mathbb{R}$ for any $x \in S$ such that $(x, \eta) \in S^*$ for all $\eta \leq \eta_x$ and hence

$$\text{relint}(S^*) \subset \text{conv}\left(\left\{(x, \eta) \;\middle|\; x \in S, (\pi'^\top, \pi'_0)(x, \eta)^\top > \alpha'\right\}\right) \subset \text{conv}(S \times \mathbb{R}).$$

Furthermore, since S is convex,

$$\begin{aligned}\text{relint}(S^*) \cap \text{epi}(z) &= \text{relint}(S^*) \cap \text{conv}(S \times \mathbb{R}) \cap \text{epi}(z) \\ &= \text{relint}(S^*) \cap (S \times \mathbb{R}) \cap \text{epi}(z) = \text{relint}(S^*) \cap \text{epi}_S(z).\end{aligned}$$

As $({\pi'}^\top, \pi'_0)(x,\eta)^\top \leq \alpha'$ is valid for $\mathrm{epi}_S(z)$, this implies that $\mathrm{relint}(S^*) \cap \mathrm{epi}(z) = \emptyset$. Since $\mathrm{epi}(z)$ is a polyhedron, we obtain from Lemma 3.37 that there exists a hyperplane $H^{=}_{(\pi^*,\pi_0^*),\alpha^*}$ such that $\mathrm{epi}(z) \subset H^{\leq}_{(\pi^*,\pi_0^*),\alpha^*}$ and $S^* \subset H^{\geq}_{(\pi^*,\pi_0^*),\alpha^*}$, but not $S^* \subset H^{=}_{(\pi^*,\pi_0^*),\alpha^*}$.

This implies that the inequality $({\pi^*}^\top, \pi_0^*)(x,\eta)^\top \leq \alpha^*$ is valid for $\mathrm{epi}(z)$. Furthermore, it holds that $\pi_0^* \leq 0$: Otherwise, if $\pi_0^* > 0$, then for any $(x,\eta) \in \mathrm{epi}(z)$ there would exist $\bar{\eta}$ large enough such that $(x, \eta+\bar{\eta}) \in \mathrm{epi}(z)$ and $({\pi^*}^\top, \pi_0^*)(x,\eta+\bar{\eta})^\top > \alpha^*$, a contradiction.

We distinguish two cases. First let us assume that $\pi_0^* < 0$. In this case, we show that $({\pi^*}^\top, \pi_0^*)(x,\eta)^\top \leq \alpha^*$ dominates $(\pi^\top, \pi_0)(x,\eta)^\top \leq \alpha$. To see this, let $x \in S$ and

$$\eta_x := \frac{{\pi'}^\top x - \alpha'}{-\pi'_0}.$$

Then, $(x,\eta_x) \in S^*$, which implies that $(x,\eta_x) \in H^{\geq}_{(\pi^*,\pi_0^*),\alpha^*}$. But this means that ${\pi^*}^\top x + \pi_0^* \eta_x \geq \alpha^*$, which, since $\pi_0^* < 0$, implies that for every $x \in S$,

$$\frac{{\pi^*}^\top x - \alpha^*}{-\pi_0^*} \geq \eta_x = \frac{{\pi'}^\top x - \alpha'}{-\pi'_0}.$$

Now, since $({\pi'}^\top, \pi'_0)(x,\eta)^\top \leq \alpha'$ dominates $(\pi^\top, \pi_0)(x,\eta)^\top \leq \alpha$, the same is true for $({\pi^*}^\top, \pi_0^*)(x,\eta)^\top \leq \alpha^*$.

On the other hand, let $\pi_0^* = 0$. Let $\varepsilon > 0$, $\pi' := \pi + \varepsilon\pi^*$ and $\alpha' := \alpha + \varepsilon\alpha^*$. Since both the inequalities $(\pi^\top, \pi_0)(x,\eta)^\top \leq \alpha$ and $(\pi^*)^\top x \leq \alpha^*$ are valid for $\mathrm{epi}(z)$ (the first by assumption, the second by our separation), the same is true for the inequality $({\pi'}^\top, \pi_0)(x,\eta)^\top \leq \alpha'$.

At the same time, we claim that the inequality $({\pi'}^\top, \pi_0)(x,\eta)^\top \leq \alpha'$ dominates $(\pi^\top, \pi_0)(x,\eta)^\top \leq \alpha$: For all $x \in S$, it holds that

$$\frac{{\pi'}^\top x + \alpha'}{-\pi_0} = \frac{\pi^\top x + \alpha}{-\pi_0} + \varepsilon \cdot \frac{{\pi^*}^\top x + \alpha^*}{-\pi_0} \geq \frac{\pi^\top x + \alpha}{-\pi_0},$$

where the last inequality follows from our separation. Since S^* is not contained in $H^{=}_{(\pi^*,\pi_0^*),\alpha^*}$, there exists $x^* \in S$ with $(\pi^*)^\top x^* > \alpha^*$, the last inequality is hence strict for $x^* \in S$ which proves the statement. □

Note that the requirement for S to be convex is indeed necessary: If, for instance, S is a discrete set (one of the typical applications of Benders decomposition), then a Pareto-optimal cut does not need to be strongly Pareto-optimal (see Fig. 3.5).

Analogously to the previous criteria, we define the criterion Pareto for a cut normal as follows:

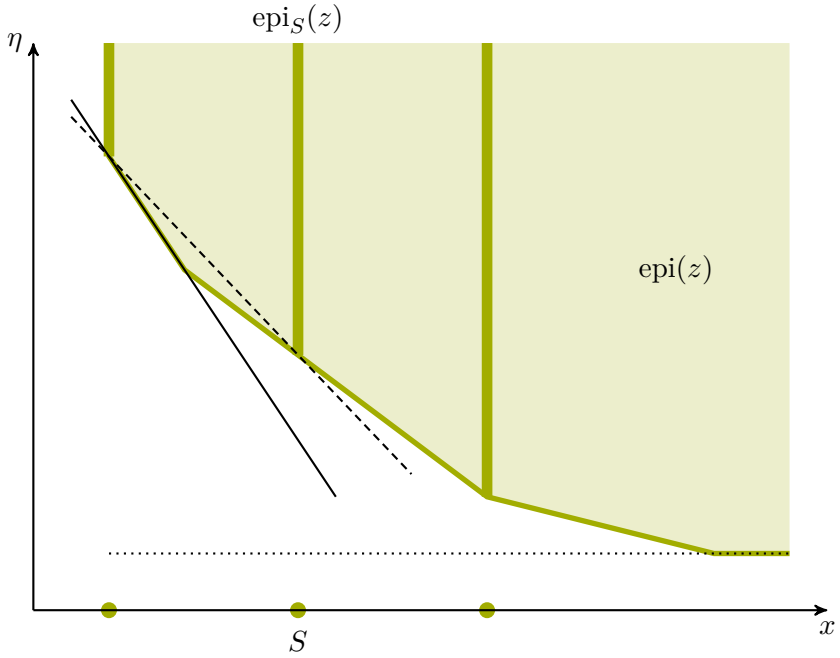


Figure 3.5: In this example, the set S consists of three discrete points, the set $\text{epi}_S(z)$ therefore consists of the three rays shown in dark green. In this case, Theorem 3.38 does not hold: The solid cut is Pareto-optimal, but it is dominated by the dashed cut, which is valid for $\text{epi}_S(z)$, but not $\text{epi}(z)$. Hence, the solid cut is not strongly Pareto-optimal. Furthermore, the dotted cut supports a facet of $\text{epi}(z)$, but does not support it anywhere in S (and actually not even in $\text{conv}(S)$).

Definition 3.39
For a problem of the form (3.1) with z as defined as in (3.2), let $(\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R}$. We say that (π, π_0) *satisfies the criterion* Pareto if there exists a scalar $\alpha \in \mathbb{R}$ such that the inequality $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ is Pareto-optimal.

This criterion seems very reasonable: If a cut is not Pareto-optimal, then it can be replaced by a different cut which is also valid for $\text{epi}(z)$ but leads to a strictly tighter relaxation. We would hence prefer to generate a stronger, Pareto-optimal cut right away.

The following theorem provides us with a characterization of Pareto-optimal cuts. It is based on the idea of [MW81, Theorem 1], which is formulated under the assumption that the subproblem is always feasible (which implies that $\pi_0 < 0$ for any cut normal (π, π_0)). While the original theorem is only concerned with sufficiency, we extend the result in a natural way to obtain a criterion that gives a complete characterization of Pareto-optimal cuts:

Theorem 3.40
For a problem of the form (3.1)*, let* $(\pi, \pi_0) \in \mathbb{R}^n \times \mathbb{R}$ *with* $\pi_0 < 0$*. The inequality* $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ *is Pareto-optimal if and only if* $H^{\leq}_{((\pi,\pi_0),\alpha)}$ *is a halfspace supporting* $\mathrm{epi}(z)$ *in a point* $(x^*, \eta^*) \in \mathrm{epi}(z) \cap \mathrm{relint}(\mathrm{conv}(S)) \times \mathbb{R}$*.*

Proof. For the *if* part, suppose for contradiction that the inequality $(\pi^\top, \pi_0)(x, \eta)^\top \leq \alpha$ is not Pareto-optimal, i. e. there exists some π', π'_0, α' such that the inequality $(\pi'^\top, \pi'_0)(x, \eta)^\top \leq \alpha'$ dominates the former inequality. This means that for all $x \in S$ (and hence all $x \in \mathrm{conv}(S)$), it holds that

$$\frac{\pi'^\top x - \alpha'}{-\pi'_0} \geq \frac{\pi^\top x - \alpha}{-\pi_0} \tag{3.46}$$

and furthermore

$$\frac{\pi'^\top \bar{x} - \alpha'}{-\pi'_0} > \frac{\pi^\top \bar{x} - \alpha}{-\pi_0} \qquad \text{for some } \bar{x} \in S.$$

Finally, since $H^{\leq}_{((\pi,\pi_0),\alpha)}$ supports $\mathrm{epi}(z)$ in (x^*, η^*),

$$\frac{\pi'^\top x^* - \alpha'}{-\pi'_0} \leq \eta^* = \frac{\pi^\top x - \alpha}{-\pi_0} \leq \frac{\pi'^\top x^* - \alpha'}{-\pi'_0}$$

and hence equality must hold everywhere in the above inequality chain. Now, as $x^* \in \mathrm{relint}(\mathrm{conv}(S))$, we can choose $\lambda > 1$ such that $\tilde{x} := \bar{x} + \lambda(x^* - \bar{x}) \in \mathrm{conv}(S)$. But then

$$\begin{aligned}\frac{\pi'^\top \tilde{x} - \alpha'}{-\pi'_0} &= \underbrace{(1-\lambda)}_{<0} \underbrace{\frac{\pi'^\top \bar{x} - \alpha'}{-\pi'_0}}_{> \frac{\pi^\top \bar{x} - \alpha}{-\pi_0}} + \lambda \underbrace{\frac{\pi'^\top x^* - \alpha'}{-\pi'_0}}_{= \frac{\pi^\top \bar{x} - \alpha}{-\pi_0}} \\ &< \frac{\pi^\top \tilde{x} - \alpha}{-\pi_0},\end{aligned}$$

contradicting (3.46).

For the *only-if* part, we first note that if $H^{\leq}_{((\pi,\pi_0),\alpha)}$ does not support $\mathrm{epi}(z)$, then it is obviously dominated by $H^{\leq}_{((\pi,\pi_0),\alpha')}$ with $\alpha' := \alpha' + \varepsilon$ for some $\varepsilon > 0$. Therefore, let $H^{\leq}_{((\pi,\pi_0),\alpha)}$ be such that it supports $\mathrm{epi}(z)$, but not in points from the set $\mathrm{epi}(z) \cap \mathrm{relint}(\mathrm{conv}(S)) \times \mathbb{R}$. Denote by $S^* := \{x \in \mathbb{R}^n \mid \exists \eta : (x, \eta) \in \mathrm{epi}(z) \cap H^{=}_{((\pi,\pi_0),\alpha)}\}$ the set of points where $H^{\leq}_{((\pi,\pi_0),\alpha)}$ supports $\mathrm{epi}(z)$.

Since $\mathrm{relint}(\mathrm{conv}(S)) \cap S^* = \emptyset$, we can use Lemma 3.37 to obtain a hyperplane separating $\mathrm{conv}(S)$ and S^* which does not contain S. Hence, there exist π^*, α^* such that $\pi^{*\top} x \geq \alpha^*$ for all $x \in S^*$ and $\pi^{*\top} x \leq \alpha^*$ for all $x \in \mathrm{conv}(S)$, where the second inequality is strict for some $x \in \mathrm{conv}(S)$ and thus also for some $x^* \in S$.

Let $\varepsilon > 0$, $\pi' := \pi - \varepsilon\pi^*$ and $\alpha' := \alpha - \varepsilon\alpha^*$. If ε is sufficiently small, then the inequality $(\pi'^\top, \pi_0)(x,\eta)^\top \leq \alpha'$ is valid for $\text{epi}(z)$: All $(x,\eta) \in \text{epi}(z)$ with $x \notin S^*$ satisfied the original inequality strictly and for all $x \in S^*$,

$$\begin{aligned}(\pi'^\top, \pi_0)(x,\eta)^\top - \alpha' &= (\pi^\top, \pi_0)(x,\eta)^\top - \alpha - \varepsilon \underbrace{(\pi^{*\top} x - \alpha^*)}_{\geq 0} \\ &\leq (\pi^\top, \pi_0)(x,\eta)^\top - \alpha \leq 0,\end{aligned}$$

since the original inequality was valid for $\text{epi}(z)$.

Finally, we claim that the inequality $(\pi^\top, \pi_0)(x,\eta)^\top \leq \alpha$ is dominated by the inequality $(\pi'^\top, \pi_0)(x,\eta)^\top \leq \alpha'$: For all $x \in S$, it holds that

$$\frac{\pi'^\top x - \alpha'}{-\pi_0} = \frac{\pi^\top x - \alpha}{-\pi_0} + \varepsilon \cdot \frac{\pi^{*\top} x - \alpha^*}{-\pi_0} \geq \frac{\pi^\top x - \alpha}{-\pi_0}.$$

Since the last inequality is strict for $x^* \in S$, this proves the statement. □

For the case where S is convex, the previous theorem immediately implies the following statement (remember that $\text{epi}_S(z) := (S \times \mathbb{R}) \cap \text{epi}(z)$):

Corollary 3.41
In particular, by the above theorem, if S is convex and $H^{\leq}_{((\pi,\pi_0),\alpha)}$ supports a facet of $\text{epi}_S(z)$, then $(\pi^\top, \pi_0)(x,\eta)^\top \leq \alpha$ is Pareto-optimal (and hence strongly Pareto-optimal).

Magnanti and Wong [MW81] also provide an algorithm that computes a Pareto-optimal cut by solving the cut-generating problem twice. While their algorithm is defined for the original Benders optimality cut (Lemma 3.4 d)), it can be adapted to work with other cut selection criteria, as well. Sherali and Lunday [SL13] present a method based on multi-objective optimization to obtain a cut that satisfies a weaker version of Pareto-optimality by solving only a single instance of the cut-generating LP.

Papadakos [Pap08] notes that, given a point in the relative interior of $\text{conv}(S)$, a Pareto-optimal cut can be generated using a single run of the cut-generating problem. Also, under certain conditions on the problem, other points not in the relative interior allow this, as well. However, the approach suggested by the authors adds Pareto-optimal cuts independently from master- or subproblem solutions, together with subproblem-generated cuts, which are generally not Pareto-optimal. This means that the Pareto-optimal cuts which are added may not even cut off the current tentative solution. The upcoming Theorem 3.43 will lead to an approach that reconciles both objectives, generating cuts that are always Pareto-optimal, but also cut off the current tentative solution.

We use a result by Cornuéjols and Lemaréchal [CL06] on the set of points exposed by a cut normal (π, π_0) to derive a method that always obtains a Pareto-optimal cut. The following lemma has been slightly generalized and rewritten to match our setting and notation, but it follows the general idea of Cornuéjols and Lemaréchal [CL06, Theorem 3.4].

Lemma 3.42
Let $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R} \setminus \operatorname{epi}(z)$, $(\omega, \omega_0) \in \operatorname{pos}(\operatorname{epi}(z) - (x^*, \eta^*)) \setminus \{0\}$ *and let* (π, π_0) *be optimal in* $(\operatorname{epi}(z) - (x^*, \eta^*))^-$ *with respect to the objective* (ω, ω_0). *Then there exists* $\alpha \in \mathbb{R}$ *such that* $H^{\leq}_{((\pi,\pi_0),\alpha)}$ *supports* $\operatorname{epi}(z)$ *in*

$$(\bar{x}, \bar{\eta}) := \frac{(\omega, \omega_0)}{-h_Q(\omega, \omega_0)} + (x^*, \eta^*),$$

where $Q := (\operatorname{epi}(z) - (x^*, \eta^*))^-$.

The case of $(\omega, \omega_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))$ was proven by Cornuéjols and Lemaréchal [CL06, Theorem 3.4]. If $(\omega, \omega_0) \in \operatorname{pos}(\operatorname{epi}(z) - (x^*, \eta^*)) \setminus \{0\}$, then there is $\mu > 0$ such that $\mu \cdot (\omega, \omega_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))$. Note that if (π, π_0) is optimal with respect to (ω, ω_0), then also with respect to $\mu \cdot (\omega, \omega_0)$. We thus have by Cornuéjols and Lemaréchal [CL06, Theorem 3.4] that there exists $\alpha \in \mathbb{R}$ such that $H^{\leq}_{(\pi,\pi_0),\alpha}$ supports $\operatorname{epi}_S(z)$ in

$$(\bar{x}, \bar{\eta}) := \frac{\mu \cdot (\omega, \omega_0)}{-h_Q(\mu\omega, \mu\omega_0)} + (x^*, \eta^*) = \frac{(\omega, \omega_0)}{-h_Q(\omega, \omega_0)} + (x^*, \eta^*).$$

We can now prove the theorem already mentioned above.

Theorem 3.43
Let $(x^*, \eta^*) \in S \times \mathbb{R}$ *and* $(\omega, \omega_0) \in \operatorname{relint}(\operatorname{conv}(\operatorname{epi}_S(z) - (x^*, \eta^*)))$. *Let* (π, π_0) *be optimal in* $(\operatorname{epi}(z) - (x^*, \eta^*))^-$ *with respect to the objective* (ω, ω_0) *and let* $\pi_0 < 0$. *Then* (π, π_0) *satisfies the criterion* PARETO.

Proof. Again, let $Q := (\operatorname{epi}(z) - (x^*, \eta^*))^-$ and let $\lambda := -(h_Q(\omega, \omega_0))^{-1}$. Observe that $(\omega, \omega_0) \in \operatorname{epi}_S(z) - (x^*, \eta^*)$ and therefore, by the definition of the reverse polar set, $h_Q(\omega, \omega_0) \leq -1$ and thus $\lambda \in (0, 1]$.

Note that, in particular, $(\omega, \omega_0) \in \operatorname{pos}(\operatorname{epi}(z) - (x^*, \eta^*)) \setminus \{0\}$. For $(\bar{x}, \bar{\eta})$ from Lemma 3.42, we thus obtain that $(\bar{x}, \bar{\eta}) = \lambda\left((\omega, \omega_0) + (x^*, \eta^*)\right) + (1-\lambda)(x^*, \eta^*)$ is a convex combination of a point $(\omega, \omega_0) + (x^*, \eta^*) \in \operatorname{relint}(\operatorname{conv}(\operatorname{epi}_S(z))) \subset \operatorname{relint}(\operatorname{conv}(S)) \times \mathbb{R}$ and $(x^*, \eta^*) \in S \times \mathbb{R}$. Since $\lambda > 0$, it holds that $\bar{x} \in \operatorname{relint}(\operatorname{conv}(S))$ and thus by Theorem 3.40 the cut defined by (π, π_0) is Pareto-optimal. □

Note that the above theorem holds for all optimal solutions (and not only for extremal optimal solutions). As we know from Theorem 3.17, optimality with respect to (ω, ω_0)

over the reverse polar set is equivalent to optimality with respect to $(H\omega, -\omega_0)$ over the alternative polyhedron and hence Theorem 3.43 can be used with the (relaxed) alternative polyhedron, as well.

To conclude this section on Pareto-optimality, we would like to point out two observations concerning our results in relation to [Pap08]:

First, Theorem 3.43 requires, again, a relative interior point of $\operatorname{conv}(\operatorname{epi}_S(z)-(x^*,\eta^*))$, just as the original approach by Magnanti and Wong [MW81] required a relative interior point of S. Naturally, since $\operatorname{conv}(\operatorname{epi}_S(z)-(x^*,\eta^*))$ is a convex set, it holds that any convex combination of a relative interior point and any other point contained in the set again yields a relative interior point. This statement is analogous to [Pap08, Theorem 8] and allows us to use an iterative update procedure for the point (ω,ω_0) similar to the procedure used in the computational experiments in [Pap08]. Furthermore, as we have mentioned above, any Pareto-optimal cut generated using Theorem 3.43 takes into account the current tentative solution (x^*,η^*), which makes sure that this solution is indeed cut off (which is not guaranteed for an arbitrary Pareto-optimal cut) and also refines the approximation of $\operatorname{epi}(z)$ in the proximity of the (infeasible) solution that yields the current lower bound.

Secondly, as Papadakos [Pap08] observed, supporting $\operatorname{epi}(z)$ in a relative interior point of $\operatorname{conv}(S)$ is sufficient, but not necessary to obtain a Pareto-optimal cut. Papadakos calls a point x a *Magnanti-Wong point*, if every cut that supports $\operatorname{epi}(z)$ in x is Pareto-optimal and they provide some additional criteria for when a point is a Magnanti-Wong point. In the context of the above theorem, this observation means that we obtain a Pareto-optimal cut whenever the point $\bar{x}$ in the proof of Theorem 3.43 is a Magnanti-Wong point. For instance, since the convex hull of a Magnanti-Wong point and the relative interior of $\operatorname{conv}(S)$ consists of Magnanti-Wong points itself ([Pap08, Theorem 8]), it is quite common that parts of the relative boundary of $\operatorname{conv}(S)$ consist of Magnanti-Wong points. If this is true for the entire relative boundary, then any $(\omega,\omega_0)\in\operatorname{conv}(\operatorname{epi}_S(z)-(x^*,\eta^*))$ yields a Pareto-optimal cut.

3.2.4 Summary

In this section, we have presented three different cut selection criteria from the literature. We have compared them to each other and discussed the problem of selecting a cut that satisfies each of the different criteria. The relationship between the different criteria can be summarized as follows:

As observed by Fischetti, Salvagnin, and Zanette [FSZ10], every cut derived from an extremal point of the alternative polyhedron corresponds to some minimal infeasible subsystem of the system (3.8) of linear inequalities (Corollary 3.27). Some of these extremal points correspond to extremal points of the reverse polar set, these are of particular interest because they lead to cuts that define facets of the set $\operatorname{epi}(z)$ (Corollary 3.31). While there is no easy way to guarantee that an extremal point of

Table 3.1: Guaranteed properties of the cut resulting from an extremal point in the alternative polyhedron which maximizes $(\tilde{\omega}, \tilde{\omega}_0)$ (under the assumption that a finite optimum exists). The checkmark in parentheses (✓) indicates that the property is almost always satisfied.

$(\tilde{\omega}, \tilde{\omega}_0)$	MIS	Facet	Pareto
$(\tilde{\omega}, \tilde{\omega}_0) \in \mathbb{R}^{m+1}$	✓	✗	✗
$(\tilde{\omega}, \tilde{\omega}_0) \in (H, -1) \cdot \mathbb{R}^{n+1}$	✓	(✓)	✗
$(\tilde{\omega}, \tilde{\omega}_0) \in (H, -1) \cdot \operatorname{relint}(\operatorname{conv}(\operatorname{epi}_S(z) - (x^*, \eta^*)))$	✓	(✓)	✓

the reverse polar set is selected as the optimum according to some linear objective over the alternative polyhedron, we can choose the objective in such a way that there always exists such a point and it is selected by *almost all* objective functions in the sense that the set of objective vectors $(\tilde{\omega}, \tilde{\omega}_0)$ that select no such point is of lower dimension (Theorem 3.32). In principle, we can also guarantee that we obtain an facet-defining cut by successively solving n different linear programs (Lemma 3.33). While this is interesting theoretically, we can think of no application where the result would justify the enormous amount of computational effort that this requires, instead we will generally be content with the high likelihood to obtain a facet-defining cut or solve the subproblem one more time to further increase the probability (Theorem 3.34). Finally, if the cut supports $\operatorname{epi}(z)$ in a relative interior point of $\operatorname{conv}(S)$ (which we can guarantee via the objective function, see Theorem 3.43), then it is furthermore Pareto-optimal (Theorem 3.40).

A summary of the results above is given in Table 3.1. We say that a criterion is *almost always satisfied*, if the set of objective vectors $(\tilde{\omega}, \tilde{\omega}_0)$ for which it is violated is of lower dimension than the specified domain.

Beyond the three criteria Facet, MIS and Pareto mentioned above, other criteria for the selection of Benders cuts have been proposed in the literature. Most notably, Saharidis, Minoux, and Ierapetritou suggest to generate cuts which include those master variables that are not *covered* by previously generated cuts (e. g., [SMI10; Aza+13; TJS13]). As a criterion about the relation between *different* cuts rather than selection criteria for individual cuts, this approach cannot easily be compared to the criteria above and may in fact even be used alongside those criteria (e. g. requiring every cut from a covering cut approach to be facet-defining). The performance of such an approach might be an interesting topic for further research.

Another very natural selection criterion is the depth of a cut (its violation by the current solution (x^*, η^*)). Cornuéjols and Lemaréchal [CL06] prove that the deepest cut (according to the Euclidean norm) is always that obtained from optimizing over the reverse polar set in the direction given by the orthogonal projection of the current solution onto the target polyhedron. They also mention personal communication with Pierre Bonami about the (euclidean) depth of cuts in the context of lift-and-project.

Bonami compared two selection criteria for cuts, where the one that led to deeper cuts than the other also slightly improved the effectiveness of the cutting process. This may be seen as giving at least some indication towards the effectiveness of deep cuts.

Finally, we want to note that Benders Decomposition shares a number of properties with the setting of disjunctive programming (see [Bal79]), both with respect to the problem classes to which it can be applied and to the solution technique itself: In both cases, we want to optimize over a set that is known to be convex, even polyhedral, but no explicit representation of the set is available or can be computed with reasonable effort. However, given a point (that represents a tentative optimal solution), the separation problem for this point and the target set can be solved *easily* and the obtained cutting planes have a number of useful properties (e. g. they can be chosen to always support the target set).

Consequently, many elements of the theory developed by Balas and others (e. g., [Bal79; Bal98; CL06]) in the context of disjunctive programming can be applied to Benders decomposition. One important example is the characterization from [Bal98] (subsequently rephrased by Cornuéjols and Lemaréchal [CL06]) of facet-defining inequalities in terms of the reverse polar set, which corresponds to our Theorem 3.30. Analogously to the situation in the case of Benders decomposition, the reverse polar set is most conveniently available as the projection of a higher-dimensional set which includes the dual variables of the subproblem.

We have mentioned previously that the convenient representation of the reverse polar set using dual variables comes with the caveat that additional vertices may appear as basic optimal solutions. These do not necessarily correspond to vertices of the reverse polar set and hence do not lead to facet-defining inequalities. The same is true in the case of disjunctive programming. The "good" optimal solutions (those which lead to facet-defining inequalities) are termed *regular* optimal solutions in Balas [Bal98]. However, the algorithmic problem of selecting these solutions is not treated in the cited paper.

3.3 Special Problem Structures

Before we turn to the application of our results to problems from the context of energy network optimization, we briefly discuss the implications of structural properties of a problem description like the one given in (3.1). We will first consider the case where the set S is polyhedral and we may hence choose to integrate the description of S directly into the subproblem. Secondly, if the matrix A is block-diagonal then the subproblem can be divided into several problems that can be solved independently. We discuss the implications and the tradeoffs that result from different possible choices regarding the aggregation of subproblems. Finally, we mention a modeling technique that is occasionally used to simplify the implementation of subproblem and cut generation. It

turns out that this technique has the useful side effect that the condition (3.25) is very easy to satisfy, albeit at the cost of a slightly larger description of the subproblem.

3.3.1 Polyhedral S

In this section, we consider the special case of problems of the form (3.1) where the set S, which restricts the x-variables independently from the y-variables, is a polyhedron for which an explicit $\mathcal{H}$-representation is available. At first sight, one might think that this would allow us to select stronger cuts. Looking at Fig. 3.6, we see that there are indeed certain cuts that separate (x^*, η^*) from $\text{epi}_S(z)$ which would violate $\text{epi}(z)$. However, it will turn out that these additional cuts are of no use practically (and against this background, we avoid overloading this section with too much formalism and focus on conveying only the general ideas).

Indeed, the best cuts according to our criteria when separating from $\text{epi}_S(z)$ are also among the best when separating from $\text{epi}(z)$. The increased set of possible cut normals instead leads to the cut selection problem for $\text{epi}_S(z)$ being unbounded in some cases where the problem for $\text{epi}(z)$ would be finite. Furthermore, once we apply our selection criteria, we actually have *more* potential cuts to choose from when separating from $\text{epi}(z)$ than when separating from $\text{epi}_S(z)$: The corresponding reverse polar set has more vertices. These additional cuts however are not Pareto-optimal and we can avoid selecting them by using a suitable objective.

In the setting mentioned above where S is a polyhedron, we can define a variant of the function z from (3.2) as follows:

$$z_S(x) := \min_{y \in \mathbb{R}^k} \left\{ d^\top y \;\middle|\; Ay \leq b - Hx, x \in S \right\}$$

This corresponds to setting $z(x)$ to $+\infty$ whenever $x \notin S$. As S is a polyhedron for which an explicit $\mathcal{H}$-representation is available, the set

$$\text{epi}(z_S) = \text{epi}(z) \cap (S \times \mathbb{R})$$

is no more difficult to deal with than the original set $\text{epi}(z)$ and all statements made above about $\text{epi}(z)$ also hold for $\text{epi}(z_S)$.

In addition, we have $\text{epi}(z_S) = \text{epi}_S(z)$. This obviously means that every cut that supports $\text{epi}(z_S)$ also supports $\text{epi}_S(z)$ and furthermore it supports a facet of $\text{epi}(z_S)$ if and only if it supports a facet of $\text{epi}_S(z)$. In particular, this implies by Theorem 3.40 that Facet $\Rightarrow$Pareto. We can use the following definition to describe the admissible cut normals to separate a tentative solution from $\text{epi}_S(z)$ directly.

Definition 3.44
Let a problem of the form (3.1) be given where $S := \{x \in \mathbb{R}^n \mid Bx \leq e\}$ is a polyhedron, and $(x^*, \eta^*) \in \mathbb{R}^n \times \mathbb{R}$. The *relaxed extended alternative polyhedron* $P_S(x^*, \eta^*)$ is given

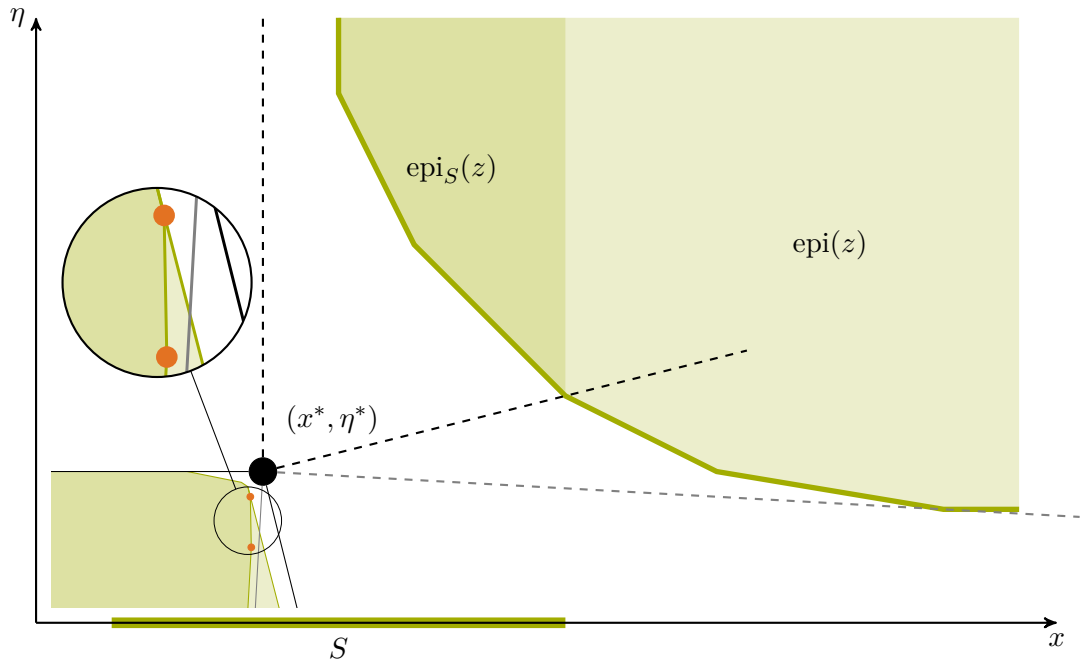


Figure 3.6: The reverse polar sets $(\text{epi}_S(z) - (x^*, \eta^*))^-$ (shaded) and $(\text{epi}(z) - (x^*, \eta^*))^-$ (darkly shaded). Both are contained in the corresponding polar cone, which is shown in black and grey, respectively. It can be observed that while $(\text{epi}(z) - (x^*, \eta^*))^- \subset (\text{epi}_S(z) - (x^*, \eta^*))^-$, the former contains two additional vertices, shown in orange. These correspond to the facets of epi(z) which are not Pareto-optimal cuts with respect to epi$_S(z)$.

by

$$P_S(x^*, \eta^*) := \left\{ \gamma, \gamma_S, \gamma_0 \geq 0 \;\middle|\; \begin{array}{c} \gamma^\top A + \gamma_S^\top B + \gamma_0 d^\top = 0 \\ \gamma^\top (b - Hx^*) + \gamma_S^\top e + \gamma_0 \eta^* \leq -1 \end{array} \right\}.$$

Analogously to Theorem 3.13, it can easily be seen that

$$(\text{epi}_S(z) - (x^*, \eta^*))^- = \begin{pmatrix} H^\top & 0 \\ 0 & -1 \end{pmatrix} \cdot P_S(x^*, \eta^*).$$

The polar cones and reverse polar sets with respect to the sets epi$_S(z)$ and epi(z) are shown in Fig. 3.6. On closer inspection, we can observe that while it is true that every vertex of the extended reverse polar set corresponds to a Pareto-optimal cut, which is not true for the original reverse polar set, this carries little practical significance: For every linear objective that would select a vertex of the reverse polar set which does not correspond to a Pareto-optimal cut, the extended reverse polar set is simply

unbounded. In the context of Table 3.1, where we restrict ourselves to objectives that yield a finite optimal solution, this means that the second and third rows coincide (every facet-defining cut normal is also Pareto-optimal).

In this sense, Pareto-optimality is much more a function of the linear objective used to generate a cut than of the feasible region over which that function is optimized: If a linear objective leads to a Pareto-optimal cut, it does so in both the original and the extended reverse polar set. If, on the other hand, a linear objective does not lead to a Pareto-optimal cut, it produces a suboptimal cut in the original reverse polar set and no cut at all in the extended reverse polar set (due to unboundedness).

In the latter case, we would have to re-solve the subproblem with a modified objective function in order to obtain a cut. While in most applications, we would probably prefer obtaining a non-Pareto-optimal cut to having to re-solve the subproblem to obtain a Pareto-optimal cut, there might be cases where we would actually prefer the latter(i. e., if we need to keep the number of cuts small), which we can achieve using Definition 3.44.

3.3.2 Multiple Subproblems and Multi-Cuts

As stated in Section 3.1, one typical application of Benders decomposition is the case where the matrix of subproblem constraints A has a block-diagonal structure. Let us denote the set of blocks by T. Once the values of x-variables have been fixed, this means that the remaining problem decomposes into a set of $|T|$ independent optimization problems over sub-vectors y^t of y for all $t \in T$.

This opens the possibility for a slightly different approach towards representing the original problem than the one used in Section 3.1: With

$$z_t(x) := \min_{y^t \in \mathbb{R}^{k_t}} \left\{ {d^t}^\top y^t \;\middle|\; A^t y^t \leq b^t - H^t x \right\},$$

the original problem can be represented as follows:

$$\begin{aligned} \min\ & c^\top x + \sum_{t \in T} z_t(x) \\ & x \in S. \end{aligned}$$

Analogously to Section 3.1, we can define

$$\operatorname{epi}_S^T(z) := \left\{ (x, \eta^1, \ldots, \eta^{|T|}) \in S \times \mathbb{R}^{|T|} \;\middle|\; \forall\ t \in T : (x, \eta^t) \in \operatorname{epi}(z_t) \right\}$$

to write

$$\begin{aligned} \min\ & c^\top x + \sum_{t \in T} \eta^t \\ & (x, \eta^1, \ldots, \eta^{|T|}) \in \operatorname{epi}_S^T(z). \end{aligned}$$

As above, the original problem can now be solved by optimizing over the set $\operatorname{epi}_S^T(z)$: We can use a version of the Benders decomposition algorithm that in every iteration generates a cut for each $t \in T$, separating the current solution from the set $\operatorname{epi}(z_t)$ until we obtain a solution that is contained in $\operatorname{epi}_S^T(z)$. The procedure to generate such a cut for an individual $\operatorname{epi}(z_t)$ is identical to that for generating a cut for $\operatorname{epi}(z)$ in the standard setting and all our results translate immediately.

It turns out (unsurprisingly) that the *multi-cut* approach outlined above yields a closer approximation of the subproblem cost function $\sum_{t \in T} z_t(x)$ (see, e. g., [BL88]). On the other hand, in each iteration, we add $|T|$ cuts instead of one to the master problem, which increases the rate at which the master problem grows between iterations. For small T, this does not pose too much of a problem, but if T is very large, it may quickly cause the master problem itself to become unpractically large and difficult to solve.

For the latter case, Birge and Louveaux [BL88] suggest to aggregate blocks of A into larger blocks. This enables us to explicitly choose the size of the set T and hence to strike different tradeoffs with respect to the conflict described above. Note that for $|T| = 1$, we consider the matrix A as a single "block", which takes us back to the original approach. With the right choice regarding the size of the set T, we might thus expect an improvement in the performance of the algorithm. However, the ideal size of T depends on the particular problem instance and we are not aware of any theoretical results regarding the optimal choice in this context. Some researchers have empirically investigated schemes where over the course of the solution process, a switch is made from single cuts to multiple cuts or vice versa [SDT14]. Their work seems to suggest that it is advantageous to start with aggregated cuts and move to more refined, separate cuts later on.

With respect to our results in this chapter, another difference between these two approaches has to be considered: In the original approach, we have to use a single objective ω for the entire subproblem in order to be able to select a cut that defines a facet of the set $\operatorname{epi}(z)$. In the multi-cut approach, we may choose a different objective ω^t for each $t \in T$ to generate a cut that defines a facet of the set $\operatorname{epi}(z)^t$. This may prove to be useful, particularly in cases where the blocks indexed by T are very different.

3.3.3 Simplified Coupling Constraints

The fact that extremal points of the alternative polyhedron do not necessarily correspond to extremal points of the reverse polar set (as pointed out in Section 3.2.2) hinges on the fact that the linear transformation from Theorem 3.13, which links the two polyhedra, is generally not a full-rank transformation. While we have developed techniques that allow us to overcome these distinctions and in most cases generate facet-defining cuts directly from the alternative polyhedron, they require some mathematical understanding of master- and subproblem as well as their connection via the interaction matrix H.

We would therefore like to point out a special situation, which is related to a modeling technique which is occasionally used in the context of Benders decomposition to simplify the notation and implementation of the subproblem, as well as the computations required for cut generation (see, e. g., [GLM99; AC00; Pap08]):

The Benders subproblem that results from fixing the x-variables is equivalent to the optimization problem

$$\begin{aligned} \min_{x,y}\ & d^\top y \\ \text{s.t. } & Ay + Hx \leq b \\ & x - x^* = 0 \\ & y \in \mathbb{R}^k, x \in \mathbb{R}^n, \end{aligned} \tag{3.47}$$

in which this fixing is taken care of by the extra constraint $x - x^* = 0$. This formulation theoretically increases the problem size, however most LP solvers are easily capable of reverting this blow-up by substituting the values of x^* during pre-processing. On the other hand, it nicely separates the complexity of the subproblem from the computation of Benders cuts: In the cut definition, only the dual variables of the constraints $x - x^* = 0$ appear as coefficients, if the objective value of the subproblem is known, then all other dual variables can be ignored. Similarly, it is easy e. g. to add constraints to the subproblem (as long as they do not introduce dependencies of additional master variables) without making any changes to the cut generation procedure.

While this technique can in certain cases make the practical implementation easier, we are not aware of any literature that discusses the implications of this technique in the context of cut selection or evaluates the performance impact of the technique. In the context of the cut selection procedure proposed by Fischetti, Salvagnin, and Zanette [FSZ10], one might be tempted to think of it as a limitation, since we can no longer choose an individual component of $\tilde{\omega}$ for each coupling constraint in the original problem. Instead, we must choose a value for each coupling *variable*, which can be translated into values for each constraint by a transformation which depends on the entries of the interaction matrix H. In particular, depending on the entries of H, it might no longer be possible to realize the 0-1-objective vector proposed in [FSZ10].

On closer inspection however, it turns out that a side effect of the above transformation is that the interaction matrix for the new subproblem takes the form

$$\begin{pmatrix} 0 \\ -I_n \end{pmatrix}$$

where I_n denote the n-dimensional identity matrix. This means that condition (3.25) on the objective function over the alternative polyhedron is satisfied by all objectives that have zero entries corresponding to the null rows of the above interaction matrix. This means that the apparent restriction of the choice of objective vectors $(\tilde{\omega}, \tilde{\omega}_0)$

coincides exactly with the restriction that is required to obtain cuts satisfying the criterion FACET in most cases.

In terms of the problem formulation from Corollary 3.22, this becomes

$$\begin{aligned} \min\ & \lambda \\ & Ay + Hx \leq b \\ & x = x^* - \lambda \cdot \bar{\omega} \\ & d^\top y \leq \eta^* - \lambda \bar{\omega}_0 \end{aligned}$$

where $\bar{\omega}$ can be chosen arbitrarily and always satisfies the condition (3.25). As a consequence, the entire theory presented in this chapter could also be derived by restricting the problem formulation to the (blown-up) form (3.47) instead of restricting the subproblem objective in the way formulated in (3.25). As we have seen, however, this blow-up is unnecessary and our approach provided us with a clearer understanding of the underlying structure. The result with respect to the formulation (3.47) results as a special case.

3.4 Benders Decomposition for the Capacity Expansion Problem

From a practical perspective, the main results in this chapter can be summed up as follows: Within the Benders cut selection framework from [FSZ10], we have investigated the effect of different choices of the objective vector $(\tilde{\omega}, \tilde{\omega}_0)$. We proved that choosing the vector from certain subsets of $\mathbb{R}^m \times \mathbb{R}$ produces cuts that satisfy desirable properties. In particular, the objective vectors that lead to these desirable cuts can be parametrized by a vector $(\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$ that lives in the same space as the original x-variables which constitute our master problem. This opens the path to using problem-specific knowledge or partial information about the optimal solution, where available, to improve the solution algorithm.

In this section, we apply the theory presented above to a version of the capacity expansion problem from Section 2.1.1, which we extend by adding variable production capacities in addition to variable transmission capacities. This section thus fulfills two distinct purposes: It serves as an example how an implementation of Benders decomposition and in particular of the improvements developed in this chapter can work in practice. In this context, we also point out some practical challenges that arise in the implementation and discuss how these can be overcome. Secondly, we aim to provide a workable template that enables researchers to use our improved version of Benders decomposition in the context of an existing capacity expansion model.

Benders decomposition has been used to solve Capacity Expansion Problems in the context of electrical power systems at least since the 1980s (see, e. g., [Blo83]). The

type of problem nicely fits the common informal characterization of problems suitable to using Benders decomposition as those with a set of *complicating variables*: If the expansion variables are fixed, the resulting dispatch problem is typically much easier to solve than the original Capacity Expansion Problem. Over the last twenty years, references to Benders decomposition in the optimization literature have multiplied [Rah+17] and power system analysis has become one of the most prominent areas of its application, some well-known examples include [RM94; AC00; BPG01; Wan+16]. Most recently, certain studies have focussed on evaluating *acceleration techniques*, including cut selection criteria, specifically on applications from the energy sector [Jen+15; LR16]. They do not, however, cover in detail the criteria discussed in this chapter.

For the sake of a clearer presentation, we begin by considering a simpler version of the optimization problems typically used in practice in the context of Energy System Optimization. In particular, we omit storage facilities and consider only a single investment decision in order to keep the notation easier to read. We also assume that the production cost vector c does not vary between different time steps (in deviation from our discussion in (2.12)). In practice, all of these restrictions can be removed and the presented techniques can be applied to more complicated optimization models which include the aspects mentioned above with only small modifications. We will discuss the details of re-including the aforementioned complexities (as well as various other constraints) and their implications in Section 3.4.3. The problem that we focus on in this section is defined as follows:

Definition 3.45
Let $\bar{p}^{\max} \in \mathbb{R}^{|\mathcal{I}|}$ and $\bar{f}^{\max} \in \mathbb{R}^{|\mathcal{L}|}$ denote vectors of upper bounds for the capacity of production units and transmission lines, respectively. Let furthermore $c^p \in \mathbb{R}^{|\mathcal{I}|}$, $c^f \in \mathbb{R}^{|\mathcal{L}|}$ and $c \in \mathbb{R}^{|\mathcal{I}|}$ denote three cost vectors, for production capacity, transmission capacity and dispatch, respectively. Finally, let $T \in \mathbb{Z}$ and let $P_t \subset \mathbb{R}^{(|\mathcal{I}|+|\mathcal{L}|)}$ be a polyhedron for all $t \in [T]$. Denote by $P := \prod_{t\in[T]} P_t$.

A vector $(\bar{p}, \bar{f}) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^{|\mathcal{L}|}$ of capacities that satisfies $\bar{p} \leq \bar{p}^{\max}$ and $\bar{f} \leq \bar{f}^{\max}$ is called *feasible at cost* η if there exist vectors $(p^t, f^t) \in P_t$ satisfying $p^t \leq \bar{p}$ and $-\bar{f} \leq f^t \leq \bar{f}$ for all $t \in [T]$ and furthermore $\sum_{t\in[T]} c^\top p^t \leq \eta$. It is called *optimal* if there is no $(\bar{p}', \bar{f}')$ feasible at cost η' such that

$$(c^p)^\top \bar{p}' + (c^f)^\top \bar{f}' + \eta' < (c^p)^\top \bar{p} + (c^f)^\top \bar{f} + \eta.$$

The polyhedra P_t can be thought of as representing a set of operating constraints that restrict the production and transmission vectors p^t and f^t for each timestep $t \in [T]$, but that do not depend on the values of $\bar{p}$ and $\bar{f}$. The most common example of such constraints are demand satisfaction constraints, but each P_t may be defined by an arbitrary selection of such constraints. Note that these constraints need not be the same for all timesteps, for instance due to changes in demand or differences in the

availability of renewable power sources. An optimal solution to the capacity expansion problem can be computed using the following linear program:

$$\min_{\substack{\bar{p},\bar{f}\\p,f}} (c^p)^\top \bar{p} + (c^f)^\top \bar{f} + \sum_{t\in[T]} c^\top p^t \tag{3.48}$$

$$\text{s.t. } \bar{p} \le \bar{p}^{\max} \tag{3.49}$$

$$\bar{f} \le \bar{f}^{\max} \tag{3.50}$$

$$p^t \le \bar{p} \qquad \forall\, t \in [T] \tag{3.51}$$

$$-\bar{f} \le f^t \le \bar{f} \qquad \forall\, t \in [T] \tag{3.52}$$

$$(p^t, f^t) \in P_t \qquad \forall\, t \in [T] \tag{3.53}$$

Note that we could replace $\bar{p}$ and $\bar{f}$ by the expressions $\max_{t\in[T]} p^t$ and $\max_{t\in[T]} |f^t|$, respectively. A standard linearization of the resulting non-linear objective function again yields the linear program (3.48) to (3.53).

While this linear program is of course theoretically solvable in polynomial time, we will assume in this chapter that the number of timesteps in T is extremely large which makes the optimization problem nonetheless difficult to solve in practice. On the other hand, for any fixed timestep t (or small subset of timesteps) and fixed values of $\bar{p}$ and $\bar{f}$, the problem consisting only of the corresponding variables p^t and f^t and constraints (3.51) to (3.53) is relatively small and can be solved quickly.

This motivates the following approach: Let

$$S := \left\{ (\bar{p}, \bar{f}) \;\middle|\; \begin{matrix} \bar{p} \le \bar{p}^{\max} \\ \bar{f} \le \bar{f}^{\max} \end{matrix} \right\} \tag{3.54}$$

and

$$z(\bar{p}, \bar{f}) := \max_{p,f} \left\{ \sum_{t\in[T]} c^\top p^t \;\middle|\; \begin{matrix} p^t \le \bar{p} & \forall\, t \in [T] \\ -\bar{f} \le f^t \le \bar{f} & \forall\, t \in [T] \\ (p^t, f^t) \in P_t & \forall\, t \in [T]. \end{matrix} \right\} \tag{3.55}$$

Instead of solving the huge $(|\mathcal{I}| + |\mathcal{L}|) \cdot (T+1)$-dimensional Linear Program (3.48) to (3.53), we solve the much smaller $(|\mathcal{I}| + |\mathcal{L}| + 1)$-dimensional problem of minimizing $(c^p)^\top \bar{p} + (c^f)^\top \bar{f} + \eta$ over the polytope $\operatorname{epi}_S(z)$ (see Proposition 3.2). Alternatively, since the matrix of constraints that define the set P is block-diagonal with each $t \in [T]$ inducing one block that consists of the constraints from P_t and the variables p^t and f^t, we can use the multi-cut technique from Section 3.3.2. By aggregating multiple timesteps into a single block, we can strike a balance between the size of the master problem and that of the subproblems that is suitable for a given instance of the problem.

Before we proceed, we should note that the constraint matrix can in fact also be understood as a block-diagonal matrix with respect to the different network regions. This enables an alternative decomposition approach where all transmission-related variables (capacities and power flows) are dealt with in the master problem, whereas all decisions within one network region (e. g. production and production capacities) are delegated to the subproblem. This decomposition exploits a different dimension of the underlying problem structure. Depending on the specific application, either approach may be computationally more useful and both can be combined to obtain a finer decomposition still.

As argued in Section 3.3.2, however, a finer decomposition is not necessarily computationally more advantageous and the relative sizes of master- and subproblem have to be carefully balanced. On the other hand, if network regions can be identified that are very weakly connected between each other, the limited interaction of different subproblems in this setting can greatly improve the performance of a decomposition algorithm. Finally, other decompositions are conceivable, e. g. by energy sector where a separate demand has to be satisfied within each sector.

In this section, we focus on the timestep-based approach outlined above, mainly because it is the one that allows the cleanest notation and because it seems computationally the most promising in our particular application. The network-based approach and other decomposition paradigms can be developed along the same lines.

The epigraph $\operatorname{epi}(z)$ of the function z as defined in (3.55) is not given explicitly but, as previously, we can easily derive a separation algorithm for $\operatorname{epi}(z)$ by applying Theorem 3.6 to the capacity expansion problem. Using Corollary 3.22, the subproblem can be written in the form of (3.39) to (3.41) as follows:

$$
\begin{aligned}
&\min\ \lambda \\
&\text{s.t. } p^t \le \bar{p}^* + \lambda\omega_p && \forall\ t \in [T] && (\gamma^{p,t}) \\
&\quad -(\bar{f}^* + \lambda\omega_f) \le f^t \le \bar{f}^* + \lambda\omega_f && \forall\ t \in [T] && (\gamma^{f,t}) \\
&\quad \sum_{t\in[T]} c^\top p^t \le \eta^* + \omega_0\lambda && && (\gamma_0) \\
&\quad \lambda \ge 0 \\
&\quad (p^t, f^t) \in P_t && \forall\ t \in [T]
\end{aligned}
\tag{3.56}
$$

The γ-variables denote the vectors of dual variables corresponding to the respective constraints. Note that in the case of $\gamma^{f,t}$, the sign of the respective dual variable depends on which side of the inequality is tight. It will turn out, however, that the sign is irrelevant for the computation of the corresponding cut, which is why we refrain from specifying the exact rule.

Now, in order to compute the optimal right-hand side for a cut according to Theorem 3.19, we need the dual variables of all the constraints, including those defining the

set P. It is useful at this point to reformulate our results in the spirit of a well-known alternative formulation of optimality cuts in the context of the original Benders decomposition algorithm (see, e. g., [GLM99; AC00] or [Con+06, Ch. 3.3]). The following result can be derived immediately from our Theorem 3.19:

Corollary 3.46
Let $(\tilde{\omega}, \tilde{\omega}_0)^\top := (H\omega, -\omega_0)^\top$ *and let* (γ, γ_0) *be an optimal solution for an optimization problem of the form* (3.27) *with objective value* D*. Then*

$$D = \gamma^\top (Hx^* - b) - \gamma_0 \eta^* = \gamma^\top Hx^* - \gamma_0 \eta^* - \gamma^\top b.$$

If $\tilde{\omega}^\top \gamma + \tilde{\omega}_0 \gamma_0 < 0$*, then by Theorem 3.19 the inequality*

$$\gamma^\top Hx - \gamma_0 \eta \leq \gamma^\top Hx^* - \gamma_0 \eta^* - D$$

supports epi(z).

The advantage of the above formulation is that all components of γ that correspond to null rows of H are immediately eliminated. This includes, in particular, all the inequalities that are part of the description of the set P. We can thus formulate our algorithm completely independently of this description: Given an optimal solution to (3.56) with objective value λ^* and where $(\gamma^{p,t}, \gamma^{f,t}, \gamma_0)$ denotes the corresponding vector of dual variables as noted above, we obtain that the cut

$$\left(\sum_{t\in[T]} \gamma^{p,t}\right)^\top \bar{p} + \left(\sum_{t\in[T]} |\gamma^{f,t}|\right)^\top \bar{f} + \gamma_0 \eta \geq \left(\sum_{t\in[T]} \gamma^{p,t}\right)^\top \bar{p}^* + \left(\sum_{t\in[T]} |\gamma^{f,t}|\right)^\top \bar{f}^* + \gamma_0 \eta^* + \lambda^*$$

supports the set epi(z). Note again that the sign of $\gamma^{f,t}$ turns out to be irrelevant. The resulting Benders decomposition algorithm for the Capacity Expansion Planning Problem is shown in Algorithm 3.

3.4.1 Upper Bounds

While the algorithm is in principle guaranteed to terminate after a finite number of iterations, convergence tends to be very slow towards the end. It therefore makes sense to abort the iteration loop as soon as a feasible solution has been found with an objective value that is sufficiently close to the optimum.

In the original Benders decomposition algorithm which uses the cut generation procedure from Lemma 3.4, this is very easy: Whenever (x, η) is a master solution

Input: Instance $(\bar{p}^{\max}, c^p, \bar{f}^{\max}, c^f, c, T, P)$ of the capacity expansion problem
Output: optimal solution $(\bar{p}^*, \bar{f}^*)$

1: initialize the set $\operatorname{epi}_S(z)^{(0)} := \left\{(\bar{p}, \bar{f}, \eta) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^{|\mathcal{L}|} \times \mathbb{R}_{\geq 0} \;\middle|\; \bar{p} \leq \bar{p}^{\max}, \bar{f} \leq \bar{f}^{\max}\right\}$
2: **for** $i = 1..$`maxit` **do**
3: choose a subproblem objective $(\omega_p^{(i)}, \omega_f^{(i)}, \omega_0^{(i)})$
4: solve the problem
$$\min\left\{(c^p)^\top \bar{p} + (c^f)^\top \bar{f} + \eta \;\middle|\; (\bar{p}, \bar{f}, \eta) \in \operatorname{epi}_S(z)^{(i)}\right\}$$
to obtain $\left(\bar{p}^{(i)}, \bar{f}^{(i)}, \eta^{(i)}\right)$
5: solve the problem (3.56) with input $\left(\bar{p}^{(i)}, \bar{f}^{(i)}, \eta^{(i)}\right)$ and $\left(\omega_p^{(i)}, \omega_f^{(i)}, \omega_0^{(i)}\right)$
6: **if** $\lambda = 0$ **then return** $\left(\bar{p}^{(i)}, \bar{f}^{(i)}\right)$
7: **else**
8: set
$$\operatorname{epi}_S(z)^{(i+1)} := \left\{(\bar{p}, \bar{f}, \eta) \in \operatorname{epi}_S(z)^{(i)} \;\middle|\; \left(\sum_{t\in[T]} \gamma^{p,t}\right)^\top (\bar{p} - \bar{p}^*) + \left(\sum_{t\in[T]} |\gamma^{f,t}|\right)^\top (\bar{f} - \bar{f}^*) + \gamma_0(\eta - \eta^*) \geq \lambda^*\right\}$$
9: **end if**
10: **end for**
11: **return** `maxit` reached, no optimal solution found

Algorithm 3: The Benders decomposition algorithm for the simplified Capacity Expansion Planning Problem.

for which we reach case d) (i. e., the LP (3.4) is feasible), denote by y an optimal solution to (3.4). Then (x, y) is a feasible solution to the original problem (3.1) and the corresponding objective value $c^\top x + d^\top y$ is an upper bound for the optimal objective value. On the other hand, $c^\top x + \eta$ is a lower bound on the optimal objective value for any master solution (x, η) (even if (3.4) is infeasible). This means that we can update the lower bound at every iteration and we occasionally obtain feasible solutions for the original problem (every time that we reach case d)) such that we can abort the algorithm if their value is close enough to the best available lower bound.

Also when we generate cuts from the alternative polyhedron, we obtain a lower bound of the optimal objective value at each iteration. Furthermore, given a solution $x \in S$ for which the un-relaxed subproblem (3.4) is feasible, we can generate a feasible solution (x, y) for the original problem in the same way described above, providing us with an upper bound for the objective value. However, while the cuts generated from the alternative polyhedron are typically *stronger* and hence increase the lower bound more quickly than the original Benders feasibility/optimality cuts, they do not enforce feasibility of the subproblem (3.4) in the same way as Benders feasibility cuts do.

This means that, unless the subproblem (3.4) is feasible for all $x \in S$, Algorithm 2 typically does not produce a solution x for which (3.4) is feasible until very late in the process. When we thus first obtain an upper bound from such a solution against which to measure the value of the current lower bound, it may turn out that the the lower bound had already been close enough to the optimal solution to abort the algorithm for a number of iterations.

This makes it necessary to spend some thought on the generation of upper bounds from master solutions for which (3.4) is *not* feasible. Fortunately, this is very easy for the simplified Capacity Expansion Problem due to the simple structure of our master problem. If we add the constraints

$$p^t \leq \bar{p}^{\max} \qquad \forall\, t \in [T] \tag{3.57}$$

$$-\bar{f}^{\max} \leq f^t \leq \bar{f}^{\max} \qquad \forall\, t \in [T] \tag{3.58}$$

to the optimization problem (3.56), then the following proposition holds:

Proposition 3.47
Let $((p^t)_{t\in[T]}, (f^t)_{t\in[T]}, \lambda)$ be feasible for (3.56) *with constraints* (3.57) *and* (3.58) *added as above. Then $(\bar{p}, \bar{f}, \eta)$ with $\bar{p} := \max_{t\in[T]} p^t$, $\bar{f} := \max_{t\in[T]} |f^t|$ and $\eta := \sum_{t\in[T]} c^\top p^t$ is a feasible master solution. The value $\bar{z}$ of this solution represents an upper bound for the value of any optimal solution. More specifically, it holds for the value z^* of the optimal solution that*

$$0 \leq \bar{z} - z^* \leq \lambda \cdot \left(\omega_0 + \omega_p^\top c^p + \omega_f^\top c^f\right)$$

Proof. Because of constraints (3.57) and (3.58), it holds that $(\bar{p}, \bar{f}) \in S$. Since the point $((p^t)_{t\in[T]}, (f^t)_{t\in[T]}, \lambda)$ is feasible for (3.56), it holds that $(p^t, f^t) \in P_t$ for all $t \in [T]$. Furthermore, $p^t \leq \bar{p}$ and $f^t \leq \bar{f}$ for all $t \in [T]$ and hence $(\bar{p}, \bar{f}, p, f)$ is feasible for (3.48) to (3.53). Let $(\bar{p}^*, \bar{f}^*, \eta^*)$ denote the master solution used in the problem (3.56). As any master solution over the course of the algorithm represents a lower bound on the optimal objective value, it holds that

$$\begin{aligned} z^* &\geq \sum_{i\in\mathcal{I}} c_i^p \bar{p}_i^* + \sum_{l\in\mathcal{L}} c_l^f \bar{f}_l^* + \eta^* = (\bar{p} - \lambda\omega_p)^\top c^p + \left(\bar{f} - \lambda\omega_f\right)^\top c^f + (\eta - \lambda\omega_0) \\ &= \bar{z} - \lambda \cdot \left(\omega_0 + \omega_p^\top c^p + \omega_f^\top c^f\right). \end{aligned}$$

□

In particular, the difference between upper bound and optimal solution value goes to 0 as $\lambda \to 0$ for fixed $(\omega_p, \omega_f, \omega_0)$. This means that any threshold for the difference between upper and lower bound will eventually be reached: As soon as the tentative master solution (x^*, η^*) is feasible, we can no longer find an (x^*, η^*)-separating cut and the problem (3.56) will have an optimal objective value of 0.

As a matter of fact, the constraints (3.57) and (3.58) are redundant for the complete optimization problem (3.48) to (3.53), which means that adding them to (3.56) does not change the validity of the resulting cuts (in fact, adding (3.57) and (3.58) amounts to including the description of S into the subproblem definition, as in Section 3.3.1). As the updated interaction matrix has null rows for the constraints (3.57) and (3.58), the cut derived based on Corollary 3.46 remains valid (although the optimal solution of the subproblem might change).

In certain cases we can use Lemma 3.42 to obtain a tighter bound than the one defined above using the following proposition, even without explicitly enforcing constraints (3.57) and (3.58):

Proposition 3.48
Let $(\bar{p}^*, \bar{f}^*, \eta^*) \notin \operatorname{epi}(z)$ be a tentative master solution, $(\omega_p, \omega_f, \omega_0) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^{|\mathcal{L}|} \times \mathbb{R}$ and let $((p^t)_{t\in[T]}, (f^t)_{t\in[T]}, \lambda)$ be an optimal solution for (3.56) with $\lambda > 0$. If $(\bar{p}^*, \bar{f}^*) + \lambda \cdot (\omega_p, \omega_f) \in S$, then

$$(\bar{p}^*, \bar{f}^*, \eta^*) + \lambda \cdot (\omega_p, \omega_f, \omega_0)$$

is a feasible master solution and its value hence constitutes an upper bound on the optimal objective value.

Proof. For any problem of the form (3.1) and tentative master solution (x^*, η^*), write $Q := (\operatorname{epi}_S(z) - (x^*, \eta^*))^-$. Let $(\omega, \omega_0) \in \mathbb{R}^n \times \mathbb{R}$ and let (λ, x, y) be an optimal solution for (3.36) to (3.38) with $\lambda > 0$. It holds that

$$\begin{aligned} h_Q(\omega, \omega_0) &= \max\Big\{\omega^\top \pi + \omega_0 \pi_0 \;\Big|\; (\pi, \pi_0) \in (\operatorname{epi}(z) - (x^*, \eta^*))^-\Big\} \\ &= \max\Big\{(H\omega)^\top \gamma - \omega_0 \gamma_0 \;\Big|\; (\gamma, \gamma_0) \in P^{\leq}(x^*, \eta^*)\Big\} \\ &= -\frac{1}{\lambda} \end{aligned}$$

where the first equality is by definition, the second follows from Theorem 3.17 and the third is a consequence of Corollary 3.22. As $\lambda > 0$, Theorem 3.15 implies that $(\omega, \omega_0) \in \operatorname{pos}(\operatorname{epi}(z) - (x^*, \eta^*))$. By Lemma 3.42, we now obtain in particular that $\lambda \cdot (\omega, \omega_0) + (x^*, \eta^*) \in \operatorname{epi}(z)$.

This proves that $(\bar{p}^*, \bar{f}^*, \eta^*) + \lambda \cdot (\omega_p, \omega_f, \omega_0) \in \operatorname{epi}(z)$ and since $(\bar{p}^*, \bar{f}^*) + \lambda \cdot (\omega_p, \omega_f) \in S$ it holds indeed that $(\bar{p}^*, \bar{f}^*, \eta^*) + \lambda \cdot (\omega_p, \omega_f, \omega_0) \in \operatorname{epi}_S(z)$. □

The antecedent is easy to check and furthermore, it is automatically satisfied if, for instance, we choose (ω, ω_0) according to Theorem 3.43 in order to obtain an Pareto-optimal cut.

3.4.2 Selection of Subproblem Objective

In each iteration of Algorithm 3, we may choose a new subproblem objective $(\omega_p, \omega_f, \omega_0)$. As noted by Fischetti, Salvagnin, and Zanette [FSZ10], the approach originally used by Benders [Ben62] to generate optimality cuts corresponds to choosing $\omega_p = \omega_f = 0$, $\omega_0 = 1$ (the original selection criterion used for feasibility cuts is unspecified and depends on the implementation of the solution algorithm). Fischetti, Salvagnin, and Zanette [FSZ10] also suggest a better selection criterion for general applications of Benders decomposition, which in the case of (3.56) corresponds to setting $\omega_p = \mathbb{1}, \omega_f = \mathbb{1}, \omega_0 = 1$. The rationale given by the authors is that this objective function gives preference to solutions where only a small number of constraints are active. Finally, the authors also note that $\tilde{\omega}_0$ can be used as a "scaling factor taking into account a wider range of variable η", but they do not go into details as to how it would be determined (beyond manually adapting it e. g. to differences in order of magnitude between objective function values and the values of decision variables which are known in advance).

A first criterion for the selection of an objective function vector is boundedness of the resulting subproblems. While we could enforce boundedness (e. g. by using a sufficiently large bounding box), the depth of the cut approaches 0 as the objective value goes to infinity, hence a cut from a solution on the boundary of such a box will be very low-depth. Furthermore, it would not support a facet of the set $\operatorname{epi}(z)$, since it does not correspond to a vertex of the original reverse polar set.

Fortunately, we have some very simple tools at hand to choose a suitable objective function. First of all, choosing $\omega_p = \mathbb{1}, \omega_f = \mathbb{1}, \omega_0 = 1$ as suggested by Fischetti, Salvagnin, and Zanette always leads to a bounded subproblem.

Alternatively, Theorem 3.43 allows us to choose any relative interior point of $\operatorname{epi}_S(z)$ to generate an objective function that not only guarantees boundedness, but that additionally always produces a Pareto-optimal cut that is very likely to also define a facet of $\operatorname{epi}_S(z)$. Especially if $\operatorname{epi}_S(z)$ is full-dimensional, then the procedure that we use in Section 3.4.1 to compute an upper bound also yields such a point (or a point that can easily be perturbed to such a point). In practice, we have even observed that the unperturbed point can often be used to the same effect (since it lying in the relative interior of $\operatorname{epi}_S(z)$ is sufficient, but not necessary for the above-mentioned properties).

Finally, we can use any other objective function vector according to our selection criteria from Section 3.2. In most practical cases, we can use prior information about the problem (e. g. monotonicity of the function $z(x)$) to choose an objective vector for which the subproblem is bounded (see also Theorem 3.15). In the rare case where a subproblem is indeed unbounded for a particular objective, we can easily recover by perturbing (or in the worst case replacing) the objective vector using one for which the subproblem is known (or guaranteed) to be bounded. In this case, we obviously have to re-solve the subproblem, which requires additional computational effort and hence should be take care of to not happen too often.

3.4.3 Additional Constraints

We now touch on some possible extensions to the version of the Capacity Expansion Problem that we have considered so far. Specifically, we will investigate the effects of adding further constraints (and variables) to (3.48) to (3.53) in order to incorporate additional aspects of the problem which are desirable in a real-world case study.

Capacity-independent operating constraints The easiest case in this context is that of additional operating constraints, which do not depend on the installed capacities. If these constraints do not span multiple timesteps, (which is the case, e. g., for time-dependent capacity factors for power generation from renewable energy sources such as wind or solar), then they could have been added to the polyhedra P_t right from the beginning. If, however, they include variables from different timesteps, then we can also add these constraints to the description of the polyhedron P without any effect on the correctness of the algorithm described above. The only part where we made explicit use of the fact that P can be written as the cartesian product of polyhedra P_t was when we applied the multi-cut technique from Section 3.3.2. This might no longer be possible and to recover a block-diagonal structure, it might be necessary to introduce additional master variables (see below).

Capacity-dependent constraints If additional constraints include the capacity variables $\bar{p}$ and $\bar{f}$, then the corresponding dual variables have to be taken into account in the cut generation procedure. To simplify the development (and maintenance) of the algorithm, it may make sense to use the simplified formulation from Section 3.3.3 and add a copy of all master problem variables to the subproblem where they can be linked to the value of the respective master variable. However, this would again destroy the block-diagonal structure of the constraints defining the subproblem since the copies of master problem variables would appear in each block.

To circumvent this issue, we can use a separate copy $\bar{p}^t$ of the master variable $\bar{p}$ for each $t \in T$: Let $P'_t := \{(p^t, f^t, \bar{p}^t, \bar{f}^t) \mid (p^t, f^t) \in P_t, p^t \leq \bar{p}^t, -\bar{f}^t \leq f^t \leq \bar{f}^t\}$. Then we can rewrite (3.56) as follows:

$$
\begin{aligned}
&\min\ \lambda \\
&\text{s.t. } \bar{p}^t = \bar{p}^* + \lambda\omega_p && \forall\ t \in [T] && (\gamma^{p,t}) \\
&\quad \bar{f}^t = \bar{f}^* + \lambda\omega_f && \forall\ t \in [T] && (\gamma^{f,t}) \\
&\quad \sum_{t\in[T]} c^\top p^t \leq \eta^* + \omega_0\lambda && && (\gamma_0) \\
&\quad \lambda \geq 0 \\
&\quad (p^t, f^t, \bar{p}^t, \bar{f}^t) \in P'_t && \forall\ t \in [T]
\end{aligned}
\tag{3.59}
$$

We are now back in the previous situation, where all new constraints can be expressed using the subproblem variables $(p^t, f^t, \bar{p}^t, \bar{f}^t)$ alone, i. e., they only include the specific variables for each individual $t \in [T]$ and do not need to explicitly reference the master capacity variables $\bar{p}$ and $\bar{f}$. This means that they may simply be added to the description of the respective polyhedron P'_t. The resulting cut takes the same form as described above (with the only difference that negative values of $\gamma^{f,t}$ now lead to negative entries, i. e., the absolute value is no longer required). If we assume furthermore that Proposition 3.47 continues to hold, then the generation of upper bounds and feasible solutions works exactly as described in Section 3.4.1.

Additional master variables If new master variables are required to express a constraint, they can be added in the same fashion: Suppose that a new vector of master variables $s \in \mathbb{R}^k$ is required. Analogously to the addition of capacity-dependent constraints, we create copies s_t of s for each individual timestep $t \in [T]$ and link them by the constraints

$$s_t = s^* + \lambda\omega_s \qquad \forall\ t \in [T], \tag{3.60}$$

which we add to (3.59). The new constraints which include the master variable s can now be incorporated in a similar fashion as above, defining $P''_t \subset \{(p^t, f^t, \bar{p}^t, \bar{f}^t, s^t) \mid (p^t, f^t, \bar{p}^t, \bar{f}^t) \in P'_t\}$ in a suitable way. Denote the dual variable corresponding to the constraint (3.60) by $\gamma^{s,t}$ for every $t \in [T]$. We can now use Algorithm 3 with the only change that the term corresponding to the constraints (3.60) has to be added to the cut definition:

$$\left(\sum_{t\in[T]} \gamma^{p,t}\right)^\top (\bar{p} - \bar{p}^*) + \left(\sum_{t\in[T]} |\gamma^{f,t}|\right)^\top (\bar{f} - \bar{f}^*) + \left(\sum_{t\in[T]} \gamma^{s,t}\right)^\top (s - s^*) + \gamma_0(\eta - \eta^*) \geq \lambda^*$$

The above situation occurs, for instance, if we include storage units into our model. One way to enforce the consistency of storage levels across time steps is to add an additional vector of variables representing the storage level after each timestep. As argued above, we can set these values in the master problem and link them to separate copies of the respective variable in the subproblems for the preceding and the following timestep using a constraint like (3.60). This restores independence between different timesteps, allowing us to use any aggregation of timesteps as discussed in the context of the multi-cut procedure in Section 3.3.2.

However, in this situation we can no longer rely on the approach from Section 3.4.1 to derive feasible solutions for the master problem and corresponding upper bounds for the optimal objective value. In the best case, we may be able to prove an equivalent theorem which allows us to use any solution to the subproblem of the form (3.59) to derive a feasible solution for the master problem with a sufficiently good objective value to allow us to reach a desired optimality threshold.

If this is not possible, we may selectively in some iterations fix certain values of ω to 0 in order to force the subproblem to satisfy the unrelaxed version of the corresponding constraints. In this way, we might, for instance, be able to enforce $s_t = s^*$ for all $t \in [T]$. If there is sufficient flexibility in the problem description, we obtain a solution that only violates constraints which contain variables that can be handled more easily in a fashion similar to Section 3.4.1, e. g. only the original capacity variables. In the worst case, however, this will lead to the problem (3.59) becoming infeasible, which does not allow us to generate a separating cut.

As a last resort in this case, we can (at least periodically) revert to the original Benders cut selection procedure by setting $(\omega, \omega_0) := (0, 1)$. This will generate feasibility cuts according to case c) from Lemma 3.4 until at some point we reach case d) again, which provides us with an upper bound for the optimal objective value as described in the beginning of Section 3.4.1.

3.5 Empirical Results

While a comprehensive study of the computational performance of different versions of the Benders Decomposition algorithm in different sets of circumstances is beyond the scope of this thesis, we have conducted a small number of tests of the approach described in this chapter in the practical setting from Section 3.4. Rather than making a claim about the performance of a particular version of the Benders decomposition algorithm, we want to demonstrate some possible approaches to optimize the algorithm that arise from our analysis.

Our empirical examples provide a way to demonstrate the effects and tradeoffs associated with these approaches, which could be used to improve the performance of the algorithm. Any such improvements are likely to be problem-specific, although selecting a smart default choice for general purpose optimization algorithms might be a worthy topic for future research. In this sense, the tests in this section should serve to validate our theoretical results in the context of a practical implementation of the decomposition algorithm as well as to provide starting points for a deeper study of the optimal choice for the parameters that affect the performance of the algorithm.

Our tests are based on the realistic instances of the Capacity Expansion Problem for the European power system from [SSH12]. Beyond the basic constraints from Section 3.4, these instances include storage facilities and all of the additional variables and constraints required to keep track of storage levels. Our main test instance is a model of the European electricity grid that consists of 102 demand regions with 587 generation units and 195 (existing and potential) transmission lines. It includes demand data and data for the availability of renewable energy sources in hourly resolution for a period of one year.

We used the approach described in Algorithm 3, which we have implemented in C++

using Gurobi 7.5. For our computations, we used eight CPU cores running at 2.5 GHz with 64 GB of main memory. To solve master problem and subproblems, we use the dual simplex algorithm (i. e., the version of the simplex algorithm that maintains dual feasibility while pivoting between bases). In each iteration, we warm-start all problems using the optimal basis from the previous iteration and solve all subproblems in parallel on the available cores. Beyond this, we run the solution algorithm with default settings, i. e., we did not undertake any computational optimizations with respect to either the algorithm itself or the solution method of master and subproblems.

Regarding the weight vector (ω, ω_0), we focussed on weight vectors that satisfy the condition (3.25) and hence are likely to produce facet-defining cuts. We have compared the following approaches:

"mixed cuts" with $(\omega, \omega_0) = (\mathbb{1}, 1)$. This corresponds most closely to the approach proposed by Fischetti, Salvagnin, and Zanette [FSZ10] (It is equivalent to their approach if we use a simplified formulation according to Section 3.3.3.)

"adaptive cuts" with $(\omega, \omega_0) = (\bar{x} - x^*, \bar{\eta} - \eta^*)$ where $(\bar{x}, \bar{\eta})$ is a (suboptimal) feasible solution computed according to Proposition 3.47. By Cornuéjols and Lemaréchal [CL06, Theorem 3.4], this results in a cut which supports $\operatorname{epi}_S(z)$ in a face which is intersected by a line through (x^*, η^*) and $(\bar{x}, \bar{\eta})$. The idea behind this criterion is that a cut like this is best suited to close the gap which currently exists between lower bound, represented by the infeasible solution (x^*, η^*), and upper bound, represented by the feasible, but suboptimal solution $(\bar{x}, \bar{\eta})$.

Before we move on to present the results of our empirical study, we must briefly discuss some issues that limit the expressiveness of the results. As soon as the problem in consideration is of non-trivial size, there are a number of factors that affect the performance of subproblems and hence the performance of the entire algorithm: The choice of algorithm (simplex vs. interior point), the use of warm-start techniques, optimality tolerance etc. all can have a substantial effect on the time required to solve an individual subproblem. Furthermore, the interplay of these factors in the context of the different cut selection criteria specified above is complex and very difficult to predict. This effect is exacerbated further by the fact that the problems in consideration often expose numerical difficulties due to the wide range of coefficients in the problem description: It may happen that one particular subproblem consumes an enormous amount of effort to solve, while in a run with slightly different parameters where this exact subproblem does not appear, no such problem is encountered.

Against this background, we have decided in this section to use both the *number of iterations* and the *total computation time* as a measure of performance. Given the variety of different factors that affect the performance of the algorithm, both are very incomplete proxies for actual computational performance by themselves. However since the number of iterations is independent at least of any choices regarding the

solution method of subproblems and the computation time gives an indication of actual performance for at least one set of such choices, their combination seemed to us the most useful choice.

Advanced cut selection criteria We begin by comparing the two approaches mentioned above on an instance covering a time horizon of one month. Specifically, we set a static refinement of the cut selection criterion from [FSZ10], enhanced by our transformation of the objective function according to Theorem 3.17 against the *adaptive cuts* approach for the selection of the weight vector (ω, ω_0) as described above. In the latter case, we update the vector (ω, ω_0) in each iteration using a (suboptimal) feasible solution computed according to Proposition 3.47. Note that this precise update mechanism should be seen as an illustrative example rather than a performance-optimized prescription.

Like all plots in this chapter, the figure shows on the (logarithmic) vertical axis the relative optimality gap, i. e., the gap between upper and lower bound relative to the optimal objective value. On the horizontal axis, the first plot shows the number of iterations and the second plot shows the total computation time. This representation takes into account the fact that in practical applications, we are often satisfied with a solution that is guaranteed to be within a certain tolerance of the optimal solution (e. g., 0.1 %), rather than a strictly optimal solution. The plot thus allows us to compare the performance of different approaches with respect to different optimality tolerances.

The results can be inspected (for two different subproblem aggregations) in Figs. 3.7 and 3.8 from the perspective of both number of iterations and computation time: As the first plot in Fig. 3.7 shows, the adaptive selection criterion substantially reduces the number of iterations required to reach a given optimality threshold. For instance, a solution that is guaranteed to be no more than 1 % from optimal can be obtained using the *adaptive cuts* approach in around 1/2 the iterations required under the *mixed cuts* selection criterion. The advantage is less pronounced for smaller optimality gaps but a small advantage remains in all cases.

In terms of computation time, the difference is even clearer: Here, the *adaptive cuts* approach enjoys an advantage of approximately a factor 2.5-3 over the *mixed cuts*. Furthermore, this advantage does not diminish over the course of the algorithm and is present at all accuracy levels.

As further tests show, moving to a different subproblem size (e. g. Fig. 3.8), we obtain qualitatively similar results: While in this case the advantage of the *adaptive cuts* approach is a little less pronounced in terms of the number of iterations, the improvement by a factor 2.5-3 in terms of the computation time remains.

Subproblem aggregation In our results above, it can already be observed that the performance of the algorithm depends not only on the selection method for the weight

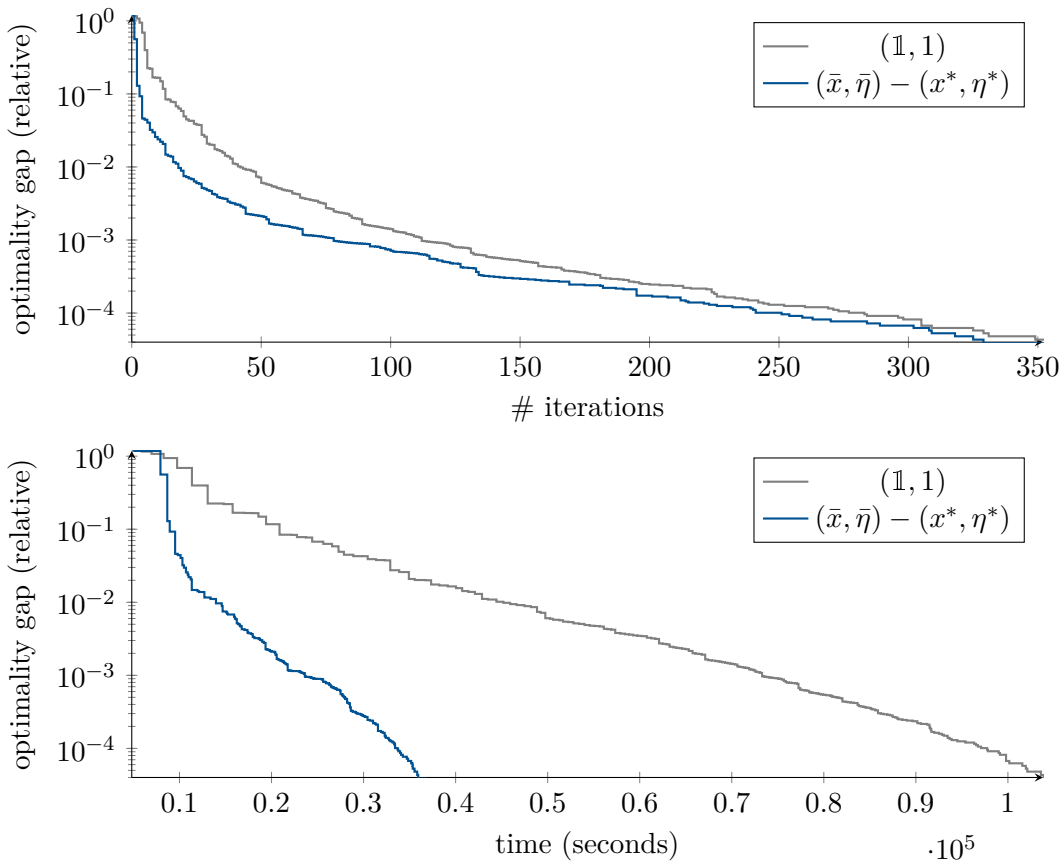


Figure 3.7: This figure shows the optimality gap vs. number of iterations and computation time for the *mixed cuts* and the *adaptive cuts* selection criteria for (ω, ω_0) from Section 3.4.2. Both versions run in a setting with 10 subproblems covering 72 timesteps each, without the bound tightening procedure from Proposition 3.48.

vector, but also on our choices regarding the aggregation of subproblems. Generally, as we have mentioned in the context of the multi-cut procedure from Section 3.3.2, smaller subproblems shift the computational burden from the subproblems to the master problem (which becomes larger due to the higher number of cuts in each iteration).

This affects the total computation time required to reach a given level of accuracy primarily in two ways: Smaller subproblems achieve a better approximation of $\operatorname{epi}(z)$ early in the algorithm thanks to the finer granularity of the generated cuts. At the same time, the computation time for early iterations decreases, because the (smaller) subproblems are much easier to solve.

On the other hand, as the algorithm progresses, this effect diminishes since the increasingly smaller changes to the subproblems between iterations together with the

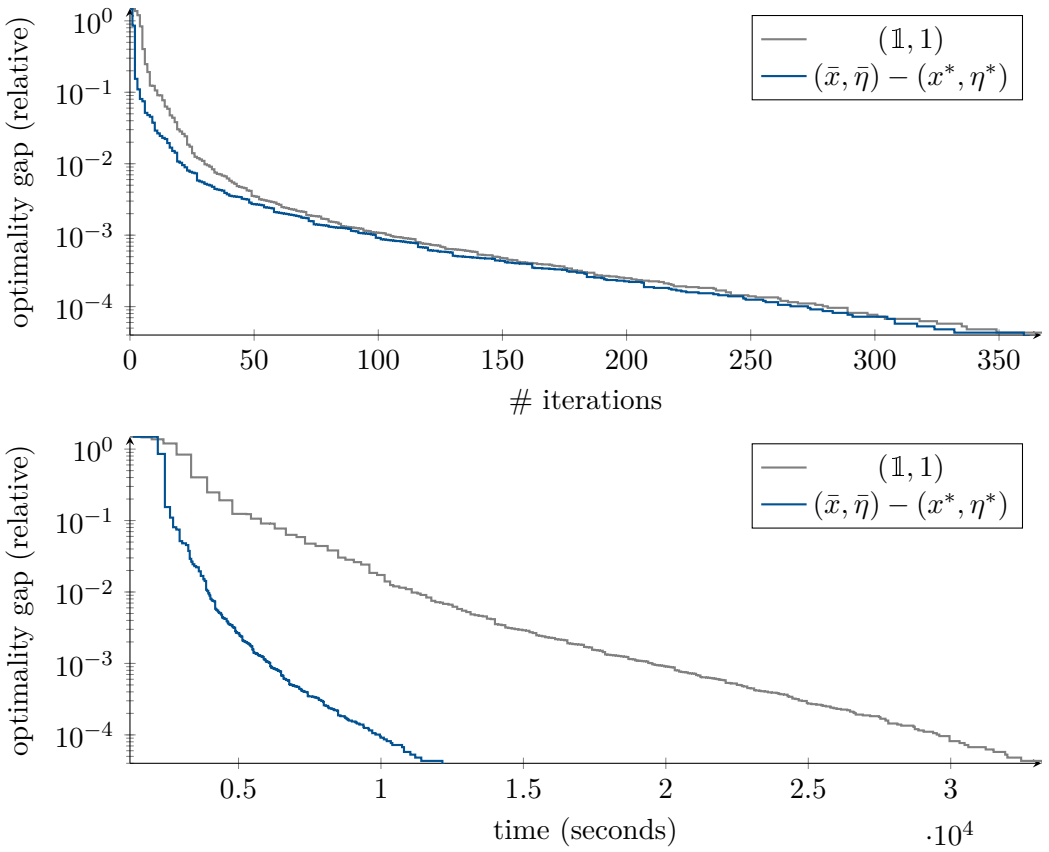


Figure 3.8: This figure shows the optimality gap vs. number of iterations and computation time for the *mixed cuts* and the *adaptive cuts* selection criteria for (ω, ω_0) from Section 3.4.2 using the basic generation of upper bounds according to Proposition 3.47. Both versions run in a setting with 30 subproblems covering 24 timesteps each, without the bound tightening procedure from Proposition 3.48.

use of simplex warm-starts means that larger subproblems can be solved relatively quickly, as well. At the same time, as cuts accumulate in the master problem, it becomes harder and harder to solve. This effect is more prominent the smaller the subproblems and hence the larger the number of cuts added in reach iteration.

The results can be observed in Fig. 3.9. If a small optimality tolerance is required, a subproblem size of 24h seems to strike the best tradeoff in the context described above. Using smaller subproblems (12h), we can reach a reasonably good solution much faster, but the algorithm then markedly slows down due to the increased size of the master problem. Using a subproblem size of 72h, on the other hand, early iterations require much more time to solve the larger subproblems, a disadvantage from which

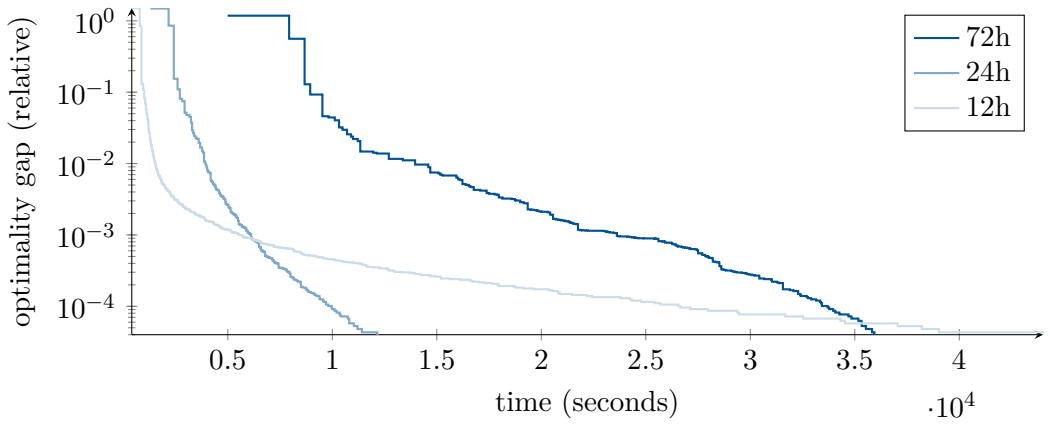


Figure 3.9: This figure shows the optimality gap vs. computation time for different subproblem sizes using the *adaptive cuts* selection criteria for (ω, ω_0) from Section 3.4.2 and the basic generation of upper bounds according to Proposition 3.47.

the algorithm can never recover (although for very small optimality tolerances, even this version outperforms the version using subproblems of 12h).

Larger instances and non-decomposed approaches Many of the above tests were also performed on larger instances covering an entire year. While the number of iterations as well as the computation time required to reach a given accuracy is obviously much larger than in the examples presented above, the general line of results (particular with respect to the relative performance of *mixed cuts* and *adaptive cuts*) is identical. Compared to conventional optimization methods, the advantage of decomposition approaches generally grows with the size of the problem under consideration. This advantage can broadly be observed in two areas: computation time and memory consumption.

Regarding computation time, the advantage of decomposition approaches is less clear. The limited exchange of information between master- and subproblems and the number of iterations thus required to solve the problem to a satisfying accuracy does introduce a notable computational overhead. On simpler problems, this overhead more than outweighs the advantage that results from the smaller size of the problems that have to be solved in any single iteration. Most practical applications of Benders decomposition thus focus e. g. on Mixed-Integer problems, that quickly become too large to be solved by conventional methods in any reasonable amount of time. Reducing the size of problems that have to be solved in a given iteration in this case dramatically reduces the computation time required, easily outweighing the overhead introduced by the decomposition framework.

The (continuous) problems that we used to evaluate our approaches can generally be

solved relatively quickly, even by conventional optimization methods. As a consequence, the advantages of the decomposition approach are less prominent and on instances covering a single month, conventional methods substantially outperform all variants of Benders decomposition that we investigated. However, on the larger instances covering an entire year, the decomposition approach quickly catches up and we expect this trend to continue on even larger problems.

On the other hand, memory consumption quickly becomes the most prominent issue in the case of larger problem sizes. Once the available memory is no longer sufficient to hold the complete representation of the problem during the run of the solution algorithm, it becomes infeasible to solve the problem within any reasonable amount of time. Interior point methods, which often provide far better performance, especially on large-scale LP problems, generally require much more memory than, e. g., the Simplex method, which can be used for the iterative updates of subproblem solutions in the context of a decomposition approach. This difference is negligible for smaller problems, but becomes more and more relevant the larger the instance under consideration. As an indication, for an instance covering an entire year, a decomposition approach with 10 days per subproblem required around 24 GB of memory, whereas an interior point algorithm on the complete problem requires at least 42 GB.

3.6 Conclusion

We conclude this chapter by a brief summery of results as well as an outlook on interesting research questions in the context of cut selection for Benders decomposition.

We have presented an improved approach for cut generation in the context of Benders decomposition. The approach is based on the relation between the alternative polyhedron, commonly used as a characterization of possible cuts in the context of Benders decomposition, and the reverse polar set, originally introduced by Balas and Ivanescu [BI64] in the context of transportation problems.

While the close similarity of the two sets is well-known, we have formally derived the exact relation between the two and have discussed its implications: In particular, the alternative polyhedron can be viewed as an extended formulation of the reverse polar set. A description of the former is much more readily available in the context of Benders decomposition, which makes it more useful as a basis for a cut generation routine. However, while all vertices of the alternative polyhedron possess some useful properties (their support corresponds to minimal infeasible subsystems of the Benders subproblem [GR90; FSZ10]), those that correspond to vertices of the reverse polar set have additional advantages: They generate facet cuts, which in particular are always supporting.

Based on this insight, we have developed a modified version of the cut generation procedure by Fischetti, Salvagnin, and Zanette [FSZ10] that produces facet cuts

for all but a sub-dimensional set of subproblem objectives without any additional computational effort. In addition, the new criterion is more robust with respect to the formulation of the problem. In particular it always generates supporting Benders cuts, which is not true for the original procedure proposed in [FSZ10].

Furthermore, our method can be parametrized by the selection of an objective vector in primal space. This can be used to leverage prior knowledge about the problem, as well as primal information obtained e. g. by a heuristic algorithm, in the context of a Benders decomposition approach. Put into context of other well-know selection criteria, most notably Pareto-optimal or facet-defining cuts, each of these criteria can be matched to a particular subset of objective functions used in the context of our cut generation framework.

Finally, we have applied our results in the context of a small case study on the Transmission Capacity Expansion Problem (TCEP). We have discussed some of the practical problems encountered in this context, as well as problem-specific topics such as the generation of upper bounds and the selection of parameters. Furthermore, we have touched on the compatibility of the algorithm with the addition of new variables and constraints to the basic TCEP.

While the original motivation for our work was very much the practical application of Benders Decomposition in the context of Energy System Optimization problems, most of our results in this chapter turned out to be very general. As such, they are applicable to all usages of the Benders Decomposition algorithm, regardless of the particular optimization problem under consideration. In future work, it would therefore be very interesting to explore the practical impact of our results on a broader set of problem instances, starting, e. g., with the mixed-integer benchmark problems used in [FSZ10].

The results of our case study may be seen as a first indication in this direction. We verified in particular the effects of different choices for the objective function used to parametrize the cut selection criterion. As an illustrative example, we show that a cut selection criterion which takes into account the current upper bound can substantially improve the performance of the algorithm.

Finally, in the context of a deeper empirical evaluation, some further choices with respect to parametrization of the algorithm as well as the implementation of certain subroutines would be interesting to analyze more deeply:

- To what extent can a-priori knowledge about the problem (or information obtained through a fast preprocessing algorithm) be leveraged to inform the selection of a (static) choice for the subproblem objective (ω, ω_0)?

- In a generic implementation of Benders Decomposition, upper bounds are computed solely to decide when the algorithm has converged and, if required, provide feasible solutions. The dynamic subproblem objective (the second approach in

our case study) provides a way to leverage upper bound information to improve the choice of parameters over the course the algorithm. In this context,

- how can our very simple update mechanism be improved, e. g. by stabilization or convex combination with some other choices for (ω, ω_0)?
- what other methods can be used to generate upper bounds and which selection of the subproblem objective in the approach from Section 3.4.1 is ideal for the computation of upper bounds in this context?

Chapter 4

Interdependent Scheduling Games

Contents

In the previous chapters, we have analyzed the problem of optimally extending a network infrastructure in order to satisfy a given demand by a set of available suppliers. The underlying assumption in all cases was that all stakeholders would cooperate maximally to implement the expansion schedule that is optimal from the point of view of the economy as a whole. This approach, with agents ignoring individual gain or cost, is of course extremely optimistic and in most cases unrealistic.

Indeed, if we assume that individual economic agents have control over the expansion of certain parts of the network, a potential for rent-seeking emerges in many places, which can lead to individually optimal outcomes that are not globally optimal (or, in economic terms, *welfare-maximizing*). This potential is all the more prominent if the economic interests of the agents are not clearly aligned with minimizing congestion in the part of the network that they control. Instead, agents may have a stake in the generation and distribution of electric power, as well. If, for instance, the owner of the transmission infrastructure is also (economically intertwined with) a generator or distributor of electrical power, then he might have an incentive to keep congestion in the network high in order to collect the resulting rents.

The most harmful realization of the sub-optimalities mentioned above is the case where a TSO decides to keep congestion in his part of the network high in order to be able to sell electricity at a higher price within his network region due to the resulting scarcity (see, e. g., [SO09]). But even if this most radical case, where beneficial infrastructure is simply not built, can be avoided through regulation (for instance by unbundling vertically integrated companies in the energy sector, see, e. g., [EU09]), TSOs may have different priorities which leads to them allocating their resources to construction projects in a way that differs from the globally optimal allocation. For instance, a TSO may decide to first build a transmission line that benefits him directly and delay the construction of a line that is of less immediate benefit to him, even though the latter line might eliminate a crucial bottleneck in the network that also affects neighboring regions and hence is more important to global welfare.

This is the scenario that we will study in the following chapter. The results in this chapter are joint work with Andres Abeliuk, Haris Aziz, Gerardo Berbeglia, Serge Gaspers, Petr Kalina, Nicholas Mattei, Dominik Peters, Kevin Schewior, Pascal Van Hentenryck and Toby Walsh. Some results have previously appeared in the proceedings of the ICJAI 2016 conference as [Abe+16], a corresponding reference is provided where this is the case. Also, the proofs for these results are generally adopted from [Abe+16] with slight modifications, unless indicated otherwise.

4.1 Introduction

The setting that we consider in this chapter was first described in [ABV15] in terms of infrastructure recovery after natural disasters and extreme weather events. The underlying idea is that different agents each control a set of *services* that they sell to the community for a fixed reward per unit of time. These services are connected by dependencies, meaning that some services need other services to be available in order to function. Note that these other services may be controlled by the same agent or by some other agent. Also, the dependency structure between services is transitive: If service a is required for service b to be available, then naturally any service c that depends on b also needs a to be available in order to function.

After a disaster, all of these services break down and have to be restored. Each agent can decide in which order he wishes to restore (or *deploy*) the services controlled by him, but a service will only be *active* (and hence generate a reward) once not only the service itself but also all other services that it depends on are deployed. This leads to a situation where the reward that any agent receives depends on both his own decisions and the decisions of other agents, who control services that his services depend on.

The problem that we consider in this chapter can be understood as a generalization of the problem described by Abeliuk, Berbeglia, and Van Hentenryck [ABV15], which is restricted two two agents of which only one may affect the other agent's payoff.

In contrast, we allow any number of agents with arbitrary interdependencies between each other. Our setting can readily be applied to other situations in the context of power systems, such as the one outlined above: Consider a city where the distribution of electricity in different areas is licensed to different companies. Each community connected to the grid generates a certain reward for the company holding the license for the respective area. However some areas may not be immediately connected to the transmission grid, instead they might need power to transfer through a neighboring region in order to reach their households. The activity of the *service* of supplying a community with electricity may thus depend on the availability of a different service, i. e., the availability of electricity in a neighboring region. Furthermore, this region could possibly be controlled by a different agent, making the reward reaped by the first agent contingent on decisions of the second agent.

4.1.1 Game-theoretic Concepts

Definition 4.1
Let $k \in \mathbb{N}$ be the number of *players*. For every player $i \in [k]$, let A_i be a finite set of *actions* and denote the *action space* by $A := \bigtimes_{i\in[k]} A_i$. We call an element $a \in A$ an *action profile*. For every player $i \in [k]$, let $u_i : A \to \mathbb{R}_{\geq 0}$ be a *utility function*. The tuple $(k, (A_1, \ldots, A_k), u)$ denotes a *normal form game*.

For a player $i \in [k]$, we write $a_{-i} \in \bigtimes_{j\in[k]\setminus\{i\}} A_j$ for an action profile's sub-vector consisting of the actions of all other players. Analogously, we allow the (formally incorrect, but very common) notation $u_i(a_i, a_{-i})$ for the utility function of player i evaluated in an action profile a' with $a'_i = a_i$ and $a'_{-i} = a_{-i}$.

Definition 4.2 (Nash Equilibrium)
Let $(k, (A_1, \ldots, A_k), u)$ be a normal form game. An action profile $a \in A$ is called *pure Nash equilibrium (PNE)* if for all players $i \in [k]$,

$$u_i(a) \geq u_i(a') \quad \text{for all } a' \in A \text{ with } a_{-i} = a'_{-i}.$$

Definition 4.3 (Price of Anarchy / Stability)
Let $G = (k, (A_1, \ldots, A_k), u)$ be a normal form game and denote by $N \subset A$ the set of action profiles that are pure Nash equilibria.

a) The *price of anarchy* of G is given by

$$\mathrm{PoA}(G) = \frac{\max_{a\in A} \sum_{i\in[k]} u_i(a)}{\min_{\bar{a}\in N} \sum_{i\in[k]} u_i(\bar{a})}.$$

b) The *price of stability* of G is given by

$$\mathrm{PoS}(G) = \frac{\max_{a\in A} \sum_{i\in[k]} u_i(a)}{\max_{\bar{a}\in N} \sum_{i\in[k]} u_i(\bar{a})}.$$

It obviously holds for every normal form game G that $\mathrm{PoS}(G) \leq \mathrm{PoA}(G)$. Furthermore, the above definitions immediately imply the following bounds for the sum of utilities in any Nash equilibrium.

Remark 4.4
Let $G = (k, (A_1, \ldots, A_k), u)$ be a normal form game and $u^* := \max_{a \in A} \sum_{i \in [k]} u_i(a)$ the maximal total utility achievable by any action profile. Let $\bar{a}$ be a Nash equilibrium. Then,

$$\frac{u^*}{\mathrm{PoA}(G)} \leq \sum_{i \in [k]} u_i(\bar{a}) \leq \frac{u^*}{\mathrm{PoS}(G)}.$$

4.1.2 Interdependent Scheduling Games

We will now define the mathematical model used to express the problems considered in this section. An interdependent scheduling game is defined as follows:

Definition 4.5
Let $k \in \mathbb{N}$ and $G := (T, E)$ be a directed, transitive and acyclic graph (see Appendix A.2) with a vertex weight function $r : T \to \mathbb{N}_0$ and a vertex labelling function $\rho : T \to [k]$. We call the tuple (k, G, r, ρ) an *interdependent scheduling game (ISG)*. If $r(v) = 1$ for all $v \in T$, we say that the game has *unit rewards*.

The set of vertices T of the graph G denotes the set of services present in the game. The set of edges E captures the dependency structure between these services. As mentioned above, we assume that these dependencies are transitive but acyclic. The weight function r assigns to each service the reward per unit of time that the service generates when it is active. The labelling function ρ assigns each service to the player $i \in [k]$ that controls it: She may decide, in which order she wants to deploy the services under her control, this order is given by the following definition of a schedule.

Definition 4.6
Let $(k, (T, E), r, \rho)$ be an interdependent scheduling game. For every $i \in [k]$, we define the set of services that player i controls by $T_i := \{v \in T \mid \rho(v) = i\}$. A *schedule* is a tuple $\pi := (\pi_1, \ldots, \pi_k)$ where each $\pi_i : T_i \to \{1, \ldots, |T_i|\}$ is a permutation of the set T_i. For every $v \in T$, we write $\pi(v) := \pi_{\rho(v)}(v)$.

We assume for reasons of simplicity that every service requires the same amount of time to deploy, this allows us to use a suitable discretization of time to say that if $\pi(v) = t$ for some $v \in T$, then v is deployed at time t. Similarly, we will assume that $|T_1| = \cdots = |T_k| =: q$ in order to be able to consider a universal time horizon q. If $|T_i| < q$ then we can extend (T, E) by a suitable number of isolated vertices assigned to player i that have no effect on the utility of any player.

Note that the assumptions of acyclicity and unit time are not entirely without loss of generality. We will discuss their effect and possible generalizations in Section 4.6.

The graph G captures the dependency structure among the set T of services: For every $v, w \in T$, we say that w *depends on* v if and only if $(v, w) \in E$ and we denote the subgraph of G induced by the vertices in T_i by $G_i = (T_i, E_i)$. As mentioned above, a service becomes active (and thus begins to generate a reward) as soon as the service itself as well as all services that it depends on are deployed. The utility for every player thus depends on the subset of services that are active at each timestep in the following way:

Definition 4.7
Let $(k, (T, E), r, \rho)$ be an interdependent scheduling game and let π be a schedule. The *activation time* of a service $v \in T$ is given by $a_\pi(v) := \max\{\pi(w) \mid w \in \{v\} \cup N^{\text{in}}(v)\}$.

The *reward* $R_i(\pi)$ of player i under schedule π is given by

$$R_i(\pi) := \sum_{t=1}^{q} \sum_{\substack{v \in T_i \\ a_\pi(v) \leq t}} r(v) = \sum_{v \in T_i} (q + 1 - a_\pi(v)) r(v).$$

The *total reward* is given by $R(\pi) := \sum_{i=1}^{k} R_i(\pi)$.

Note that transitivity of the graph (T, E) allows us to restrict the definition of the activation time for a service v to the immediate neighborhood of v, since every service that is required for activation of a service on which v depends is already contained in $N^{\text{in}}(v)$ by transitivity.

Finally, observe that an ISG $(k, (T, E), r, \rho)$ is indeed a normal form game according to Definition 4.1: With $[k]$ as the set of players and the set of permutations of T_i as the set of actions of player i, every schedule π is an action profile and the function R_i is the utility function for player i.

We can use the following graphical representation for both instances and solutions of Interdependent Scheduling Games (see Fig. 4.1): Given an Interdependent Scheduling Game $(k, (T, E), r, \rho)$, every service $v \in T$ is represented by a node in a grid where the nodes in the i-th row correspond to the services in T_i (those which are controlled by player i). Every service $v \in T$ is labeled with its reward $r(v)$ and its dependencies are indicated by the arcs shown between different services. For ease of presentation, we might omit arcs that are implied by transitivity of the dependency relation, hence the transitive closure of the shown graph will yield the original graph (T, E).

Note that in this representation, a service v is identified only by the player that controls it, the value $r(v)$ and the edges in $N_G(v)$. While there may be services that are indistinguishable by these three characteristics, these will also be interchangeable in any solution to the Interdependent Scheduling Game without affecting the utility of any player and hence whenever two different solutions yield the same graphical

Figure 4.1: The instance from Example 4.8 and two possible solutions. While player 1 receives a higher reward under the schedule π^A (33 vs. 15), the total reward is higher under π^B (336 vs. 516).

representation, we may say that these solutions are *equivalent*. Alternatively, if the rewards are irrelevant (e. g. if we consider a uniform ISG), we might label some services with an identifier such that we can refer to them in the text.

An example of an Interdependent Scheduling Game, two possible solutions and their graphical representations are presented in the following example.

Example 4.8
Consider the following example: Let $k = 2$ and define a graph (T, E) by $T = \{v_1, v_2, v_3, w_1, w_2, w_3\}$ and $E = \{(v_1, w_1), (v_2, v_3), (v_2, w_2), (v_2, w_3)\}$. Furthermore, let $\rho(v_1) = \rho(v_2) = \rho(v_3) = 1$ and $\rho(w_1) = \rho(w_2) = \rho(w_3) = 2$, as well as $r(v_1) = 10$, $r(v_2) = r(v_3) = r(w_1) = 1$, and $r(w_2) = r(w_3) = 100$. A graphical representation of the instance and of two solutions π^A and π^B is shown in Fig. 4.1.

For π^A, we have $R_1(\pi^A) = 3{\cdot}10{+}2{\cdot}1{+}1 = 33$ and $R_2(\pi^A) = 3{\cdot}1{+}2{\cdot}100{+}100 = 303$. For π^B on the other hand, player 1 receives $R_1(\pi^B) = 3 \cdot 1 + 2 \cdot 1 + 10 = 15$ whereas player 2 receives $R_2(\pi^B) = 3 \cdot 100 + 2 \cdot 100 + 1 = 501$. We see that while player 1 receives a higher reward under the schedule π^A, he can sacrifice some of it to enable an alternative solution π^B with higher total reward.

4.2 Related Work

The most immediate precursor to the work presented in this chapter (as mentioned above) is [ABV15], where a limited variant of the setting presented above was considered. The (significant) restriction imposed by the authors is that only two players are considered and the reward of one of these players is independent of decisions made by the other player, i. e., there is only one player whose services depend on services controlled by the other player.

The general setting of interdependent scheduling games bears a strong resemblance with the area of machine scheduling. Indeed, a problem instance as defined above can be seen as a scheduling problem on k identical parallel machines with unit processing times according to the classification of [LRB77] with two additional types of constraints: machine eligibility restrictions and *soft precedence constraints*. This connection is worth exploring a little further, since we will be able to apply some results from the

scheduling literature to prove complexity results about our particular setting. In fact, it will turn out that our setting fits right between two classes of well-known scheduling problems, one of which is efficiently solvable whereas the other one is known to be $\mathbb{NP}$-hard.

Interpreting an ISG in the context of machine scheduling problems, every player is represented by a *machine*. The deployment of each service corresponds to a *job* with processing time 1, independently of the machine used. However, the unique machine that is eligible for any job is determined by the labelling function ρ. The graph (T, E) specifies a precedence relation between the jobs, these constraints however are not binding, rather a penalty (in the form of forgone reward) has to be paid that depends on the violation of the constraint.

From the objective functions typically considered in the scheduling literature, the *weighted sum of completion times* $\sum_{v \in T} r(v)a_\pi(v)$ is most closely related to the total reward as defined in Definition 4.7. More specifically, it holds that

$$C_\Sigma(\pi) := \sum_{v \in T} r(v)a_\pi(v) = \sum_{i=1}^{k} \sum_{v \in T_i} (q+1)r(v) - R_i(\pi) = \sum_{v \in T} (q+1)r(v) - R(\pi). \quad (4.1)$$

Similarly, the unweighted sum of completion times naturally corresponds to the total reward in the same way for the case of unit rewards ($r(v) = 1$ for all $v \in T$). It can be observed that maximization of the total reward and minimization of the weighted sum of completion times differ only by the constant $\sum_{v \in T}(q+1)r(v)$, which is independent of the schedule and hence for most of the problems considered in this chapter, both objective functions are equivalent.

The constant does, however, depend on the size of the instance, which means that both objective functions behave differently once we consider their asymptotic values for a series of instances with growing size. We will point out these differences whenever they are relevant.

The base version of the problem of minimizing the weighted sum of completion times on identical parallel machines with unit processing times has long been known to be efficiently solvable (see, e. g., [Law64]). Machine eligibility restrictions have been considered in [Pin16], where it is proved that makespan minimization with unit processing times is efficiently solvable if eligibility constraints obey a special structure. The minimization of weighted sum of completion times (again, for unit processing times) is also possible in polynomial time [BJK97]. Alternatively, eligibility constraints can be modeled by imposing sufficiently high processing times if a job is processed on the wrong machine (which however leaves both the domain of identical machines and that of unit processing times).

Precedence constraints also appear regularly in the scheduling literature. However, they are usually assumed to be binding in the sense that a job *cannot* be processed before its dependencies are completed. In contrast, the setting that we consider does

allow for a service to be deployed before its dependencies are satisfied, albeit at the cost of slightly reducing the reward obtained from that service.

Whenever (hard) precedence constraints are introduced, the resulting problem almost always becomes $\mathbb{NP}$-hard, even for very restricted cases (see, e. g., [LR78] for the case of a single machine and unit processing time or unit rewards, [LRB77] for the case of m machines with both unit processing time and unit reward). One of the questions that are answered in this chapter is therefore whether the transition to soft precedence constraints (with a fixed penalty per unit of violation equal to the job's weight) makes previously $\mathbb{NP}$-hard problems efficiently solvable.

A different dimension considered in this chapter is that of game-theoretic solution concepts for interdependent scheduling games. There exists a significant body of literature on "Scheduling Games", however the settings that are typically considered differ markedly from the scenario treated in this chapter:

Most works can be classified into the domains of complete and incomplete information (see [HMU07]). In the latter case, agents represent machines (e. g., [Kou14]) or (collections of) jobs (e. g., [ABP06]) and may decide to reveal information about their parameters. In the case of complete information, the typical setting is that of agents choosing a machine on which to process the jobs that they control. Here, the order in which jobs are performed on one machine is determined by some previously known scheduling policy. This raises the mechanism design question of determining a scheduling policy for each machine in such a way as to incentivize agents to chose an assignment of their jobs to machines that is as close to globally optimal as possible (e. g., [CKN09]).

All of the literature cited above does not assume the existence of any precedence constraints. Indeed, game-theoretical properties of scheduling contexts with precedence constraints seem to have not received much attention hitherto. The only strain of literature in this direction that we are aware of considers multi-agent project scheduling where agents may decide to speed up (at additional cost) the completion of individual services that they control in order to achieve an earlier termination of the entire project, for which they in turn receive a previously determined reward [BB11; Agn+15].

4.3 Reward Maximization and Best Responses

We first consider two natural decisions problems in the context of Interdependent Scheduling Games.

Name:	ISG Reward Maximization
Input:	An ISG $(k, (T, E), r, \rho)$ and an integer W.
Question:	Is there a schedule π such that $R(\pi) \geq W$?

Name:	ISG Best Response
Input:	An ISG $(k, (T, E), r, \rho)$, player $i \in [k]$, schedule π_{-i} for all players $\{1, \dots, k\} \setminus \{i\}$ and an integer W.
Question:	Is there a schedule π_i for player i such that $R_i(\pi_{-i}, \pi_i) \geq W$?

Note that ISG Best Response is a generalization of ISG Reward Maximization for a single player where additional soft dependencies may prevent a service from generating a reward before a fixed timeslot.

Ignoring machine eligibility constraints for a moment, one might be tempted to think that ISG Reward Maximization is in fact equivalent to the ($\mathbb{NP}$-hard) scheduling problem $(n|m|I, prec, p_i = 1|\sum C_i)$ in the notation from [LRB77]: the problem of minimizing the (unweighted) sum of completion times in a setting with m identical machines and n jobs with unit processing times under precedence constraints. As we have seen in (4.1), minimizing the sum of completion times is the same as maximizing the total reward. If we could thus assume that every reward-maximizing schedule was such that all services activate immediately, then we would in fact be looking for a schedule satisfying all dependencies as strict precedence constraints.

And indeed, for Interdependent Scheduling Games with uniform rewards, the following simple lemma holds:

Lemma 4.9
Let $(k, (T, E), r, \rho)$ be an Interdependent Scheduling Game with uniform rewards and $|T_1| = |T_2| = \cdots = |T_k| =: q$. Let π be an arbitrary solution. Then,

$$kq \leq R(\pi) \leq \frac{kq(q+1)}{2}.$$

Proof. The instance contains kq services, each contributing a reward of 1 for every timestep in which the respective service is active. Every service activates at the latest in the last timestep. In this case, each service contributes a reward of exactly 1, totaling kq. On the other hand, if every service activates immediately, the total reward per player is $\sum_{t=1}^{q}(q+1-t) = \sum_{t=1}^{q} t = \frac{q(q+1)}{2}$. □

For uniform ISGs, a conflict-free schedule (if it exists) thus also maximizes total reward, as it achieves the maximal possible reward of $\frac{k \cdot q(q+1)}{2}$. However, in general such a schedule need not exist. Furthermore, if we drop the assumption of uniformity, then there might be no welfare-maximizing schedule that is also conflict-free, even if a conflict-free schedule exists, as the following example will show:

Definition 4.10
Let $(k, (T, E), r, \rho)$ be an ISG and π a corresponding schedule. If $a_\pi(v) = \pi(v)$ for all $v \in T$, then π is called *conflict-free*.

Figure 4.2: Two schedules for an ISG that demonstrate that a conflict-free schedule might not be welfare-maximizing (see Example 4.11): The unique (up to permutation of equivalent services) welfare-maximizing schedule π^B has a conflict, while the unique (up to permutation of equivalent services) conflict-free schedule has lower welfare.

Example 4.11 ([Abe+16])
Consider the ISG and two schedules π^A and π^B shown in Fig. 4.2. The schedule π^A on the left is conflict-free while the schedule π^B on the right has a conflict (the first service of player 2 only activates at time 2). Despite the conflict, π^B has higher welfare; the two services with reward 100 become active simultaneously in step two, providing a higher reward to player 2 and more total welfare. The schedule π^A, on the other hand, is the only conflict-free schedule (up to permutation of equivalent services): The order of services for player 1 is uniquely determined by the internal dependencies and player 2 has only one service that can immediately activate if it is scheduled first.

The following lemma will prove useful in several of our proofs in this chapter. It captures the fact that for an individual player, it is always advantageous to deploy a service only after at least those of its prerequisites have been deployed that are controlled by the same player. In other words, we can always find a best response that is "conflict-free" with respect to the services controlled by a single player.

Lemma 4.12 ([Abe+16])
Let $(k, (T, E), r, \rho)$ be an ISG and π_{-i} a schedule for all players except player i. Then, there exists a best response π_i for player i such that

$$\pi_i(u) < \pi_i(v) \text{ for all } (u, v) \in E_i. \tag{4.2}$$

Proof. For any player $i \in [k]$ and schedule π_i of that player, let $\sigma(\pi_i) := |\{v \in T_i : \exists (u, v) \in E_i \text{ with } \pi_i(u) > \pi_i(v)\}|$ denote the number of services in T_i that depend on another service in T_i which is scheduled later. Let π_i' denote a best response of player i such that $\sigma(\pi_i')$ is minimal among all best responses. We suppose for contradiction that the statement is false, therefore $\sigma(\pi_i') \geq 1$. Choose $(u, v) \in E_i$ with $\pi_i'(u) > \pi_i'(v)$ in such a way that there is no u' with $(u', v) \in E_i$ and $\pi_i'(u') > \pi_i'(u)$.

Figure 4.3: Construction of the schedule π_i' in the proof for Lemma 4.12.

Consider the following modified schedule π_i^* for player i (see Fig. 4.3):

$$\pi_i^*(w) := \begin{cases} \pi_i'(w) - 1 & \pi_i'(v) + 1 \leq \pi_i'(w) \leq \pi_i'(u) \\ \pi_i'(u) & w = v \\ \pi_i'(w) & \text{else.} \end{cases}$$

The following two properties hold:

a) The schedule π_i^* is also a best response. The only service that is scheduled to a later time in π_i^* (and hence could cause itself or services depending on it to generate a smaller reward) is v. However, v did not activate before time step $\pi_i'(u)$ under π' and as $\pi_i^*(v) = \pi_i'(u)$ the reward generated by v does not change. The same holds for all services that depend on v.

b) $\sigma(\pi_i^*) < \sigma(\pi_i')$. First, note that v does not contribute towards σ anymore as $\pi_i^*(v) > \pi_i^*(u)$ (the same holds for all other services that v depends upon by the maximality of u). Now, consider any service w that did not contribute to $\sigma(\pi_i')$. As the ordering among all services except v remains the same, such a w can only contribute to $\sigma(\pi_i^*)$ if it depends on v and $\pi_i'(v) < \pi_i'(w) < \pi_i^*(v) = \pi_i'(u)$. But then, it must also depend on u by transitivity and hence it contributed to $\sigma(\pi_i')$ already.

From a) and b), we obtain a contradiction to minimality of $\sigma(\pi_i')$, which concludes the proof. □

Looking at the proof of this lemma, it is easy to see that, in contrast to Example 4.11, conflict-freeness is a necessary condition for reward-maximization in the case of a single player.

Theorem 4.13 ([Abe+16])
For one player and general rewards, every welfare-maximizing schedule is a conflict-free schedule.

Proof. This follows by an observation about the proof of Lemma 4.12: In the one-player case, service u activates immediately under schedule π_i' by its maximality among dependencies for which v has to wait. Hence, it also activates immediately under schedule π_i^*, which is one time step earlier than under schedule π_i'. Schedule π_i^* hence generates strictly more reward than schedule π_i'. Thus, any schedule π with $\sigma(\pi) > 0$ cannot be welfare-maximizing. □

The following two theorems prove that the problem ISG Reward Maximization is $\mathbb{NP}$-complete for quite restricted instances as soon as either rewards are non-uniform or the number of players can be arbitrarily large. This is in line with the complexity results for strict precedence constraints mentioned above, which shows that softness of the precedence constraints does not make the corresponding problems computationally easier. We will use reductions from the following two problems, which are known to be $\mathbb{NP}$-hard:

Name:	Min 2SAT [KKM94]
Input:	An integer k and a Boolean formula F in Conjunctive Normal Form where each clause contains exactly two literals.
Question:	Is there an assignment to the variables of F such that at most k clauses are satisfied?

Name:	Single Machine Weighted Completion Time [LR78]
Input:	A set of jobs $J_i \in J$. Each job has a weight w_i and a processing time $p_i = 1$. A precedence relation $\prec$ such that if $i \prec j$ then J_j cannot be scheduled before J_i. An integer k.
Question:	If C_i is the completion time of job J_i, is there an ordering of the jobs such that $\sum_{i \in J} w_i C_i \leq k$?

Theorem 4.14 ([Abe+16])
ISG Reward Maximization *is* $\mathbb{NP}$*-complete, even when the rewards are uniform and each player has two services.*

Proof. The problem is in $\mathbb{NP}$ since we can efficiently calculate the welfare of a given schedule. We reduce from the $\mathbb{NP}$-hard problem Min 2SAT.

For each variable x in F, create a player P_x with services $T_x = \{x, \neg x\}$. For each clause c in F, create a player P_c with services $T_c = \{c_1, c_2\}$. For each clause $c = (\ell_1 \vee \ell_2)$, the precedence graph contains (c_1, c_2), (ℓ_1, c_1), and (ℓ_2, c_1). Rewards are uniform, and we set $W = 3n + 3m - k$, where n and m are the number of variables and clauses of F.

It remains to prove that F has an assignment satisfying at most k clauses if and only if the ISG has a schedule generating a reward of at least W. For the forward direction, suppose F has an assignment $\alpha : \mathrm{var}(F) \to \{\mathrm{true}, \mathrm{false}\}$ satisfying at most k

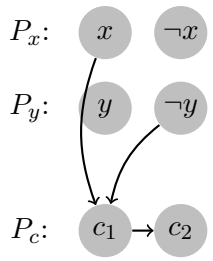


Figure 4.4: An example for the substructure used in Theorem 4.14 to represent the clause $c = (x \vee \neg y)$. The schedule generates maximal reward if and only if all literals that appear in the clause are scheduled to the first slot.

clauses. Consider the schedule where, for each variable x, the player P_x schedules first the literal of x that is set to false by α, i. e., x is scheduled before $\neg x$ iff $\alpha(x) = \text{false}$. Additionally, for each clause c, the service c_1 is scheduled before c_2. This schedule generates a reward of 3 for each variable player: a reward of 1 at the first time step and a reward of 2 at the second time step. For a satisfied clause c, the schedule generates a reward of 2: at the first time step no reward is generated since the literal satisfying the clause is scheduled at the second time step and there is an arc from that literal to c_1, and a reward of 2 is generated at the second time step. For an unsatisfied clause c, the schedule generates a reward of 3: since both literals are set to false, they are both scheduled at the first time step. Thus, the total reward generated for this schedule is at least $3n + 3m - k$.

For the reverse direction, let π be a schedule generating a reward of at least W. Consider the assignment $\alpha : \text{var}(F) \to \{\text{true}, \text{false}\}$ with $\alpha(x) = \text{false}$ iff player P_x schedules x at the first time step. Note that at the second time step, each player generates a reward of 2. Also, each player corresponding to a variable generates an additional reward of 1 at the first time step since his services have in-degree 0. So, at least $3n + 3m - k - (3n + 2m) = m - k$ additional clause players generate a reward of 1 at the first time step. But, for each such clause c, the service c_1 is scheduled before c_2 and both literals occurring in c have to be scheduled at the first time step, which means that the assignment α sets these literals to false. Therefore, α does not satisfy the clause c. We conclude that α satisfies at most k clauses. □

Theorem 4.15 ([Abe+16])
For general rewards, ISG Reward Maximization *is* $\mathbb{NP}$*-complete even for a single player.*

Proof. We give a reduction from the $\mathbb{NP}$-hard problem Single machine weighted completion time.

For each job $J_i \in J$, create service t_i with reward $r_i = w_i$ and consider the same precedence graph as the one given for jobs. We set $W = (|J| + 1) \sum_{i \in J} w_i - k$. We

prove that an ordering π of jobs has a weighted completion time of at most k if and only if the schedule π in the ISG satisfies $R(\pi) \geq W$.

For an ordering π, let $C_i(\pi)$ denote the completion time of job J_i in the machine scheduling problem and note that in the ISG, the service t_i is also completed at time $C_i(\pi)$. Then, the weighted sum of completion times for π in the scheduling instance is given by $\sum_{i \in T} w_i C_i = \sum_{i \in J} r_i C_i$.

If π is a solution for the scheduling problem, then in particular it satisfies all precedence constraints. Thus, in the ISG, all services activate immediately and $R(\pi) = \sum_{i \in T} (|T| + 1 - C_i(\pi))\, r_i = (|T|+1) \sum_{i \in T} r_i - \sum_{i \in T} r_i C_i(\pi)$. But this implies that if $\sum_{i \in J} w_i C_i(\pi) \leq k$, then $R(\pi) \geq W$.

Conversely, if π is a schedule for the ISG with $R(\pi) \geq W$, then by Theorem 4.13, there exists a schedule π' which is conflict free with $R(\pi') \geq R(\pi) \geq W$. But now, using the same argument as above, we obtain that $\sum_{i \in J} w_i C_i(\pi') \leq k$, which concludes the proof. □

Turning to the problem ISG Best Response, which has to take into account precedence constraints imposed by other players, we recall that the problem is a generalization of ISG Reward Maximization for a single player. Theorem 4.15 thus immediately implies the following corollary.

Corollary 4.16
For general rewards, ISG Best Response *is* $\mathbb{NP}$*-complete.*

For uniform rewards, on the other hand, a player can efficiently compute his best response to the schedule of all other players' services. In the following lemma, given an ISG $(k, (T, E), r, \rho)$, a player $i \in [k]$ and a service $v \in T \setminus T_i$, we write $\pi_{-i}(v)$ for the time that service v is deployed under any schedule (π_i, π_{-i}), since this time does not depend on π_i (although its activation time might).

Lemma 4.17
Let $(k, (T, E), r, \rho)$ be an ISG and let $i \in [k]$ and let π_{-i} be a vector of schedules for all players except i. Let $\eta(v) := \max\{\pi_{-i}(w) : w \in T \setminus T_i, (w, v) \in E\}$. Then, π_i is a best response to π_{-i} if the following conditions hold for all $v \in T_i$:

a) π_i satisfies (4.2)*, and*

b) if $w \in T_i$ with $\pi_i(w) \geq \pi_i(v)$ and $\pi_i(w') \leq \pi_i(v)$ for all $w' \in T_i$ with $(w', w) \in E_i$, then $\eta(v) \leq \eta(w)$.

In the statement of the above lemma, $\eta(v)$ represents the lower bound on v's activation time imposed by π_{-i}. Note that, in particular, $(u, w) \in E_i$ implies that $\eta(w) \geq \eta(u)$ by transitivity of E.

Proof. Let π_i be a schedule for player i satisfying the conditions above. Suppose for contradiction that π_i is not a best response. Let π_i^* be a best response satisfying condition (4.2) (which exists by Lemma 4.12) which maximizes the first index for which any such best response differs from π_i. Formally, there exists $n \in \mathbb{N}$ such that $(\pi_i^*)^{-1}(i) = (\pi_i)^{-1}(i)$ for all $i < n$ and there is no best response π_i' satisfying condition (4.2) with $(\pi_i')^{-1}(i) = (\pi_i)^{-1}(i)$ for all $i < n+1$.

We will prove that there exists a best response coinciding with π_i in the first $n+1$ entries, which yields a contradiction. Let $a := (\pi_i^*)^{-1}(n)$ and $b := (\pi_i)^{-1}(n)$ denote the services scheduled at timestep n by the schedules π_i^* and π_i, respectively. Since π_i and π_i^* coincide in the first n entries, it holds that $\pi_i^*(a) < \pi_i^*(b)$. We distinguish three cases:

i) $\eta(b) \leq \pi_i^*(a)$. In this case, we adapt π_i^* by moving b to the position of a and moving all services between $\pi_i^*(a)$ and $\pi_i^*(b)$ by one position. Formally,

$$\pi_i^{**}(w) := \begin{cases} \pi_i^*(w) + 1 & \pi_i^*(a) \leq \pi_i^*(w) < \pi_i^*(b) \\ \pi_i^*(a) = k & w = b \\ \pi_i^*(w) & \text{else.} \end{cases}$$

Because of condition a), π_i^{**} satisfies condition (4.2) and b activates immediately under the schedule π_i^{**}. The reward generated by service b thus increases by $\pi_i^*(b) - \pi_i^*(a)$ and at the same time the reward generated by at most $\pi_i^*(b) - \pi_i^*(a)$ other services decreases by 1. Hence, π_i^{**} is still a best response and satisfies condition (4.2). Furthermore, $(\pi_i)^{-1}(n) = (\pi_i^{**})^{-1}(n)$, contradicting maximality of π_i^*.

ii) $\pi_i^*(a) < \eta(b) < \pi_i^*(b)$. We use the procedure from above to move b to the position $\eta(b)$, thereby constructing a new best response $\tilde{\pi}_i^*$ with $\tilde{\pi}_i^*(b) = \eta(b)$, which still maximizes the first index for which any best response satisfying (4.2) differs from π_i. Then, proceed as in iii) below.

iii) $\pi_i^*(b) \leq \eta(b)$. Then, by condition b) above and since π_i^* satisfies (4.2), it holds that $\eta(b) \leq \eta(a)$. Therefore, $\eta(a) \geq \eta(b) \geq \pi_i^*(b)$, which means that a will never activate before time $\pi_i^*(b)$. We move b to the position of a (without any effect on the generated reward, since both a and b will not activate before $\pi_i^*(b)$), making sure that all successors of a can be re-scheduled to a position no later than $\pi_i^*(b)$, as well.

Formally, we construct a schedule π_i^{**} as follows: Set $\pi_i^{**}(b) := \pi_i^*(a)$. Let $a' = \operatorname{argmin}\{\pi_i^*(v) \mid v \in N^{\text{out}}(a)\}$ be the earliest successor of a. If $\pi_i^*(a') > \pi_i^*(b)$, set $\pi_i^{**}(a) := \pi_i^*(b)$ and $\pi_i^{**}(w) := \pi_i^*(w)$ for all other services. Otherwise, set $\pi_i^{**}(a) := \pi_i^*(a')$ and let a'' be the earliest successor of a'. Proceed with a'' (and possibly its earliest successor) as above until either no more successors are left

or an earliest successor a^* satisfies $\pi_i^*(a^*) > \pi_i^*(b)$. Since $\eta(a) \geq \pi_i^*(b)$, none of the services that were moved activated before $\pi_i^*(b)$ anyway and thus all rewards stay the same and the resulting schedule π^{**} is still a best response. Furthermore, it still satisfies condition (4.2). However, $(\pi_i)^{-1}(n) = (\pi_i^{**})^{-1}(n)$, contradicting maximality of π_i^*.

In all three cases, we reach a contradiction which proves that our assumption was wrong and π_i is indeed a best response. □

Using this lemma, we can prove the following theorem:

Theorem 4.18 ([Abe+16])
For ISG with uniform rewards, the problem ISG Best Response *is solvable in polynomial time.*

Proof. We give a greedy algorithm that solves the problem optimally. Consider the subgraph G_i of G induced by the set T_i of services belonging to player i and let π_{-i} be a vector of schedules for all other players. For any $v \in T_i$, let $\eta(v)$ be defined as in Lemma 4.17.

Starting from the first time step, successively schedule a service which minimizes η among all services with no incoming edges in G_i. Such a service always exists, as G (and hence all subgraphs) is acyclic. The service with all its (outgoing) edges is then removed from G_i.

The resulting schedule obviously satisfies both conditions from Lemma 4.17 and is hence a best response. □

Again, as ISG Best Response covers all instances of ISG Reward Maximization with only a single player,Theorem 4.18 immediately yields the following.

Corollary 4.19 ([Abe+16])
For uniform rewards and a single player, ISG Reward Maximization *can be solved in polynomial time.*

4.4 Nash Dynamics and Equilibria

The problem of ISG Reward Maximization presented in the previous section dealt with determining the globally optimal schedule for an Interdependent Scheduling Game. However, in order to implement such a schedule, a centralized entity needs to be able not only to compute it but also to compel the agents to act according to this schedule.

In many real-life applications, such an entity does not exist or does not dispose of sufficient coercive power. Instead, agents are free to chose a schedule that suits their own interest as well as possible.

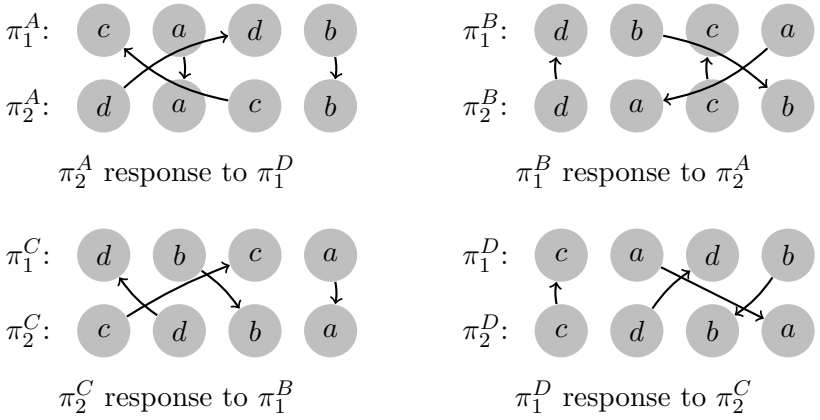


Figure 4.5: The instance from Example 4.20. The figure shows four schedules $\pi^A, \pi^B, \pi^C, \pi^D$ where each schedule arises from one player choosing a best response to the previous schedule while the other player keeps his current schedule.

We discussed the computational complexity of determining such an individually optimal schedule (a best response) in the previous section. In this section, we will investigate the consequences of different agents each acting in order to maximize their own reward while observing (or making assumptions about) the actions of other agents.

A natural question is whether a steady state always emerges from a situation where the schedule chosen by each agent is public knowledge and all agents may update their decision in reaction to the information about other agents' choices. This need not be the case, as the following example shows.

Example 4.20 ([Abe+16])
Consider the instance of an ISG and the sequence of best responses shown in Fig. 4.5: From π^A to π^B, player 1 updates his schedule to choose a best response, from π^B to π^C, player 2 updates his schedule, from π^C to π^D, player 1 updates his schedule and from π^C to π^D, player 2 updates his schedule. Now, player 1 can choose a best response that takes us back to the schedule π^A, completing the cycle. Note that each of the best responses is non-unique and indeed, choosing a different best response might break the cycle.

In light of Example 4.20, where a steady state does not arise automatically from players iteratively selecting best responses to each others' actions, the question arises whether such a state exists in the first place. Such a schedule where no player has an incentive to unilaterally deviate to a different schedule is known as a pure Nash equilibrium (see Definition 4.2).

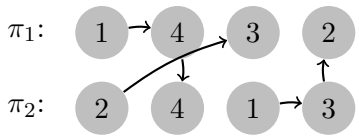


Figure 4.6: An ISG for two players with general rewards which does not admit a pure Nash equilibrium (see Theorem 4.21).

4.4.1 ISGs with General Rewards

In the case of general rewards, this is not necessarily the case, even for two players.

Theorem 4.21 ([Abe+16])
There exists an ISG with two players and general rewards which does not admit a pure Nash equilibrium (PNE).

Proof. Consider the the instance shown in Fig. 4.6. Note that the node labels indicate the reward associated with each service, but we will also use them as identifiers for the respective service, i. e., service 1 for player 1 is the service which is controlled by player 1 and has reward 1.

Assume that this game admits a PNE. Any best response of player 1 must satisfy that service 4, being the highest reward service, is scheduled immediately after service 1. Therefore, any possible best response of player 1 has to adopt one of the following schedule configurations (we denote, e. g., by $(1, 4, *, *)$ the set of schedules that have service 1 at the first position and service 4 at the second):

a) $\pi_1 \in (1, 4, *, *)$,

b) $\pi_1 \in (*, 1, 4, *)$ or

c) $\pi_1 \in (*, *, 1, 4)$.

For the same reason, in any best response of player 2, service 4 must be scheduled as soon as possible. These observations narrow the set of possible pure Nash equilibria to three cases, each corresponding to player 1 playing a schedule conforming to one of the patterns listed above:

a) If in a pure Nash equilibrium, player 1 selects a schedule of the form $\pi_1 \in (1, 4, *, *)$, then player 2's unique best response is $\pi_2 = (2, 4, 1, 3)$. But then, player 1's unique best response is $\pi_1 = (3, 1, 4, 2)$, a contradiction.

b) If in a pure Nash equilibrium, player 1 selects a schedule of the form $\pi_1 \in (*, 1, 4, *)$, then player 2's unique best response is $\pi_2 = (1, 3, 4, 2)$. But then, player 1's unique best response of $\pi_1 = (1, 4, 2, 3)$, a contradiction.

c) If in a pure Nash equilibrium, player 1 selects a schedule of the form $\pi_1 \in (*, *, 1, 4)$, then player 2's best response must satisfy $\pi_2 \in \{(2, 1, 3, 4), (1, 3, 2, 4)\}$. But if $\pi_2 = (2, 1, 3, 4)$, then $\pi_1 = (3, 1, 4, 2)$ is player 1's unique best response and if $\pi_2 = (1, 3, 2, 4)$, then $\pi_1 = (1, 4, 3, 2)$ is player 1's unique best response. In both cases, we arrive at a contradiction.

Hence, none of the three cases can be a Nash equilibrium, which concludes the proof.□

We can use the above example to construct a reduction from the well-known $\mathbb{NP}$-hard problem 3SAT to show that deciding whether a pure Nash equilibrium exists for a particular instance of an ISG is $\mathbb{NP}$-hard.

Name:	3SAT [Kar72]
Input:	A Boolean formula F in Conjunctive Normal Form where each clause contains exactly 3 literals.
Question:	Is there an assignment to the variables of F such that all clauses are satisfied?

Theorem 4.22 ([Abe+16])
Deciding whether an ISG with general rewards admits a pure Nash equilibrium is $\mathbb{NP}$-hard, even when each player has at most 4 services.

Proof. We give a reduction from the $\mathbb{NP}$-hard problem 3SAT.

For each variable x in the formula F, create a player P_x with services $T_x = \{x, \neg x\}$. Both services have the same reward $r(x) = r(\neg x) = 1$ (we may add two dummy services with reward 0). For each clause c in F, create a player P_c with services $T_c = \{c_1, c_2, c_3, d_c\}$ and set rewards to be $r(d_c) = 3$ and $r(c_1) = r(c_2) = r(c_3) = 4$. For each clause c, we create a gadget G_c corresponding to a copy of the ISG from Theorem 4.21 which admits no PNE and consists of 2 players with 4 services each. For each clause $c = (\ell_1 \vee \ell_2 \vee \ell_3)$ in F, the precedence graph contains arcs (ℓ_1, c_1),(ℓ_2, c_2),(ℓ_3, c_3) and arcs from service d_c to the 8 services of gadget G_c (see Fig. 4.7).

It remains to prove that F has an assignment satisfying all clauses if and only if the ISG admits a pure Nash equilibrium. For the forward direction, suppose F has an assignment $\alpha : \text{var}(F) \to \{\text{true}, \text{false}\}$ satisfying all clauses. Consider the schedule where, for each variable x, the player P_x schedules first the literal of x that is set to true by α, i. e., x is scheduled before $\neg x$ iff $\alpha(x) = \text{true}$. For each clause c, the player P_c schedules first its services that depend on literals set to true, then the services that depend on literals set to false and finally service d_c. Services in gadget G_c can be scheduled arbitrarily. This schedule is a pure Nash equilibrium: All services of each variable and clause player activate immediately when they are scheduled and the higher-valued services are scheduled first. Finally, the players in gadget G_c are

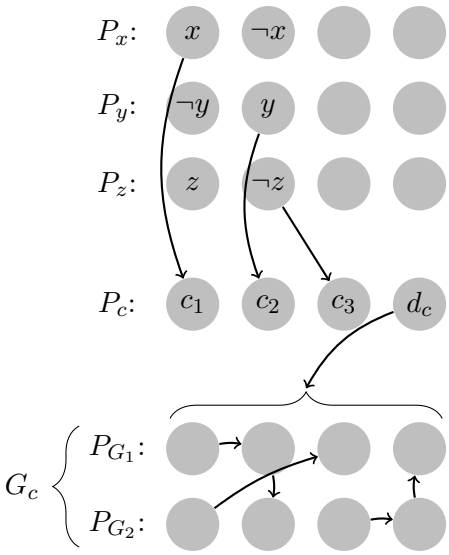


Figure 4.7: An example for the substructure used in Theorem 4.22 to represent the clause $c = (x \vee y \vee \neg z)$. If none of the literals that appear in c is scheduled in the first slot, any best response of P_c will schedule d_c in the first position, thereby exposing the structure from Theorem 4.21.

indifferent between all schedules because their services all become active in the last time step, given that service d_c was scheduled at the end.

For the reverse direction, suppose conversely that the game has a pure Nash equilibrium. Consider the assignment $\alpha : \mathrm{var}(F) \to \{\text{true}, \text{false}\}$ with $\alpha(x) = \text{true}$ iff player P_x schedules x at the first time step. We show that the assignment α satisfies F. Suppose some clause c is not satisfied. Then, none of its literal services will be activated before the second time step, which implies that service d_c is the only service of player P_c that can activate in the first time step. Hence, all best responses for the clause player P_c put service d_c into the first time slot, giving the player a reward of 36. This means that services in gadget G_c have no restrictions imposed from the outside (i. e., from the service d_c). But G_c for itself does not admit a Nash equilibrium, and hence the entire game does not either, a contradiction. Thus, all clauses are satisfied by the assignment α. □

4.4.2 ISGs with Uniform Rewards

In the previous section, we saw that in general we can not expect a pure Nash equilibrium to exist in every ISG. Indeed, even determining whether this is the case for a particular instance turned out to be an $\mathbb{NP}$-hard problem.

However, these questions get significantly easier to solve if we restrict rewards to be uniform: In this case a pure Nash equilibrium always exists and it can be efficiently computed.

Theorem 4.23 ([Abe+16])
Any ISG with uniform rewards admits a pure Nash Equilibrium which can be computed in polynomial time.

Proof. We iteratively construct a schedule in a way which guarantees that every player's schedule is a best response. For each service v, let

$$N_i^{\text{in}}(v) := (N^{\text{in}}(v) \cup \{v\}) \cap T_i$$

denote those services that v depends on and that are controlled by player i (including v itself in the case where $\rho(v) = i$). Denote by $T_i^{(t)}$ the set of services of player i that are already scheduled before iteration t (we set $T_i^1 = \emptyset$ for all $i \in [k]$). Let $\alpha_i^{(t)} := |T_i^{(t)}|$ denote the number of such services.

In every iteration, we will choose a service and schedule it together with all other (remaining) services that it depends on. By transitivity, this means that for a service $v \in T_i^{(t)}$, $a_\pi(v)$ is well-defined during iteration t, since all of its prerequisites have already been scheduled. We can therefore define

$$\mu_i^{(t)}(v) := \begin{cases} \max\{a_\pi(w) \mid w \in N_i^{\text{in}}(v)\}, & \text{if } N_i^{\text{in}}(v) \subset T_i^{(t)} \\ \alpha_i^{(t)} + |N_i^{\text{in}}(v) \setminus T_i^{(t)}|, & \text{else,} \end{cases}$$

which represents a lower bound on the earliest possible activation time of the service v based on its dependency on services from T_i. Specifically, if all of the prerequisites of v which are controlled by player i have already been scheduled (the first case), then the maximum of their activation times is the earliest time that v itself can become active and begin to generate reward (although it might itself be scheduled before that) Also note that services controlled by players other than i might still keep v inactive after that. Otherwise, in the second case (which in particular always applies if $v \in T_i \setminus T_i^{(t)}$), the service v cannot activate until itself and all of its prerequisites have been scheduled themselves. Since the first $\alpha_i^{(t)}$ slots in player i's schedule are already filled by services in $T_i^{(t)}$, the remaining prerequisites from $N_i^{\text{in}}(v) \setminus T_i^{(t)}$ must be scheduled after that.

Furthermore, we define $\mu^{(t)}(v) := \max_{i \in [k]} \mu_i^{(t)}(v)$ which represents a lower bound for the activation time of v in any schedule which is a "completion" of the partial schedule from iteration t. Note in particular that this bound is tight: It is achieved if v and all remaining prerequisites are scheduled immediately as the next services. Note that $\mu^{(t)}(v)$ can hence only increase from one iteration to the next and that it reaches the value $a_\pi(v)$ as soon as service v and all its predecessors are scheduled and is constant after that.

We will prove that the following conditions hold for every iteration t and player i:

i) If $v, w \in T_i^{(t)}$, thencondition (4.2) holds. If $v \in T_i^{(t)}$ and $w \in T_i \setminus T_i^{(t)}$, then $(w, v) \notin E_i$.

ii) If $v \in T_i^{(t+1)} \setminus T_i^{(t)}$, then $\mu^{(t)}(v)$ is minimal among all services from the set $T_i \setminus T_i^{(t)}$ and $\mu^{(t)}(v) \geq \mu^{(t+1)}(w)$ for all $w \in T_i^{(t)}$.

To see that these conditions are also sufficient for the resulting schedule π_i to be a best response, observe the following: Since all services from T_i are eventually added to the set $T_i^{(t)}$, this means that at the end, i) implies that (4.2) holds for all $v, w \in T_i$.

By Lemma 4.17, if the resulting schedule for each player satisfies condition b) from that lemma, as well, then each player's schedule is a best response and we have thus found a pure Nash equilibrium. Suppose otherwise and let i denote a player such that condition b) is violated for a service $v^* \in T_i$, i. e. there exists $w^* \in T_i$ with $\pi_i(w^*) \geq \pi_i(v^*)$ and $\pi_i(w) \leq \pi_i(v^*)$ for all $w \in T_i$ with $(w, w^*) \in E_i$, but $\eta(w^*) < \eta(v^*)$. Assume w. l. o. g. that v^* is such that $\pi(v^*)$ is minimal among all such services.

Let t denote the iteration where the service v^* is scheduled. Then the above implies that $w^* \in T_i \setminus T_i^{(t)}$ (otherwise $\pi_i(w^*) < \pi_i(v^*)$). Furthermore, all predecessors w of w^* within T_i are either already scheduled or will be scheduled in iteration t (otherwise, $\pi_i(w) > \pi_i(v^*)$).

Observe that for all players $j \in [k] \setminus \{i\}$ and $v \in T_i$, it holds that $\mu_j^{(t)}(v) \leq \eta(v)$:

- If $N_j^{\text{in}}(v) \subset T_j^{(t)}$, then the last predecessor of v cannot be scheduled any earlier than $\alpha_j^{(t)} + |N_j^{\text{in}}(v) \setminus T_j^{(t)}|$ and hence $\mu_j^{(t)}(v) \leq \eta(v)$.

- If, on the other hand, $N_j^{\text{in}}(v) \setminus T_j^{(t)} = \emptyset$ (i. e. all predecessors of v that are controlled by player j have already been scheduled), then this implies that all of their predecessors have already been scheduled, as well. Thus, for every $w \in T_j$ with $(w, v) \in E$, it holds that $a_\pi(w) \leq \max\{\eta(v)\}$

Overall, we obtain that $\mu^{(t)}(v) \leq \max\{\eta(v), \mu_i^{(t)}(v)\}$.

On the other hand, since v^* and all its predecessors are scheduled in iteration t, this implies, as argued above, that the bound given by $\mu^{(t)}$ on its activation time is tight, which implies that $\mu^{(t)}(v^*) = a_\pi(v^*) \geq \eta(v^*) > \eta(w^*)$. But since by ii), $\mu^{(t)}(w^*) \geq \mu^{(t)}(v^*)$, we must have that $\mu_i^{(t)}(w^*) > \eta(w^*)$ and thus $\mu^{(t)}(w^*) = \mu_i^{(t)}(w^*)$.

Furthermore, since we have assumed that $\pi_i(w) \leq \pi_i(v^*)$ for all $w \in T_i$ with $(w, w^*) \in E_i$, we have that $N_i^{\text{in}}(w^*) \setminus T_i^{(t)} = \{w^*\}$ and thus $\mu^{(t)}(w^*) = \mu_i^{(t)}(w^*) = \alpha_i^{(t)} + 1$. On the other hand, $\mu^{(t)}(w^*) \geq \mu^{(t)}(v^*) \geq \alpha_i^{(t)} + 1$ (the last inequality holds for any $v \in T_i \setminus T_i^{(t)}$), from which we obtain $\mu^{(t)}(w^*) = \mu^{(t)}(v^*) = \alpha_i^{(t)} + 1$. Both v^* and w^* would thus activate immediately when they are scheduled, and moving w^* to v^*'s

position, shifting all services in between by one slot analogously to our construction in Lemma 4.17, we obtain a schedule π_i^* with $R_i(\pi_i^*, \pi_{-i}) = R_i(\pi_i, \pi_{-i})$. The schedule π_i is hence a best response if and only if π_i^* is a best response. Iterating this argument (and observing that, if v' is the service for which condition b) from Lemma 4.17 is violated, then $\pi_i^*(v') > \pi_i(v^*)$), we obtain a schedule satisfying the prerequisites from Lemma 4.17, which proves that π_i was indeed a best response.

We now prove that both of the above conditions hold by induction over t, for $t = 1$ the set $T_i^{(t)}$ is empty for every player $i \in [k]$ and the conditions trivially hold.

Assume now that the conditions are satisfied for iteration t and proceed in the following way: Choose a service v^* that minimizes $\mu^{(t)}$ over all services not yet scheduled and that has no incoming edges from services belonging to the same player. Such a service must exist, since if $(w, v^*) \in E$ for some service w, then $\mu^{(t)}(w) \leq \mu^{(t)}(v^*)$. Let i be the player such that $v^* \in T_i$.

Denote by $S := \{v^*\} \cup N^{\text{in}}(v^*)$ the set consisting of v^* and all the services that v^* depends on (it may be the case that $S = \{v^*\}$). By induction, scheduling all services in S (respecting the ordering required by edges in E_i if necessary) satisfies condition (4.2) for all players i and $v, w \in T_i^{(t+1)}$. Furthermore, for every $v \notin S$ and $w \in S$, $(v, w) \notin E_i$ as otherwise $w \in S$. This means that condition i) is satisfied for iteration $t + 1$.

Regarding condition ii) note that for every $v \in S$, $\mu^{(t)}(v) = \mu^{(t)}(v^*)$ by minimality of v^* and the dependency of v^* on v. Hence for every player i, if $v \in T_i^{(t+1)} \setminus T_i^{(t)} \subset S$, then $\mu^{(t)}(v)$ is minimal among all services from the set $T_i \setminus T_i^{(t)}$.

Finally, for every $v \notin S$ and $w \in S$, it holds that $\mu^{(t+1)}(w) = \mu^{(t)}(w) = \mu^{(t)}(v^*) \leq \mu^{(t)}(v) \leq \mu^{(t+1)}(v)$ where the first equality holds because w and all its dependencies are scheduled in iteration t, the second equality was shown above and the inequalities follows by minimality of v^* and monotonicity of $\mu^{(t)}(v)$ in t. This proves that ii) is also satisfied for iteration $(t + 1)$.

By induction, this proves that the described procedure indeed constructs a pure Nash equilibrium for the given game. The number of iterations is bounded by T and hence the procedure runs in time polynomial in $|T|$. □

4.5 Price of Anarchy and Price of Stability

In the last section of this chapter, we investigate the relation between the different solution concepts considered in the previous sections: If a pure Nash equilibrium exists, can we provide any estimate on how close the total reward in the equilibrium is, compared to the maximal value that can be achieved by any schedule? As different Nash equilibria may exist that yield different total rewards, we consider both the price of anarchy and the price of stability (see Definition 4.3). The results in this section are original joint work with Kevin Schewior and have not previously been published (unless noted otherwise).

From our review of related literature in the area of machine scheduling (see Section 4.2), we recall that the reward of every player differs from the negative weighted sum of completion times by only a constant (see (4.1)). This means that all results presented so far equally hold if we assume that every player aims to minimize his weighted sum of completion times instead of maximizing his reward. However, the constant difference between the two objective functions does begin to play a role if we consider the asymptotic values of price of anarchy and price of stability, as we will do in this section. We will hence formulate all results with respect to one of the following two objective functions:

Let $(k, (T, E), r, \rho)$ be an instance of an Interdependent Scheduling Game with $|T_1| = |T_2| = \cdots = |T_k| =: q$ and π a feasible schedule. We consider the following objective functions, the relation of which we have already discussed briefly in Section 4.2:

$$R(\pi) = \sum_{i=1}^{k} R_i(\pi) = \sum_{i=1}^{k} \sum_{t=1}^{q} \sum_{v \in T_i, t \geq a_\pi(v)} r(v) = \sum_{i=1}^{k} \sum_{v \in T_i} (q + 1 - a(v))\, r(v)$$

$$C_\Sigma(\pi) = \sum_{i=1}^{k} \sum_{v \in T_i} a_\pi(v) r(v) = \sum_{i=1}^{k} \sum_{v \in T_i} (q+1) r(v) - R_i(\pi) = \sum_{v \in T} (q+1) r(v) - R(\pi)$$

We refer to $R(\pi)$ as the *total reward* for schedule π and to $C_\Sigma(\pi)$ as the *weighted sum of completion times* for schedule π. Note that for the case of uniform rewards, the latter simplifies to $kq(q+1) - R(\pi)$. Note further that the problem with respect to the weighted sum of completion times is a minimization problem, hence in this case we have to use the following slightly modified version of Definition 4.3 (A denotes the set of all schedules and N is the set of schedules that are pure Nash equilibria):

$$\mathrm{PoA}'(k, (T, E), r, \rho) := \frac{\max_{\pi \in N} C_\Sigma(\pi)}{\min_{\pi \in A} C_\Sigma(\pi)}$$

$$\mathrm{PoS}'(k, (T, E), r, \rho) := \frac{\min_{\pi \in N} C_\Sigma(\pi)}{\min_{\pi \in A} C_\Sigma(\pi)}$$

Remark 4.24
Given an Interdependent Scheduling Game with uniform rewards, Lemma 4.9 implies for any two solutions π, π' with $R(\pi) \geq R(\pi')$ that

$$\frac{R(\pi)}{R(\pi')} \leq \frac{q+1}{2}.$$

Furthermore, regarding the weighted sum of completion times, we obtain

$$\frac{C_\Sigma(\pi)}{C_\Sigma(\pi')} = \frac{kq(q+1) - R(\pi)}{kq(q+1) - R(\pi')} \leq \frac{kq(q+1) - kq}{kq(q+1) - \frac{kq(q+1)}{2}} = \frac{2q}{q+1}.$$

Remark 4.25
For general rewards, we can obtain the following bound with respect to the total reward for any two solutions π, π' with $R(\pi) \geq R(\pi')$ (the inequality in both cases follows from $a_{\pi'}(v) \geq 1$ and $a_{\pi'}(v) \leq q$):

$$\frac{R(\pi)}{R(\pi')} = \frac{\sum_{i=1}^{k} \sum_{v \in T_i} (q + 1 - a_\pi(v))\, r(v)}{\sum_{i=1}^{k} \sum_{v \in T_i} (q + 1 - a_{\pi'}(v))\, r(v)} \leq \frac{\sum_{i=1}^{k} \sum_{v \in T_i} qr(v)}{\sum_{i=1}^{k} \sum_{v \in T_i} r(v)} = q$$

Analogously, for the weighted sum of completion times, we obtain

$$\frac{C_\Sigma(\pi)}{C_\Sigma(\pi')} = \frac{\sum_{i=1}^{k} \sum_{v \in T_i} a_\pi(v) r(v)}{\sum_{i=1}^{k} \sum_{v \in T_i} a_{\pi'}(v) r(v)} \leq \frac{\sum_{i=1}^{k} \sum_{v \in T_i} qr(v)}{\sum_{i=1}^{k} \sum_{v \in T_i} r(v)} = q.$$

Applying the previous remarks to the price of anarchy immediately yields upper bounds for all combinations of uniform/general rewards on one hand and total reward/weighted sum of completion times on the other hand. Note that the case b)(i) previously appeared in [Abe+16, Theorem 10].

Theorem 4.26
Let $(k, (T, E), r, \rho)$ be an Interdependent Scheduling Game and $|T_1| = |T_2| = \cdots = |T_k| =: q$. The following holds:

a) The price of anarchy is bounded by q for both total reward and weighted sum of completion times.

b) If rewards are uniform (i. e., $r(v) = 1$ for all $v \in T$), then the price of anarchy is bounded

(i) with respect to total reward by $\frac{q+1}{2}$, and

(ii) with respect to weighted sum of completion times by $\frac{2q}{q-1}$.

The following example from [Abe+16, Theorem 9] easily shows that the bounds for cases b)(i) and b)(ii) are asymptotically tight.

Example 4.27
Consider the ISG and schedule π shown in Fig. 4.8. The worst pure Nash equilibrium is obtained (as shown) by scheduling the service, on which all other services of other players depend, at the end; as opposed to the welfare-maximizing schedule π^* achieved by moving this service to the beginning (which happens to be a pure Nash equilibrium, as well). The ratio between the total rewards for an instance with k players and q services each is

$$\frac{R(\pi^*)}{R(\pi)} = \frac{k \cdot q(q+1)/2}{q(q+1)/2 + (k-1)q} = \frac{k(q+1)}{q + 2k - 1} \xrightarrow{k \to \infty} \frac{q+1}{2}.$$

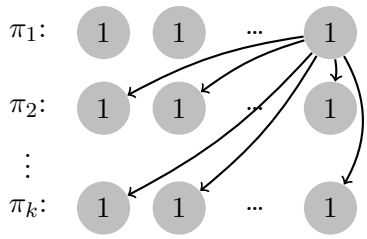


Figure 4.8: For this ISG, the upper bound on the price of anarchy from Theorem 4.26 is tight (see Example 4.27).

Similarly, for the weighted sum of completion times, we obtain

$$\frac{C_\Sigma(\pi)}{C_\Sigma(\pi^*)} = \frac{\sum_{i=1}^{k} \sum_{v \in T_i} a_\pi(v) r(v)}{\sum_{i=1}^{k} \sum_{v \in T_i} a_{\pi^*}(v) r(v)} = \frac{q(q+1)/2 + kq^2}{k \cdot q(q+1)/2} \xrightarrow{k \to \infty} 2.$$

Setting the rewards to 1 for one service of all but the first player and 0 for all other services, we obtain that the bound for the case a) is also asymptotically tight. However, the Nash equilibrium in the example from [Abe+16, Theorem 9] is weak and obviously not welfare-optimal. The welfare-maximizing Nash equilibrium on the other hand is also globally welfare-optimal. This raises the question of which total reward may be achieved by any pure Nash equilibrium. We first note that this question is computationally hard to answer, by the following simple corollary from Theorem 4.14.

Corollary 4.28
Given an ISG with uniform rewards and an integer k, determining whether a pure Nash equilibrium exists that achieves total reward at least k (or weighted sum of completion times at most k) is $\mathbb{NP}$-hard.

Nonetheless, can we determine a bound on the factor, by which an *optimal* pure Nash equilibrium (as opposed to *any* pure Nash equilibrium) may underperform the globally optimal solution? It turns out, as the following theorem shows, that the optimal pure Nash equilibrium may in general be no better than what is already guaranteed for all Nash equilibria (and indeed any solution) by Theorem 4.26. In other words, the bounds in the above theorem are asymptotically tight already with respect to the price of stability (it holds that price of stability $\leq$ price of anarchy).

Theorem 4.29
a) For every $q \geq 3$, there exists an instance $I_q^1 = (k, (T, E), r, \rho)$ of an Interdependent Scheduling game with general rewards and $|T_1| = |T_2| = \cdots = |T_k| = q$ such that
(i) $PoS(I_q^1) = q + \mathcal{O}(1)$ with respect to total reward, and
(ii) $PoS(I_q^1) = \frac{1}{2}q + \mathcal{O}(1)$ with respect to weighted sum of completion times.

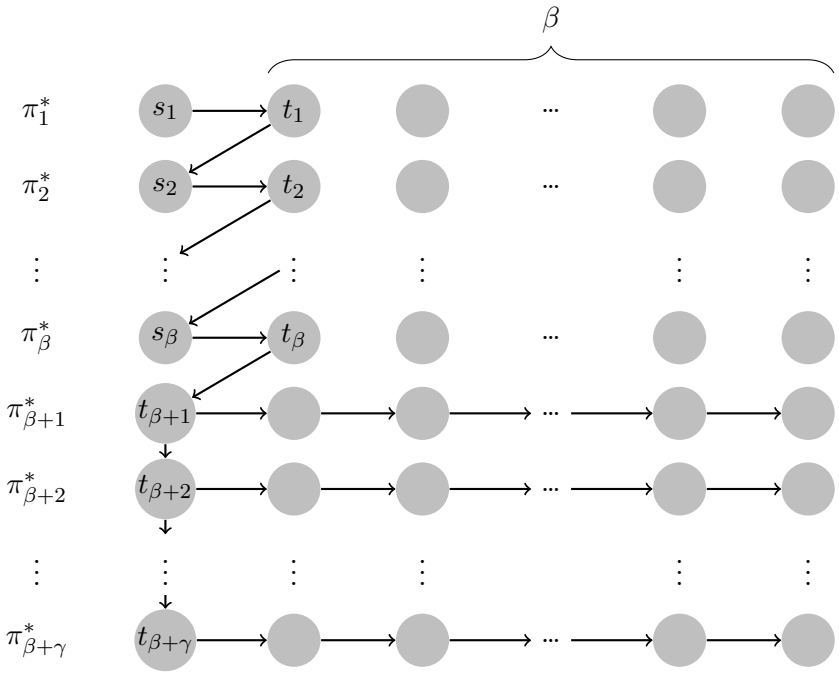


Figure 4.9: The ISG instance used in the proof for Theorem 4.29. The shown schedule π^* is welfare-maximizing and yields an total reward (for the case of uniform rewards) of $\frac{(\beta+1)(\beta+2)}{2} + (\beta-1)\cdot\left(\beta + \frac{\beta(\beta+1)}{2}\right) + \gamma\cdot\left(\beta + \frac{\beta(\beta+1)}{2}\right)$.

b) For every $q \geq 3$, there exists an instance $I_q^2 = (k, (T, E), r, \rho)$ of an Interdependent Scheduling game with uniform rewards and $|T_1| = |T_2| = \cdots = |T_k| = q$ such that

(i) $PoS(I_q^2) = \frac{1}{2}q + \mathcal{O}(1)$ with respect to total reward, and

(ii) $PoS(I_q^2) = 2 + \mathcal{O}(q^{-1})$ with respect to weighted sum of completion times.

Proof. Consider an instance of an Interdependent Scheduling Game as shown in Fig. 4.9. We have $q = \beta + 1$.

We first consider the case of uniform rewards. Let π be a Nash equilibrium schedule (one possible such schedule is shown in Fig. 4.10). First, note that $\pi_1(s_1) < \pi_1(t_1)$, since this is the only way for player 1 to achieve $R_1(\pi) \geq \frac{(\beta+1)(\beta+2)}{2}$. Now, let $i \in \{2, \ldots, \beta\}$. If $\max_{j<i}\{\pi_j(t_j), \pi_j(s_j)\} \leq \beta$, then player i can achieve $R_i(\pi) \geq \frac{(\beta+1)(\beta+2)}{2}$ only by choosing π_i such that $\max_{j<i}\{\pi_j(t_j), \pi_j(s_j)\} \leq \pi_i(s_i) < \pi_i(t_i)$.

If all players play a best response, we therefore have that

$$\pi_1(s_1) < \pi_1(t_1) \leq \pi_2(s_2) < \pi_2(t_2) \leq \cdots \leq \pi_\beta(s_\beta) < \pi_\beta(t_\beta)$$

and hence by the strict inequalities $\pi_\beta(t_\beta) \geq \pi_1(s_1) + \beta = \beta + 1$.

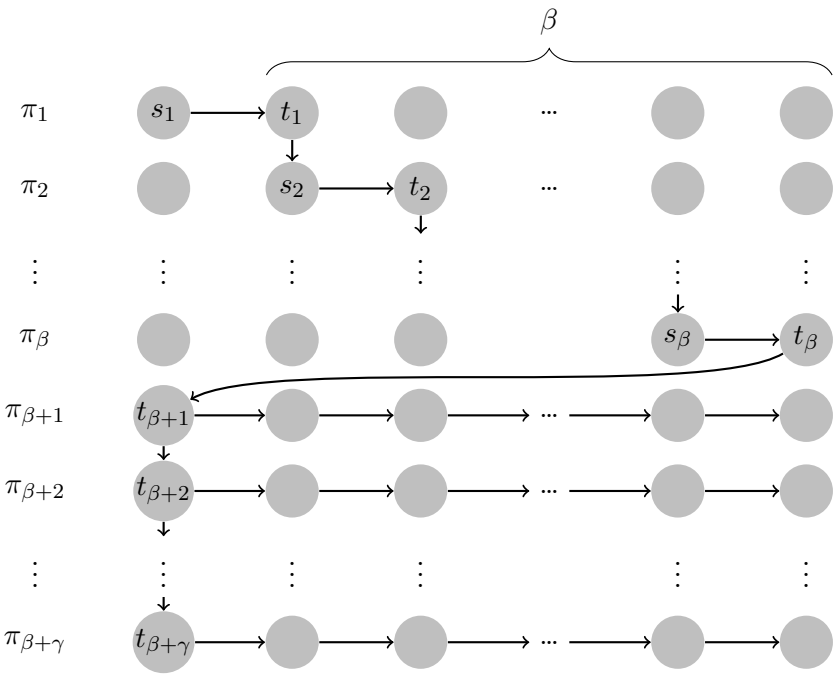


Figure 4.10: Equilibrium schedule for the ISG instance used in the proof of Theorem 4.29. The shown schedule achieves a total reward (in the case of uniform rewards) of $\beta \cdot \frac{(\beta+1)(\beta+2)}{2} + \gamma \cdot (\beta+1)$.

This implies that $t_{\beta+1}$ (and by transitivity $t_{\beta+2}, \ldots, t_{\beta+\gamma}$ as well as all services of players $\beta+1$ to $\beta+\gamma$ that depend on these) will only activate in the last timestep.

Therefore, for all $i \in \{\beta+1, \ldots, \beta+\gamma\}$, we get

$$R_i(\pi) \leq \beta + 1$$

by which we obtain $R(\pi) \leq \beta \cdot \frac{(\beta+1)(\beta+2)}{2} + \gamma \cdot (\beta+1) = \gamma \cdot (\beta+1) + \mathcal{O}(\beta^3)$.

On the other hand, for the schedule π^* shown in Fig. 4.9, we have

$$\begin{aligned} R(\pi^*) &= \frac{(\beta+1)(\beta+2)}{2} + (\beta-1) \cdot \left(\beta + \frac{\beta(\beta+1)}{2}\right) + \gamma \cdot \left(\beta + \frac{\beta(\beta+1)}{2}\right) \\ &= \gamma \cdot \left(\beta + \frac{\beta(\beta+1)}{2}\right) + \mathcal{O}(\beta^3). \end{aligned}$$

For any optimal schedule π^* and equilibrium schedule π in the given instances, we hence obtain (by choosing γ large enough, e. g. $\gamma := \beta^3$)

$$\frac{R(\pi^*)}{R(\pi)} \geq \frac{\gamma \cdot \left(\beta + \frac{\beta(\beta+1)}{2}\right) + \mathcal{O}(\beta^3)}{\gamma \cdot (\beta+1) + \mathcal{O}(\beta^3)} = \frac{\frac{1}{2}\beta^5 + \mathcal{O}(\beta^4)}{\beta^4 + \mathcal{O}(\beta^3)} = \frac{1}{2}\beta + \mathcal{O}(1).$$

and

$$\begin{aligned}\frac{C_\Sigma(\pi)}{C_\Sigma(\pi^*)} &= \frac{\sum_{i=1}^{k}\sum_{v\in T_i} a_\pi(v)}{\sum_{i=1}^{k}\sum_{v\in T_i} a_{\pi^*}(v)} = \frac{(\beta+\gamma)(\beta+1)(\beta+2)-R(\pi)}{(\beta+\gamma)(\beta+1)(\beta+2)-R(\pi^*)} \\ &\geq \frac{\mathcal{O}(\beta^3)+\gamma\cdot(\beta^2+2\beta+1)}{\mathcal{O}(\beta^3)+\gamma\cdot\left((\beta+1)(\beta+2)-\beta-\frac{\beta(\beta+1)}{2}\right)} \\ &= \frac{\beta^5+\mathcal{O}(\beta^4)}{\frac{1}{2}\beta^5+\mathcal{O}(\beta^4)} = 2+\mathcal{O}(\beta^{-1}).\end{aligned}$$

In the case of general rewards, we can set $r(t_j) = 1$ for all $j \in \{\beta+1, \beta+2, \ldots, \beta+\gamma\}$ and $r(v) = 0$ for all other services v. Note that the considerations above remain valid and hence, as above, $t_{\beta+1}$ (and by transitivity $t_{\beta+2}, \ldots, t_{\beta+\gamma}$ as well as all services of players $\beta+1$ to $\beta+\gamma$ that depend on these) will only activate in the last timestep. We obtain

$$R_i(\pi) \leq \begin{cases} 0 & i \in \{1,\ldots,\beta\} \\ 1 & i \in \{\beta+1,\ldots,\beta+\gamma\} \end{cases}$$

and hence $R(\pi) \leq \gamma$. On the other hand, for the schedule π^* shown in Fig. 4.9, we have

$$R(\pi^*) = \gamma\cdot\beta.$$

We can choose e. g. $\gamma := \beta$ to obtain

$$\frac{R(\pi^*)}{R(\pi)} \geq \beta$$

and

$$\begin{aligned}\frac{C_\Sigma(\pi)}{C_\Sigma(\pi^*)} &= \frac{\sum_{i=1}^{k}\sum_{v\in T_i} a_\pi(v)r(v)}{\sum_{i=1}^{k}\sum_{v\in T_i} a_{\pi^*}(v)r(v)} = \frac{\gamma\cdot(\beta+2)-R(\pi)}{\gamma\cdot(\beta+2)-R(\pi^*)} \\ &\geq \frac{\gamma\cdot(\beta+1)}{2\gamma} = \frac{\beta^2+\mathcal{O}(\beta)}{2\beta} = \beta+\mathcal{O}(1).\end{aligned}$$

□

4.6 Conclusion

To conclude this chapter, we will revisit some of the principal restrictions of the setting that we have considered in this chapter. Specifically, we have made several assumptions in defining what we call an ISG instance (Section 4.1.2), some of which can legitimately be questioned. In this section, we will briefly discuss the two most important assumptions (unit-time and acyclicity) and discuss the applicability of our results to the setting where these assumptions are relaxed.

unit time One serious limitation for the practical applicability of our setting is the assumption that every service requires the same amount of time to be deployed. This assumption can be relaxed quite easily, we can model any service v that requires multiple units of time to deploy as a corresponding series of services that require unit time. Each of these auxiliary services does not generate any reward, but is a prerequisite for the original service. Incoming and outgoing dependencies of the original service can be kept the same.

As the auxiliary services do not generate any reward and are no prerequisites for any other services (except for v), we can assume without loss of generality that the auxiliary services are scheduled right before v (in an arbitrary order). The only downside of this transition is that, even if the original instance had uniform rewards, we enter the setting of general rewards with all of the corresponding consequences regarding computational complexity and existence of Nash equilibria mentioned above.

acyclicity The consequences of dropping the assumption of acyclicity are less clear. Cycles within the services controlled by an individual player can be reduced to the above case of non-unit time: A set S of services that belong to the same player and that are cyclically dependent on each other will only activate (all at the same time) once the last service from that set has been deployed. It is thus equivalent to a single service that generates a reward of $\sum_{v \in S} r(v)$ and that requires $|S|$ units of time to deploy (services that depend on any of the services in S by transitivity depend on all services in S). By extension of the above argument, we can reduce the setting that allows dependency cycles within an individual player's services (and optionally that also allows non-unit time services) to the setting of general rewards (but unit time) that we have already covered.

When several players are involved in a single cycle C, the argument becomes slightly more involved, but again the two main problems discussed in this chapter (ISG Reward Maximization and ISG Best Response) can be expressed equivalently without that cycle:

For ISG Reward Maximization, we can choose one service $v^* \in C$ and assign to it the sum of all rewards, i. e., $r(v^*) = \sum_{v \in C} r(v)$. Furthermore, this service depends on all other services in the cycle. As the problem ISG Reward Maximization does not differentiate between rewards won by different players, the resulting problem is equivalent to the original problem, but removes the cycle.

For ISG Best Response and a fixed player i, we select one service $v^* \in C \cap T_i$ from the subset of services in C that are controlled by i. We assign to it the reward $r(v^*) = \sum_{v \in C \cap T_i} r(v)$ and let it depend on all other services in C. As a result, player i will start to accumulate the reward $\sum_{v \in C \cap T_i} r(v)$ as soon as all the services in the cycle have been deployed, as was the case in the original setting.

We have seen that both major assumptions from Section 4.1.2 can be dropped without seriously altering the definition of the related problems. However, both transformations require that we leave the domain of uniform rewards, to which most of the positive results in this chapter are constrained. The resulting computational problems are thus likely to be $\mathbb{NP}$-hard, as well. On the other hand, both cases do not cover the full generality of general rewards, particularly in interaction with processing times: For instance, the service with a higher reward that results from removing a precedence cycle only occurs together with a set of corresponding size of services with reward zero.

It might hence be worthwhile to investigate the special instances that result from the transformations described above to see if any of the positive results presented in this chapter can be translated.

Chapter 5

Summary and Outlook

In this thesis, we have approached the topic of power system analysis from three perspectives:

- representation of electrical power flows (Chapter 2)
- handling of large, structured optimization problems (Chapter 3)
- incentives in coupled scheduling environments (Chapter 4)

In all three areas, the questions that we started from were mostly algorithmic, asking how to address those topics *computationally* in the most efficient way. In this context, however, we have obtained a number of theoretical results, which are of interest in their own right. These include

- a generalization of a well-known characterization of extremal network flow solutions to *differential flows*,
- a precise connection between different cut selection criteria for Benders decomposition from the literature together with an overarching framework that encompasses them,
- a study of the boundary of computational hardness in scheduling problems with precedence constraints.

Our results in each chapter raise a number of interesting questions for future research within each theory area, which we have already discussed at the end of the respective chapter. Adding to those, we briefly highlight two directions of future research, which arise at the boundary between our results.

Benders Decomposition and Stochastic Programming In the context of stochastic programming, Benders decomposition (sometimes known as the *L-shaped method*) is a standard approach to deal with optimization problems covering a large number of stochastic scenarios. In the context of Energy System Analysis, however, a typical

example of a stochastic programming problem is computationally extremely difficult to solve, even using that method:

When analyzing the behavior of power systems in the context of intermittent generation from renewable sources, a key aspect of uncertainty is the availability of renewable energy sources, as well as the magnitude of demand. Since these vary on a relatively short timescale (typical models resolve to no more than 60 minutes) and consecutive steps are linked, e. g. by storage contents, the scenario tree resulting from a corresponding multi-stage stochastic programming model for a single year, even with only two scenarios per time step (e. g., high-wind and low-wind), consists of $2^{8760} \approx 10^{2637}$ vertices. Since on the other hand, this tree contains a lot of structure (e. g. transition probabilities are typically independent of any events beyond the immediately preceding time step), a popular technique in this context is *Stochastic Dual Dynamic Programming* (see, e. g., [PP91; PG08; Sha11]).

This technique, also known as *Nested Benders Decomposition* (see [Reb16]) bears close similarity with Benders decomposition, in particular each subproblem contains an approximation of the *future cost function*, representing the expected value of the cost incurred in all future timesteps. In each iteration, this approximation is refined by adding constraints to the epigraph, which essentially correspond to Benders cuts. As a consequence, many of our results on Benders decomposition are immediately applicable in this setting, as well. This application, among others, is also investigated in our ongoing research project *DecEnSys*.

Incentives in Decomposition and DC Power Flow Finally, incentives in the context of power system analysis, expansion and operation are a very broad topic, which we have only touched in this thesis. It is of interest to both economists and engineers and spans various lines of conflict between different agents: Between consumers and suppliers, between producers and transmission system operators and finally between different competitors in all of these sectors. At the same time, it can be formalized in a way that lends itself to provable mathematical statements.

In Chapter 4, we have focussed on one particular setting, that of selfish operators of interdependent infrastructure, who compete for rewards from the customers that they serve. While this setting is particularly interesting due to its connection with well-known problems from the scheduling domain, it leaves aside many of the other areas of competition outlined above. In connection with the other two main topics of this thesis, DC power flow and decomposition, the following questions seem particularly interesting for future research:

In terms of power flow models, most literature concerned with incentive structures in electricity markets today considers the simplified Transport model (if transmission capacities play a role at all). Indeed, the peculiarities of electrical power flows, even approximated at the level of the DC model present some additional challenges in

analyzing the underlying incentive structures: For instance, it may happen that an increase in transmission capacity along one corridor (due to the associated increase in susceptance) leads to a *reduced* transmission capacity between other nodes in the network. These challenges certainly warrant further research in this area.

With respect to decomposition, some researchers have attempted to introduce artificial "information barriers" between different parts of a problem to model selfish behavior: In the context of a decomposition approach, each subproblem can be thought to represents an individual agent, optimizing her private objective function. Established methods such as Benders decomposition ensure in this context, that the exchange of information between subproblems is sufficient to ensure convergence to a global optimum. Using a different information exchange scheme, one may attempt to replicate the limited coordination between different agents as it takes place in reality.

In this context, however, the significance of the results of such an algorithm is often unclear: For which input data can the algorithm be expected to converge at all? Can the result be described theoretically and does it represent an equilibrium as it could be observed in the real world? Which properties can a resulting solution be expected to have and how robust is it to choices in the precise implementation of the information exchange scheme?

On the other hand, if a suitable information exchange scheme can be developed and analyzed, such approaches may offer a natural way to deal with game-theoretic problems at a large scale. In such a setting, we might not be able to replicate the strong convergence properties of Benders decomposition, but it might be possible to at least obtain feasible solutions together with a corresponding lower bound, certifying some *a posteriori* quality guarantee for the solution.

Conclusion

A major challenge in the analysis and design of power systems consists in solving large-scale (transmission and generation) capacity expansion problems. Most immediately, these are used to devise economically optimal expansion paths from a given infrastructure while satisfying certain objectives. These objectives can range from satisfying a certain demand to limiting the consumption of certain resources or using a prescribed percentage of renewable energy sources.

But also beyond this immediate application, capacity expansion problems are useful to answer many questions that touch on the structure of a future energy system: Even if a research question does not immediately have anything to do with capacity expansion (e. g., investigating the economics of providing storage capacity as a business model, or the viability of a community-level microgrid), the question of the available infrastructure arises. In order to provide a meaningful answer, we must make some assumptions about the broader energy system in the context of which our study should take place.

While infrastructure decisions in reality are by no means (and arguably should not be) taken exclusively according to economic criteria, the *economically optimal* infrastructure often provides a useful baseline, especially in the absence of any other, more reliable source of information. This is especially true for questions about the state of an energy system in the far future, where assumptions drawn from current political priorities may no longer yield useful advice.

As we have argued in Chapter 1, the requirements imposed on capacity expansion problems used in this context become more and more challenging, partly as a consequence of changes that are currently underway in energy systems globally, but maybe particularly in Europe: Renewable energy sources force us to consider intermittent supply at a high temporal resolution, growing storage capacities make it essential to respect the exact order in which different load cases occur. As the locations of supply can no longer be chosen to be close to centers of demand (think of off-shore wind turbines and solar power plants in the desert), the constraints of the transmission grid are exacerbated. The coupling of different sectors in the energy system introduces a new level of interdependency.

In this context, our thesis provides several avenues to dealing with these challenges: Using simplified network models from Chapter 2, we can get an idea of the constraints imposed by the transmission grid, without resorting to the full complexity of optimization models required to accurately represent their behavior. The improvements to the decomposition framework *Benders decomposition* from Chapter 3 allow us to deal with larger optimization problems overall, particularly if they consist of weakly coupled segments, such as different sectors of the energy system or limited international (or inter-continental) transmission capacities. Finally, our results with respect to incentives (Chapter 4) give us an estimate of the additional complexity and the potential welfare losses introduced by selfish behavior of different agents in the energy system.

While we have already undertaken first steps towards the implementation of the techniques mentioned above together with our partners from *Chair of Renewable and Sustainable Energy Systems* at TUM, important challenges remain. Simplified transmission networks and decomposition may make it easier to imagine a model that encompasses all interesting aspects of Power System Optimization at the same time, enabling us to evaluate game-theoretical strategy equilibria in large-scale coupled power systems. However, much research still needs to be done in order to be able to compute meaningful solutions within any reasonable timeframe.

A different question that arises from methodological improvements in this context is that of their economical (or technological) interpretation: Do the mathematical structures exposed by optimization methods for problems in the area of power system analysis hold interesting insights into desirable structures of a future power system itself? This argument is being made, for instance, in the context of the transmission grid (see, e. g., [HU18]), but whether the resulting structures can be beneficial for the power system as a whole remains open for debate.

Appendix

A.1 Notation

We will find in many places that it is inconvenient to number the elements of a finite set S in order to allow us to map them to their corresponding entry of a vector $a \in \mathbb{R}^{|S|}$. In these cases, we use the notation $a \in \mathbb{R}^S$ to denote a vector a of dimension $|S|$, the components of which are indexed directly by the elements of the set S (rather than by natural numbers), i. e., if $a \in \mathbb{R}^S$ and $s \in S$, then $a_s \in \mathbb{R}$. This is in line with the notation T^S that is commonly used for the set of mappings from the set S into T. As S is finite, the vector space $\mathbb{R}^S$ is obviously isomorphic to the vector space $\mathbb{R}^{|S|}$ and our notation is hence equivalent to numbering the elements of $S = \{s_1, \ldots, s_{|S|}\}$ and denoting the entry of a which corresponds to s_i by a_i.

A.2 Graph Theory

Definition A.1 (graph)
Let V be a vertex set and $E' \subset \binom{V}{2}$ a set of undirected edges. We call $G' = (V, E')$ an *undirected graph*. With a set $E \subset V^2 \setminus \{(v, v) : v \in V\}$ of *directed* edges, we call $G = (V, E)$ a *directed graph*.

If G is a directed graph, we denote for every vertex $v \in V$ the *in-neighborhood* of v by $N^{\text{in}}(v) := \{w \in V \mid (w, v) \in E\}$ and the *out-neighborhood* by $N^{\text{out}}(v) := \{w \in V \mid (v, w) \in E\}$. The *neighborhood* of v is the union $N(v) := N^{\text{in}}(v) \cup N^{\text{out}}(v)$. Analogously, we define the set of incoming edges by $\delta^{\text{in}}(v) := \{(w, v) \in E \mid w \in V\}$, the set of outgoing edges by $\delta^{\text{out}}(v) := \{(v, w) \in E \mid w \in V\}$ and the set of all neighboring edges by $\delta(v) := \delta^{\text{in}}(v) \cup \delta_v^{\text{out}}$.

We call a directed graph *anti-symmetric* if for all $v, w \in V$, it holds that $(v, w) \in E \Rightarrow (w, v) \notin E$, i. e., there are no two edges that connect the same pair of vertices in opposite directions.

To simplify the notation, we define the *incidence matrix* A by

$$A = (a_{ve})_{\substack{v \in V \\ e \in E}} := \begin{cases} 1 & e \in \delta^{\text{in}}(v) \\ -1 & e \in \delta^{\text{out}}(v) \\ 0 & \text{otherwise.} \end{cases} \tag{A.1}$$

Definition A.2 (paths)
Let $G = (V, E)$ be a directed graph. We say that a set of edges $P \subset E$ is an undirected v-w-path if there exists $k \in \mathbb{N}$, $(v = v_1, v_2, \ldots, v_k = w) \in V^k$ and $\mathcal{I} \subset [k-1]$ such that $v_i \neq v_j$ for all $i, j \in [k]$ and $P = \{(v_i, v_{i+1}) \mid i \in \mathcal{I}\} \cup \{(v_{i+1}, v_i) \mid i \in [k-1] \setminus \mathcal{I}\}$. If $\mathcal{I} = [k-1]$, then P is a directed v_1-v_k-path (or simply a v_1-v_k-path) in G.

If $G = (V, E)$ is an undirected graph, then $P \subset E$ is a v-w-path if there exists $k \in \mathbb{N}$ and $(v = v_1, v_2, \ldots, v_k = w) \in V^k$ such that $v_i \neq v_j$ for all $i, j \in [k]$ and $P = \{\{v_i, v_{i+1}\} \mid i \in [k-1]\}$.

Definition A.3 (cycles)
Let $G = (V, E)$ be a directed graph. We say that a set of edges $C \subset E$ is a simple undirected cycle if there exists $k \in \mathbb{N}$, $(v_1, v_2, \ldots, v_k) \in V^k$ and $\mathcal{I} \subset [k]$ such that $v_i \neq v_j$ for all $i, j \in [k]$ and $C = \{(v_i, v_{i+1}) \mid i \in \mathcal{I}\} \cup \{(v_{i+1}, v_i) \mid i \in [k] \setminus \mathcal{I}\}$ (writing $v_{k+1} := v_1$). If $\mathcal{I} = [k]$, then P is a directed cycle (or simply a cycle) in G.

If $G = (V, E)$ is an undirected graph, then $C \subset E$ is a simple cycle if there exists $k \in \mathbb{N}$, $(v_1, v_2, \ldots, v_k) \in V^k$ such that $v_i \neq v_j$ for all $i, j \in [k]$ and $C = \{\{v_i, v_{i+1}\} \mid i \in [k]\}$ (writing $v_{k+1} := v_1$).

If there is no cycle in G, then we say that G is *acyclic*.

Definition A.4 (connectivity)
A (directed or undirected) graph $G = (V, E)$ is *connected*, if for every $v, w \in V$ there exists a v-w-path in G. A directed graph is *weakly connected*, if for every $v, w \in V$ there exists an undirected v-w-path in G.

Definition A.5 (subgraphs and components)
Let $G := (V, E)$ and $G' := (V', E')$ be two (directed) graphs. If $V' \subset V$ and $E' \subset E$, then we call (V', E') a *subgraph* of (V, E) and we write $v \in G'$ if and only if $v \in V'$. If $E' = \{(v, w) \in E \mid v, w \in V'\}$, then (V', E') is the subgraph *induced by* V'.

If G' is a connected subgraph of G and there exists no connected subgraph G'' of G such that $G'' \neq G'$ and G' is a subgraph of G'', then G' is a *connected component* of G.

If G' is a weakly connected subgraph of G and there exists no weakly connected subgraph $G''ofG$ such that $G'' \neq G'$ and G' is a subgraph of G'', then G' is a *weakly connected component* of G.

Definition A.6 (transitivity)
The directed graph $G = (V, E)$ is *transitive* if for all $(u, v), (v, w) \in E$, we have $(u, w) \in E$. The *transitive closure* of G is a graph $G' = (V, E')$ where $(u, v) \in E'$ if and only if there is a directed u-v-path in G.

A.3 Convex Geometry

Definition A.7 (convex set and lineality space)
A set $C \subset \mathbb{R}^n$ is *convex* if for all $x, y \in C$ and $\lambda \in [0,1]$, it holds that $\lambda x + (1-\lambda)y \in C$.

The set $\{y \in \mathbb{R}^n \mid x + \lambda y \in C \text{ for all } x \in C, \lambda \in \mathbb{R}\}$ is called the *lineality space* of C. If the lineality space of C is $\{0\}$, then we call C *line-free*.

Definition A.8 (halfspace and hyperplane)
Let $\pi \in \mathbb{R}^n \setminus \{0\}$, $\alpha \in \mathbb{R}$. We denote by

$$H^{\leq}_{(\pi,\alpha)} := \left\{ x \in \mathbb{R}^n \;\middle|\; \pi^\top x \leq \alpha \right\}$$

the *halfspace* induced by (π, α). Similarly, we write

$$H^{=}_{(\pi,\alpha)} := \left\{ x \in \mathbb{R}^n \;\middle|\; \pi^\top x = \alpha \right\}$$

for the corresponding *hyperplane* that is the boundary of $H^{\leq}_{(\pi,\alpha)}$.

Definition A.9 (supporting and separating hyperplanes and halfspaces)
Let $C \subset \mathbb{R}^n$ be a closed convex set.

a) The function $h_C(c) := \sup\{c^\top x \mid x \in C\}$ is called the *support function* of C.

b) A halfspace $H^{\leq}_{(\pi,\alpha)}$ *supports* C if $C \subset H^{\leq}_{(\pi,\alpha)}$ and $C \cap H^{=}_{(\pi,\alpha)} \neq \emptyset$.

c) A hyperplane $H^{=}_{(\pi,\alpha)}$ *strongly separates* C from another convex set $C' \subset \mathbb{R}^n$ if there exists $\varepsilon > 0$ such that

$$\pi^\top x \leq \alpha - \varepsilon \qquad \forall\, x \in C$$

and

$$\pi^\top x \geq \alpha + \varepsilon \qquad \forall\, x \in C'$$

or vice versa.

d) Let $x \in \mathbb{R}^n \setminus C$. A halfspace $H^{\leq}_{(\pi,\alpha)}$ is *x-separating* for C if $C \subset H^{\leq}_{(\pi,\alpha)}$ and $x \notin H^{\leq}_{(\pi,\alpha)}$.

Definition A.10 (cone)
Let C be a convex set such that for all $x \in C$ and $\lambda > 0$, it holds that $\lambda x \in C$. Then, C is called a *cone*. If furthermore for all $x \in C \setminus \{0\}$ it is true that $-x \notin C$, then C is called a *pointed cone*.

Definition A.11 (recession cone)
Let $C \subset \mathbb{R}^n$ be a convex set. The set

$$\mathrm{rec}(C) := \{y \in \mathbb{R}^n \mid C + y \subset C\}$$

is called the *recession cone* of C.

Definition A.12 (epigraph)
Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. The set

$$\mathrm{epi}(h) := \{(x, \eta) \mid \eta \geq h(x)\}$$

is called the *epigraph of* h. For a subset $C \subset \mathbb{R}^n$, we define

$$\mathrm{epi}_C(h) := \mathrm{epi}(h) \cap (C \times \mathbb{R}).$$

Definition A.13 (convex functions and subdifferentials)
Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. The function h is called *convex* if its epigraph $\mathrm{epi}(h)$ is a convex set.

Let $C \subset \mathbb{R}^n$ be a convex set and $h : C \to \mathbb{R} \cup \{\pm\infty\}$ a convex function. Let $x^* \in C$. The set

$$\partial h(x^*) := \left\{ a \in \mathbb{R}^n \;\middle|\; h(x^*) + a^\top (x - x^*) \leq h(x) \forall\ x \in C \right\}$$

is called the *subdifferential of* h *in* x^*. A vector $a \in \partial h(x^*)$ is called *subgradient of* h *in* x^*.

Index

Bibliography

[Abe+16] A. Abeliuk, H. Aziz, G. Berbeglia, S. Gaspers, P. Kalina, N. Mattei, D. Peters, P. Stursberg, P. Van Hentenryck, and T. Walsh. "Interdependent scheduling games". In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 2016, pp. 2–9.

[ABV15] A. Abeliuk, G. Berbeglia, and P. Van Hentenryck. "Bargaining mechanisms for one-way games". In: *Games* 6.3 (2015), pp. 347–367.

[Agn+15] A. Agnetis, C. Briand, J.-C. Billaut, and P. Šůcha. "Nash equilibria for the multi-agent project scheduling problem with controllable processing times". In: *Journal of Scheduling* 18.1 (2015), pp. 15–27.

[AS13] P. Ahlhaus and P. Stursberg. "Transmission capacity expansion: An improved Transport Model". In: *4th IEEE/PES Innovative Smart Grid Technologies Europe*. 2013, pp. 1–5.

[AMO93] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, 1993.

[AC00] N. Alguacil and A. J. Conejo. "Multiperiod optimal power flow using Benders decomposition". In: *IEEE Transactions on Power Systems* 15.1 (2000), pp. 196–201.

[AMC03] N. Alguacil, A. L. Motto, and A. J. Conejo. "Transmission expansion planning: A mixed-integer LP approach". In: *IEEE Transactions on Power Systems* 18.3 (2003), pp. 1070–1077.

[ABP06] E. Angel, E. Bampis, and F. Pascual. "Truthful algorithms for scheduling selfish tasks on parallel machines". In: *Theoretical Computer Science* 369.1-3 (2006), pp. 157–168.

[Aza+13] N. Azad, G. K. D. Saharidis, H. Davoudpour, H. Malekly, and S. A. Yektamaram. "Strategies for protecting supply chain networks against facility and transportation disruptions: an improved Benders decomposition approach". In: *Annals of Operations Research* 210.1 (2013), pp. 125–163.

[Bah+01] L. Bahiense, G. C. Oliveira, M. V. F. Pereira, and S. Granville. "A mixed integer disjunctive model for transmission network expansion". In: *IEEE Transactions on Power Systems* 16.3 (2001), pp. 560–565.

[Bai+08] X. Bai, H. Wei, K. Fujisawa, and Y. Wang. "Semidefinite programming for optimal power flow problems". In: *International Journal of Electrical Power & Energy Systems* 30.6 (2008), pp. 383–392.

[Bal75] E. Balas. "Facets of the knapsack polytope". In: *Mathematical Programming* 8.1 (1975), pp. 146–164.

[Bal79] E. Balas. "Disjunctive programming". In: *Annals of Discrete Mathematics* 5 (1979), pp. 3–51.

[Bal98] E. Balas. "Disjunctive programming: Properties of the convex hull of feasible points". In: *Discrete Applied Mathematics* 89.1 (1998), pp. 3–44.

[BI64] E. Balas and P. L. Ivanescu. "On the generalized transportation problem". In: *Management Science* 11.1 (1964), pp. 188–202.

[Bap14] R. B. Bapat. *Graphs and matrices*. 2nd ed. Universitext. Springer, 2014.

[Ben62] J. F. Benders. "Partitioning procedures for solving mixed-variables programming problems". In: *Numerische Mathematik* 4.1 (1962), pp. 238–252.

[BGdV04] F. Berger, P. Gritzmann, and S. de Vries. "Minimum cycle bases for network graphs". In: *Algorithmica* 40.1 (2004), pp. 51–62.

[BPG01] S. Binato, M. V. F. Pereira, and S. Granville. "A new Benders decomposition approach to solve power transmission network design problems". In: *IEEE Transactions on Power Systems* 16.2 (2001), pp. 235–240.

[BL88] J. R. Birge and F. V. Louveaux. "A multicut algorithm for two-stage stochastic linear programs". In: *European Journal of Operational Research* 34.3 (1988), pp. 384–392.

[Blo83] J. A. Bloom. "Solving an electricity generating capacity expansion planning problem by generalized Benders' decomposition". In: *Operations Research* 31.1 (1983), pp. 84–100.

[Bos+11] S. Bose, D. F. Gayme, S. Low, and K. M. Chandy. "Optimal power flow over tree networks". In: *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*. 2011, pp. 1342–1348.

[BS19a] R. Brandenberg and P. Stursberg. "A generalization of spanning trees for network flows with differential constraints". In preparation. 2019.

[BS19b] R. Brandenberg and P. Stursberg. "Cut selection for Benders decomposition". In preparation. 2019.

[BLS99] A. Brandstädt, V. B. Le, and J. P. Spinrad. *Graph classes: a survey*. Vol. 3. Discrete Mathematics and Applications. Siam, 1999.

[BB11] C. Briand and J. Billaut. “Cooperative project scheduling with controllable processing times: a game theory framework”. In: *Proceedings of the 16th IEEE Conference on Emerging Technologies & Factory Automation*. 2011, pp. 1–7.

[BJK97] P. Brucker, B. Jurisch, and A. Krämer. “Complexity of scheduling problems with multi-purpose machines”. In: *Annals of Operations Research* 70 (1997), pp. 57–73.

[Chr+11] P. Christiano, J. A. Kelner, A. Madry, D. A. Spielman, and S.-H. Teng. “Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs”. In: *Proceedings of the forty-third annual ACM symposium on Theory of computing*. 2011, pp. 273–282.

[CKN09] G. Christodoulou, E. Koutsoupias, and A. Nanavati. “Coordination mechanisms”. In: *Theoretical Computer Science* 410.36 (2009), pp. 3327–3336.

[Chu97] F. R. Chung. *Spectral graph theory*. Vol. 92. CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997.

[CV14] C. Coffrin and P. Van Hentenryck. “A linear-programming approximation of AC power flows”. In: *INFORMS Journal on Computing* 26.4 (2014), pp. 718–734.

[Con+06] A. J. Conejo, E. Castillo, R. Minguez, and R. García-Bertrand. *Decomposition techniques in mathematical programming: engineering and science applications*. Springer, 2006.

[CW18] M. Conforti and L. A. Wolsey. “‘Facet’ separation with one linear program”. In: *Mathematical Programming, Series A* (2018), pp. 1–20.

[Coo+98] W. Cook, W. Cunningham, W. Pullyblank, and A. Schrijver. *Combinatorial Optimization. Series in Discrete Mathematics and Optimization*. Wiley-Interscience, 1998.

[CL06] G. Cornuéjols and C. Lemaréchal. “A convex-analysis perspective on disjunctive cuts”. In: *Mathematical Programming, Series A* 106.3 (2006), pp. 567–586.

[DLM10] C. D’Ambrosio, A. Lodi, and S. Martello. “Piecewise linear approximation of functions of two variables in MILP models”. In: *Operations Research Letters* 38.1 (2010), pp. 39–46.

[DW61] G. B. Dantzig and P. Wolfe. “The decomposition algorithm for linear programs”. In: *Econometrica: Journal of the Econometric Society* 29.4 (1961), pp. 767–778.

[Die17] R. Diestel. *Graph theory*. 5th ed. Graduate Texts in Mathematics. Springer, 2017.

[EU09] "Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC". In: *Official Journal of the European Union* (2009).

[Duf65] R. J. Duffin. "Topology of series-parallel networks". In: *Journal of Mathematical Analysis and Applications* 10.2 (1965), pp. 303–318.

[Ege16] J. Egerer. *Open source electricity model for Germany (ELMOD-DE)*. Tech. rep. Deutsches Institut für Wirtschaftsforschung, 2016.

[Ege+14] J. Egerer, C. Gerbaulet, R. Ihlenburg, F. Kunz, B. Reinhard, C. von Hirschhausen, A. Weber, and J. Weibezahn. *Electricity sector data for policy-relevant modeling: Data documentation and applications to the german and european electricity markets*. Tech. rep. Data Documentation, DIW, 2014.

[FSZ09] M. Fischetti, D. Salvagnin, and A. Zanette. "Minimal infeasible subsystems and Benders cuts". `http://www.math.uwaterloo.ca/~bico/bellairs/Benders.pdf`. 2009.

[FSZ10] M. Fischetti, D. Salvagnin, and A. Zanette. "A note on the selection of Benders cuts". In: *Mathematical Programming, Series B* 124.1-2 (2010), pp. 175–182.

[Fri12] M. Fripp. "Switch: a planning tool for power systems with large shares of intermittent renewable energy". In: *Environmental Science & Technology* 46.11 (2012), pp. 6371–6378.

[FSL05] Y. Fu, M. Shahidehpour, and Z. Li. "Security-constrained unit commitment with AC constraints". In: *IEEE Transactions on Power Systems* 20.2 (2005), pp. 1001–1013.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and intractability*. W. H. Freeman and Company, 1979.

[Gar70] L. L. Garver. "Transmission network estimation using linear programming". In: *IEEE Transactions on Power Apparatus and Systems* 7 (1970), pp. 1688–1697.

[Geo70] A. M. Geoffrion. "Elements of large-scale mathematical programming Part I: Concepts". In: *Management Science* 16.11 (1970), pp. 652–675.

[Geo72] A. M. Geoffrion. "Generalized Benders decomposition". In: *Journal of optimization theory and applications* 10.4 (1972), pp. 237–260.

[Geo74] A. M. Geoffrion. "Lagrangean relaxation for integer programming". In: *Approaches to Integer Programming* 2 (1974), pp. 82–114.

[GR90] J. Gleeson and J. Ryan. “Identifying minimally infeasible subsystems of inequalities”. In: *ORSA Journal on Computing* 2.1 (1990), pp. 61–63.

[GT89] A. V. Goldberg and R. E. Tarjan. “Finding minimum-cost circulations by canceling negative cycles”. In: *Journal of the ACM* 36.4 (1989), pp. 873–886.

[GLM99] A. Grothey, S. Leyffer, and K. McKinnon. “A note on feasibility in Benders decomposition”. In: *Numerical Analysis Report NA/188, Dundee University* (1999).

[GLS12] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*. Vol. 2. Springer, 2012.

[GMD79] R. Gutman, P. Marchenko, and R. Dunlop. “Analytical development of loadability characteristics for EHV and UHV transmission lines”. In: *IEEE Transactions on Power Apparatus and Systems* 2 (1979), pp. 606–617.

[HC94] B. Hamann and J.-L. Chen. “Data point selection for piecewise linear curve approximation”. In: *Computer Aided Geometric Design* 11.3 (1994), pp. 289–301.

[HK86] W. H. Hayt and J. E. Kemmerly. *Engineering circuit analysis*. 4th ed. McGraw-Hill series in electrical engineering. McGraw-Hill, 1986.

[Hew+16] D. Hewes, S. Altschaeffl, I. Boiarchuk, and R. Witzmann. “Development of a dynamic model of the European transmission system using publicly available data”. In: *2016 IEEE International Energy Conference*. 2016, pp. 1–6.

[HMU07] B. Heydenreich, R. Müller, and M. Uetz. “Games and mechanism design in machine scheduling – an introduction”. In: *Production and Operations Management* 16.4 (2007), pp. 437–454.

[HO03] J. N. Hooker and G. Ottosson. “Logic-based Benders decomposition”. In: *Mathematical Programming, Series A* 96.1 (2003), pp. 33–60.

[HU18] M. Hotz and W. Utschick. “The hybrid transmission grid architecture: Benefits in nodal pricing”. In: *IEEE Transactions on Power Systems* 33.2 (2018), pp. 1431–1442.

[IEEE79] “IEEE Reliability Test System”. In: *IEEE Transactions on Power Apparatus and Systems* PAS-98.6 (1979), pp. 2047–2054.

[Ili92] M. Ilić. “Network theoretic conditions for existence and uniqueness of steady state solutions to electric power circuits”. In: *Proceedings of the 1992 IEEE International Symposium on Circuits and Systems*. Vol. 6. 1992, pp. 2821–2828.

[Jab06] R. A. Jabr. "Radial distribution load flow using conic programming". In: *IEEE transactions on power systems* 21.3 (2006), pp. 1458–1459.

[Jen+15] M. Jenabi, S. M. T. Fatemi Ghomi, S. A. Torabi, and S. H. Hosseinian. "Acceleration strategies of Benders decomposition for the security constraints power system expansion planning". In: *Annals of Operations Research* 235.1 (2015), pp. 337–369.

[Kar72] R. M. Karp. "Reducibility among combinatorial problems". In: *Complexity of computer computations*. Ed. by R. E. Miller and J. W. Thatcher. Plenum Press, 1972, pp. 85–103.

[Kav+04] T. Kavitha, K. Mehlhorn, D. Michail, and K. Paluch. "A faster algorithm for minimum cycle basis of graphs". In: *International Colloquium on Automata, Languages, and Programming*. 2004, pp. 846–857.

[Kha+08] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich. "Generating all vertices of a polyhedron Is hard". In: *Discrete & Computational Geometry* 39.1 (2008), pp. 174–190.

[KNK11] F. Kießling, P. Nefzger, and U. Kaintzyk. *Freileitungen: Planung, Berechnung, Ausführung*. Springer, 2011.

[Kir47] G. Kirchhoff. "Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird". In: *Annalen der Physik und Chemie*. Ed. by J. C. Poggendorfer. Vol. 72. Verlag von Johann Ambrosius Barth, 1847.

[Kni72] U. G. Knight. *Power systems engineering and mathematics*. 1st ed. Pergamon Press, 1972.

[KKM94] R. Kohli, R. Krishnamurti, and P. Mirchandani. "The Minimum Satisfiability Problem". In: *SIAM Journal on Discrete Mathematics* 7.2 (1994), pp. 275–283.

[KV08] B. Korte and J. Vygen. *Combinatorial optimization*. 4th ed. Springer, 2008.

[Kou14] E. Koutsoupias. "Scheduling without payments". In: *Theory of Computing Systems* 54.3 (2014), pp. 375–387.

[Lat+03] G. Latorre, R. D. Cruz, J. M. Areiza, and A. Villegas. "Classification of publications and models on transmission expansion planning". In: *IEEE Transactions on Power Systems* 18.2 (2003), pp. 938–946.

[LL12] J. Lavaei and S. H. Low. "Zero duality gap in optimal power flow problem". In: *IEEE Transactions on Power Systems* 27.1 (2012), pp. 92–107.

[LTZ14] J. Lavaei, D. Tse, and B. Zhang. "Geometry of power flows and optimization in distribution networks". In: *IEEE Transactions on Power Systems* 29.2 (2014), pp. 572–583.

[Law64] E. L. Lawler. “On scheduling problems with deferral costs”. In: *Management Science* 11.2 (1964), pp. 280–288.

[LF92] F. N. Lee and Q. Feng. “Multi-area unit commitment”. In: *IEEE Transactions on Power Systems* 7.2 (1992), pp. 591–599.

[Lei+15] T. Leibfried, T. Mchedlidze, N. Meyer-Hübner, M. Nöllenburg, I. Rutter, P. Sanders, D. Wagner, and F. Wegner. “Operating power grids with few flow control buses”. In: *Proceedings of the 2015 ACM Sixth International Conference on Future Energy Systems*. 2015, pp. 289–294.

[LR78] J. K. Lenstra and A. H. G. Rinnooy Kan. “Complexity of scheduling under precedence constraints”. In: *Operations Research* 26.1 (1978), pp. 22–35.

[LRB77] J. K. Lenstra, A. H. G. Rinnooy Kan, and P. Brucker. “Complexity of machine scheduling problems”. In: *Annals of Discrete Mathematics* 1 (1977), pp. 343–362.

[LH05] B. C. Lesieutre and I. Hiskens. “Convexity of the set of feasible injections and revenue adequacy in FTR markets”. In: *IEEE Transactions on Power Systems* 20.4 (2005), pp. 1790–1798.

[Low14a] S. 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.

[Low14b] S. 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.

[LR16] S. Lumbreras and A. Ramos. “How to solve the transmission expansion planning problem faster: acceleration techniques applied to Benders decomposition”. In: *IET Generation, Transmission & Distribution* 10.10 (2016), pp. 2351–2359.

[MSL15] R. Madani, S. Sojoudi, and J. Lavaei. “Convex relaxation for optimal power flow problem: Mesh networks”. In: *IEEE Transactions on Power Systems* 30.1 (2015), pp. 199–211.

[MW81] T. L. Magnanti and R. T. Wong. “Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria”. In: *Operations Research* 29.3 (1981), pp. 464–484.

[EC88] E. S. El-Mallah and C. J. Colbourn. “The complexity of some edge deletion problems”. In: *IEEE Transactions on Circuits and Systems* 35.3 (1988), pp. 354–362.

[Nas51] J. Nash. “Non-cooperative games”. In: *Annals of Mathematics* (1951), pp. 286–295.

[NW88] G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience, 1988.

[NN94] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM Studies in Applied and Numerical Mathematics. Society for Industrial and Applied Mathematics, 1994.

[Nis+07] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, eds. *Algorithmic game theory*. Cambridge University Press, 2007.

[Pap08] N. Papadakos. "Practical enhancements to the Magnanti–Wong method". In: *Operations Research Letters* 36.4 (2008), pp. 444–449.

[PS98] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: Algorithms and complexity*. Courier Corporation, 1998.

[PP91] M. V. F. Pereira and L. M. V. G. Pinto. "Multi-stage stochastic optimization applied to energy planning". In: *Mathematical Programming* 52.1 (1991), pp. 359–375.

[PG08] A. Philpott and Z. Guan. "On the convergence of stochastic dual dynamic programming and related methods". In: *Operations Research Letters* 36.4 (2008), pp. 450–455.

[Pin16] M. L. Pinedo. *Scheduling: Theory, algorithms, and systems*. 5th ed. Springer, 2016.

[Rah+17] R. Rahmaniani, T. G. Crainic, M. Gendreau, and W. Rei. "The Benders decomposition algorithm: A literature review". In: *European Journal of Operational Research* 259.3 (2017), pp. 801–817.

[Reb16] S. Rebennack. "Combining sampling-based and scenario-based nested Benders decomposition methods: application to stochastic dual dynamic programming". In: *Mathematical Programming, Series A* 156.1 (2016), pp. 343–389.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Roc84] R. T. Rockafellar. *Network flows and monotropic optimization*. Pure and Applied Mathematics. Wiley-Interscience, 1984.

[RM94] R. Romero and A. Monticelli. "A hierarchical decomposition approach for transmission network expansion planning". In: *IEEE Transactions on Power Systems* 9.1 (1994), pp. 373–380.

[Rom+02] R. Romero, A. Monticelli, A. Garcia, and S. Haffner. "Test systems and mathematical models for transmission network expansion planning". In: *IEE Proceedings – Generation, Transmission and Distribution*. Vol. 149. 2002, pp. 27–36.

[SMI10] G. K. D. Saharidis, M. Minoux, and M. G. Ierapetritou. “Accelerating Benders method using covering cut bundle generation”. In: *International Transactions in Operational Research* 17.2 (2010), pp. 221–237.

[SO09] E. E. Sauma and S. S. Oren. “Do generation firms in restructured electricity markets have incentives to support social-welfare-improving transmission investments?” In: *Energy Economics* 31.5 (2009), pp. 676–689.

[SSH12] K. Schaber, F. Steinke, and T. Hamacher. “Transmission grid extensions for the integration of variable renewable energies in Europe: Who benefits where?” In: *Energy Policy* 43 (2012), pp. 123–135.

[Sch03] A. Schrijver. *Combinatorial optimization*. Algorithms and Combinatorics. Springer, 2003.

[SF05] M. Shahidehpour and Y. Fu. “Benders decomposition: applying Benders decomposition to power systems”. In: *IEEE Power and Energy Magazine* 3.2 (2005), pp. 20–21.

[Sha11] A. Shapiro. “Analysis of stochastic dual dynamic programming method”. In: *European Journal of Operational Research* 209.1 (2011), pp. 63–72.

[SL13] H. D. Sherali and B. J. Lunday. “On generating maximal nondominated Benders cuts”. In: *Annals of Operations Research* 210.1 (2013), pp. 57–72.

[SDT14] C. Skar, G. Doorman, and A. Tomasgard. “Large-scale power system planning using enhanced Benders decomposition”. In: *Proceedings of the 18th Power System Computations Conference*. 2014.

[SL12] S. Sojoudi and J. Lavaei. *Network topologies guaranteeing zero duality gap for optimal power flow problem*. Tech. rep. University of California, Berkeley, 2012.

[SL14] S. Sojoudi and J. Lavaei. “Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure”. In: *SIAM Journal on Optimization* 24.4 (2014), pp. 1746–1778.

[StC53] H. P. St. Clair. “Practical concepts in capability and performance of transmission lines”. In: *Transactions of the AIEE. Part III: Power Apparatus and Systems* 72.6 (1953), pp. 1152–1157.

[SJA09] B. Stott, J. Jardim, and O. Alsaç. “DC power flow revisited”. In: *IEEE Transactions on Power Systems* 24.3 (2009), pp. 1290–1300.

[TCL15] C. W. Tan, D. W. Cai, and X. Lou. “Resistive network optimal power flow: uniqueness and algorithms”. In: *IEEE Transactions on Power Systems* 30.1 (2015), pp. 263–273.

[TJS13] L. Tang, W. Jiang, and G. K. D. Saharidis. "An improved Benders decomposition algorithm for the logistics facility location problem with capacity expansions". In: *Annals of Operations Research* 210.1 (2013), pp. 165–190.

[Tru77] K. Truemper. "On max flows with gains and pure min-cost flows". In: *SIAM Journal on Applied Mathematics* 32.2 (1977), pp. 450–456.

[Tru78] K. Truemper. "Optimal flows in nonlinear gain networks". In: *Networks* 8.1 (1978), pp. 17–36.

[Tuo+09] A. Tuohy, P. Meibom, E. Denny, and M. O'Malley. "Unit commitment for systems with significant wind penetration". In: *IEEE Transactions on Power Systems* 24.2 (2009), pp. 592–601.

[VW10] F. Vanderbeck and L. A. Wolsey. "Reformulation and decomposition of integer programs". In: *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. Ed. by M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey. Springer, 2010, pp. 431–502.

[VGS85] R. Villasana, L. L. Garver, and S. J. Salon. "Transmission network planning using linear programming". In: *IEEE Transactions on Power Apparatus and Systems* 104 (1985), pp. 349–356.

[VM44] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

[Wan+16] Q. Wang, J. D. McCalley, T. Zheng, and E. Litvinov. "Solving corrective risk-based security-constrained optimal power flow with Lagrangian relaxation and Benders decomposition". In: *International Journal of Electrical Power & Energy Systems* 75 (2016), pp. 255–264.

[Whe43] C. Wheatstone. "The Bakerian Lecture: an account of several new instruments and processes for determining the constants of a voltaic circuit". In: *Philosophical Transactions of the Royal Society of London* 133 (1843), pp. 303–327.

[Wid18] P. Widmann. "Vergleich von DC- und TR-Modellen zur Stromnetzwerkoptimierung". Bachelor's Thesis. 2018.

[Wol81] L. A. Wolsey. "A resource decomposition algorithm for general mathematical programs". In: *Mathematical Programming at Oberwolfach*. Springer, 1981, pp. 244–257.

[WW84] A. J. Wood and B. F. Wollenberg. *Power generation, operation, and control*. John Wiley & Sons, 1984.

[ZT13] B. Zhang and D. Tse. "Geometry of injection regions of power networks". In: *IEEE Transactions on Power Systems* 28.2 (2013), pp. 788–797.

[Zha+12] H. Zhang, V. Vittal, G. T. Heydt, and J. Quintero. “A mixed-integer linear programming approach for multi-stage security-constrained transmission expansion planning”. In: *IEEE Transactions on Power Systems* 27.2 (2012), pp. 1125–1133.