"Analysis of renewable energy integration in transmission-constrained electricity markets using parallel computing"

Aravena Solís, Ignacio Andrés

ABSTRACT

Renewable energy integration has progressively challenged the operating assumptions of electricity markets, questioning at first the deterministic nature of day-ahead and other forward electricity markets. As the integration of renewable resources became deeper, transmission grids began to experience congestion in unforeseen, dynamic and uncertain patterns, questioning now the management of transmission constraints in electricity markets. This dissertation proposes detailed models and algorithms for analyzing the impacts of uncertain renewable supply on different transmission-constrained electricity market designs, using parallel computing extensively in order to solve large-scale mathematical programs and to carry out simulations under different operating conditions for realistic systems. The contributions of the dissertation are organized in three chapters. Chapter 2 presents an asynchronous distributed algorithm for solving the stochastic unit commitment problem. The algorithm, deployed in parallel, is able to solve realistic instances of stochastic unit commitment within operationally acceptable tolerances and solution times. Chapter 3 proposes a consistent framework for modelling different zonal electricity markets. We develop cutting-plane algorithms to incorporate the N-1 criterion into this framework and we conduct simulations using single-period models for the Central Western European system under 768 000 different operating conditions. In chapter 4, we propose a hierarchy of mathematical programs that model the European zonal electricity market organization in d...

CITE THIS VERSION

Université Catholique de Louvain
Ecole Polytechnique de Louvain

Center for Operations Research and Econometrics

Analysis of renewable energy integration in transmission-constrained electricity markets using parallel computing

Ignacio Andrés Aravena Solís

Doctoral Thesis

Supervisor:
Anthony Papavasiliou
(UCL, Belgium)

Jury:
Philippe Chevalier (UCL, Belgium)
Guglielmo Lulli (Lancaster University, UK)
Yves Smeers (UCL, Belgium)
Mathieu Van Vyve (UCL, Belgium)

President:
Philippe Chevalier (UCL, Belgium)

PhD Organization

Ignacio Andrés Aravena Solís
Université catholique de Louvain
École Politechnique de Louvain
Center of Operations Research and Econometrics

Thesis Supervisor

Anthony Papavasiliou
Associate Professor, Université catholique de Louvain
École Polytechnique de Louvain
Center of Operations Research and Econometrics

Supervisory Committee

Yves Smeers
Emeritus Professor, Université catholique de Louvain
Center of Operations Research and Econometrics

Mathieu Van Vyve
Professor, Université catholique de Louvain
Louvain School of Management
Center of Operations Research and Econometrics

REPUESTA DEL ORÁCULO: Hagas lo que hagas te arrepentirás.

by Nicanor Parra

Abstract

Renewable energy integration has progressively challenged the operating assumptions of electricity markets, questioning at first the deterministic nature of day-ahead and other forward electricity markets. As the integration of renewable resources became deeper, transmission grids began to experience congestion in unforeseen, dynamic and uncertain patterns, questioning now the management of transmission constraints in electricity markets. This dissertation proposes detailed models and algorithms for analyzing the impacts of uncertain renewable supply on different transmission-constrained electricity market designs, using parallel computing extensively in order to solve large-scale mathematical programs and to carry out simulations under different operating conditions for realistic systems.

The contributions of the dissertation are organized in three chapters. Chapter 2 presents an asynchronous distributed algorithm for solving the stochastic unit commitment problem. The algorithm, deployed in parallel, is able to solve realistic instances of stochastic unit commitment within operationally acceptable tolerances and solution times. Chapter 3 proposes a consistent framework for modelling different zonal electricity markets. We develop cutting-plane algorithms to incorporate the N-1 criterion into this framework and we conduct simulations using single-period models for the Central Western European system under 768 000 different operating conditions. In chapter 4, we propose a hierarchy of mathematical programs that model the European zonal electricity market organization in detail in a multi-period setting. We compare the performance of this organization to the performances of a nodal market and the stochastic unit commitment model on the Central Western European system. Numerical results in chapters 3 and 4 indicate that failing to account for the limitations of the transmission grid, in a regime of large-scale renewable energy integration, can undermine system performance to much a larger extent than failing to account for uncertainty in nodal electricity markets.

Acknowledgements

I must admit that takes me some serious introspection to believe that I am finally writing this section. Tracing back the steps that took me to this point and revisiting all the points at which the slightest change would have led to me towards a different path is mind-blowing. Yet, here – it seems – I am, and it is actually a pleasure to write this section, which will be certainly the most read part of my Thesis.

First, I would like to extend my deepest gratitude to my advisor Pr. Anthony Papavasiliou, for his dedication, patience, commitment and rigor over these 4.5 years. Although it was difficult in the beginning, gradually he convinced me that the key to good quality work was to face it straight to its harshest criticism and improve it block-by-block. Having learnt this is already paying-off and there is no way for me to thank him enough for having the resolve to play a kind devil's advocate on each of our multiple meetings. At the same time, he showed me that greek people can be very fun and made my PhD experience an outlier among students. Thank you very much for your support and encouragement, there is no doubt in my mind that without you this work would not have been possible.

I would also like to extend my gratitude to Pr. Yves Smeers for his constant attention to my work and his insightful comments, that very often surprised me in my ignorance, and to Pr. Mathieu Van Vyve for his help in understanding the operations of the European electricity market. Thank you also to Pr. Philippe Chevalier and Pr. Guglielmo Lulli for taking part of the Jury and providing feedback on the original draft of this Thesis.

I thank Pr. Shmuel Oren (UC Berkeley) and Dr. Deepak Rajan (Lawrence Livermore National Laboratory) for the opportunity of visiting/intern with them in the summers of 2015 and 2017, and for the fun work we did in those occasions.

I am very grateful, as well, to my wife Lucy for all her support and comprehension during all these years we spent apart. Thank you for taking care of all the many things I ignored in order to concentrate on the PhD and for always bringing happiness (and *llamas*) to my life.

To my friends at CORE: thank you for making of everyday a freaky and nerdy culinary adventure at Sablon or avoiding Sablon. I could not have survived these years without the teas with Sinem; beers with Aditi, Stefan and Robert; long-standing debates with Mery (yes, I mentioned you!), Guz, Manuel, Jonas and Andras; Italian weekends with Manuela and Elissa; *cumbia* afternoons with Abdel and Mikel; deep cultural interchanges with Yuting and Huan; tender conversations with Claudio, Sabina, Valeria and Erika; strange conversations with Dagmara; and all the people at CORE that made my days happier. I would like to make an special mention to Catherine who cared for me almost as a baby sitter from my first day at CORE until nowadays. From outside

CORE, I must thank my only belgian friend Esther for making of our gatherings a surrealistic experience.

To my parents Juany and Ignacio, thank you for guiding me into this path and all the lessons you taught me throughout it. To my siblings Pame, Naty and Juan: we are all finally done with studying! (right?)

Finally, I would like to give special thanks to the Lawrence Livermore National Laboratory and to the Consortium des Équipements de Calcul Intensif for granting us computing time at their high-performance computing clusters, which we use extensively to conduct numerical experiments throughout this Thesis.

This work was supported by UC Louvain through FSR Grant and by the ENGIE Chair in Energy Economics & Energy Risk Management.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Governments and regulatory agencies worldwide have set forth a series of policies, over the past decade, towards reducing CO_2 and other greenhouse gases emissions. These policies have led to the integration of more than 129GW of solar and wind capacity in the US power grid, 277GW in Europe and 295GW in China, with other countries following their lead [Int18]. These renewable integration levels will continue to increase for most power grids following the Paris Agreement of 2015, entering into force in 2020 [Uni16].

Renewable resources pose several challenges to power system operations. Solar and wind power supply can vary significantly between forecast and real time operation, making it critical to determine adequate reserve levels in order to operate power systems reliably. This have motivated the industry and academic community to investigate efficient approaches for scheduling reserves in power systems under uncertainty (see [TvAFL15] and references therein for a survey). At the same time, the fast pace at which renewable resources have been integrated into power systems has left little time for the transmission grids to adapt to new flow patterns. As a consequence, congestion resulting from the new distribution of installed capacity and the uncontrollable fluctuations of renewable energy resources have lead to wind curtailment in the US (e.g. in ERCOT, [HZPA15]), Germany [Mor18] and China [HZPA15].

Coping with these challenges has been especially difficult in Europe due to (i) the zonal design of its energy market, which suffers from significant unscheduled flows as evidenced by the externalities of renewable power integration in Germany on neighboring networks (e.g. Poland and other countries [CPMS13]), and to (ii) the decentralized paradigm used to operate the system in real time. European Transmission System Operators (TSOs) have identified these problems and are currently undergoing efforts to develop coordinated congestion

management measures [Umb16] and to harmonize the definition and management of reserves in Europe [HAE+14]. In the same fashion, Independent System Operators (ISOs) and Regional Transmission Organizations (RTOs) in the US have recognized the need for developing cooperation strategies among balancing authorities in wide-area interconnections [MDE+10]. Ultimately, the introduction of renewable resources is inducing spatial and temporal coordination requirements on system operators.

This dissertation seeks to analyze the impacts of uncertain renewable supply on different electricity market designs under congested conditions. We focus on the European electricity market, where renewable integration is one of the highest and where the impacts of renewable integration are exacerbated by the discrepancies between the day-ahead and real-time representations of the physical system in the electricity market. Our approach consists in developing detailed models that avoid simplifying assumptions often made in economic analyses and that aim at being as realistic as possible in their encompassing of current and future legal obligations of the European electricity market[1]. The same level of detail is used for ideal market models serving as benchmarks in our analytics, which we conduct using realistic instances in order to obtain meaningful performance metrics. The price to pay for this lack of simplification is computational both in terms of algorithmic tools and machine resources: we develop practically-efficient algorithms for solving the mathematical programs on which state-of-the-art techniques fail, and we deploy these algorithms and carry out simulations using parallel computing infrastructure.

The present chapter recollects relevant previous results that lay the foundations for this dissertation. Section 1.2 introduces the basis of the algorithmic techniques developed in this dissertation. Section 1.3 presents models for transmission-constrained electricity markets in their simplest form.Section 1.4 presents a brief overview of the peculiarities of European electricity markets and describes the instance of the Central Western European system (CWE) we use as case study. Section 1.5 concludes this chapter by outlining the structure the rest of the dissertation.

1.2 Preliminaries on optimization

Optimization problems are a powerful tool for studying and simulating electricity markets. In this dissertation, we model electricity markets as optimization problems of the following form:

[1]Models for European electricity markets can only be idiosyncratic because of the peculiar institutions within it. However, with its coverage of 989 GW [Eur17b] and about 500 million consumers [Eur18] the European market is significantly larger than all the restructured US markets (CAISO 60 GW, ERCOT 75GW) and also than the largest US RTOs like PJM (166 GW) and MISO (174 GW) [Com18].

$$p^* = \max_{x \in X} f(\boldsymbol{x}) \tag{1.1}$$

$$\text{s.t.} \quad A\boldsymbol{x} \leq \boldsymbol{b} \qquad [\boldsymbol{\lambda}], \tag{1.2}$$

where $\boldsymbol{x} \in X \subset \mathbb{R}^n$ correspond to variables, f is a continuous concave objective function, $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ are parameters, $\boldsymbol{\lambda} \in \mathbb{R}^m$ are Lagrange multipliers and p^* is the optimal value. X is a bounded set that might contain binary constraints, that is $x_i \in \{0, 1\}\ \forall i \in I \subseteq \{1, \ldots, n\}$. Throughout this dissertation, unless indicated otherwise, we use lowercase for denoting scalars, bold lowercase for vectors and uppercase for matrices and sets.

While there exist general purpose techniques for solving problems of the type (1.1) – (1.2), e.g. branch-and-bound or branch-and-cut for linear and quadratic objectives, these techniques may fail to solve problems directly whenever n, m or both are very large, nowadays, in the order of millions. Optimization problems arising from electricity markets under uncertainty typically surpass the scale where general purpose techniques are effective [TvAFL15] and, therefore, require the design of specialized decomposition techniques that exploit the structure of the problem at hand to achieve a solution. In this section, we introduce the basis for the decomposition techniques developed in this dissertation in a generic context, including Lagrangian duality [BV04, Chapter 5], the subgradient [Nes04] and stochastic subgradient methods [Erm83], the cutting-plane method [Nes04] and the column-generation method [DW60].

1.2.1 Lagrangian duality

Lagrangian duality allows to relax constraints by moving them into the objective function, weighted by the Lagrange multipliers $\boldsymbol{\lambda}$. We can define the Lagrange dual function for problem (1.1) – (1.2) as:

$$g(\boldsymbol{\lambda}) = \sup_{x \in X} f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top (A\boldsymbol{x} - \boldsymbol{b}). \tag{1.3}$$

$g(\boldsymbol{\lambda})$ provides an upper bound on p^* if $\boldsymbol{\lambda} \geq \boldsymbol{0}$ and it is a convex function, even if X is a non-convex set [BV04, Chapter 5]. In order to obtain the best upper bound on p^* we can minimize $g(\boldsymbol{\lambda})$, leading to the Lagrange dual problem:

$$d^* = \min_{\lambda \geq 0} g(\boldsymbol{\lambda}). \tag{1.4}$$

The dual optimal value d^* is an upper bound on p^*, which might or might not be tight. Regardless, the dual problem (1.4) allows us to solve, at least approximately, the primal problem (1.1) – (1.2): assume that there is a practically efficient method to solve the dual problem and that a feasible point $\tilde{\boldsymbol{x}}$ for the primal problem can be recovered from the dual solution $\boldsymbol{\lambda}^*$, then $\tilde{\boldsymbol{x}}$ is

a solution to the primal problem within an optimality gap of $d^* - f(\tilde{\boldsymbol{x}})$. This approach is known as Lagrange relaxation and it is commonly used in Integer Programming for solving problems where optimizing over X is much simpler than optimizing over X and $A\boldsymbol{x} \leq \boldsymbol{b}$ [Geo10].

Note that it might not be possible to solve the dual problem (1.4) using differentiable optimization algorithms, since the Lagrangian dual function (1.3) might be non-differentiable if X is not convex or if f is not strictly concave. In general, in such cases, it becomes necessary to use subgradient methods (introduced in sections 1.2.2 and 1.2.3) or other non-differentiable optimization algorithms [Nes04]. On the other hand, if X is a convex set, then we can define a smooth approximation of g following [Nes05]:

$$g_\mu(\boldsymbol{\lambda}) = \sup_{x \in X} \; f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top (A\boldsymbol{x} - \boldsymbol{b}) - \frac{1}{2}\mu \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2, \tag{1.5}$$

where $\mu > 0$ is the smoothness parameter and $\boldsymbol{x}_0 \in \mathbb{R}^n$ is the point around which the approximation is built. g_μ is continuously differentiable and its gradient is Lipschitz continuous with constant $L^\mu = 1/\mu \cdot \|A\|_2^2$ [Nes05, Theorem 1]. In other words, the larger μ the smoother the approximation will be, at the cost of a higher error with respect to the original function.

g_μ can be used in place of g in order to find an approximate solution to the dual problem (1.4) faster than with non-differentiable optimization methods. However, since $g_\mu(\boldsymbol{\lambda}) \leq g(\boldsymbol{\lambda})$, we must evaluate the original dual function at the approximate minimizer to obtain a non-trivial upper bound on the primal problem.

1.2.2 Subgradient method

Assume, now, that we would like to solve problem (1.4). If f is strictly concave and X is convex, then we can use the projected gradient method to perform the minimization because g is differentiable and its gradient at $\boldsymbol{\lambda}$ is $(A\boldsymbol{x}^*)^\top$, where $\boldsymbol{x}^*$ is the supremizer of (1.3). On the contrary, if the aforementioned conditions are not met, then g might be non-differentiable. Luckily, we can define an extension of the gradient for non-differentiable convex functions known as the subgradient [Nes04]. $\boldsymbol{h}$ is a subgradient of g at $\boldsymbol{\lambda}_0$ if and only if

$$g(\boldsymbol{\lambda}) \geq g(\boldsymbol{\lambda}_0) + \boldsymbol{h}^\top (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) \quad \forall \boldsymbol{\lambda} \in \operatorname{dom} g,$$

in other words, a subgradient is the slope of a supporting hyperplane of g at $\boldsymbol{\lambda}_0$. Subgradients are not unique; the set of all subgradients is called the subdifferential and the subdifferential of g at $\boldsymbol{\lambda}_0$ is denoted $\partial g(\boldsymbol{\lambda_0})$. In the case of problem (1.4), the subdifferential of the objective function can be written as $\partial g(\boldsymbol{\lambda}) = \{(A\boldsymbol{x})^\top \; \forall \boldsymbol{x} \in X^*(\boldsymbol{\lambda})\}$, where $X^*(\boldsymbol{\lambda})$ denotes the set of supremizers of (1.3).

Algorithm 1 Projected subgradient method.

Initialize $\boldsymbol{\lambda}^1 := \boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}^{\text{best}} := \boldsymbol{\lambda}_0$, $d^{\text{best}} := \infty$
for $k = 1, \ldots, N$ **do**
  Evaluate $g(\boldsymbol{\lambda}^k)$ and obtain $\boldsymbol{h} \in \partial g(\boldsymbol{\lambda}^k)$
  if $g(\boldsymbol{\lambda}^k) < d^{\text{best}}$ **then**
    $\boldsymbol{\lambda}^{\text{best}} := \boldsymbol{\lambda}^k$, $d^{\text{best}} := g(\boldsymbol{\lambda}^k)$
  end if
  $\boldsymbol{\lambda}^{k+1} := \mathcal{P}_\Lambda[\boldsymbol{\lambda}^k - \alpha_k \cdot \boldsymbol{h}]$
end for
Output: $\boldsymbol{\lambda}^{\text{best}}$, d^{best}

Subgradients can be used in an analogous way to gradients for minimizing a non-differentiable convex function, leading to the projected subgradient method presented in Alg. 1. Here Λ corresponds to the domain over which we are minimizing ($\Lambda = \mathbb{R}^m_+$ for problem (1.4)), $\{\alpha_k\}_{k=1}^N$ is a sequence of stepsizes ($\alpha_k \geq 0$ $k = 1, \ldots, N$), and $\boldsymbol{\lambda}^{\text{best}}, d^{\text{best}}$ are the best iterate and upper bound on d^* respectively, which must be tracked since the subgradient method is not guaranteed to descend at every iteration. Convergence of the subgradient method, in the sense that $d^{\text{best}} - d^* \to 0$ as $N \to \infty$, is guaranteed for the following stepsize choices [Pol69]:

- Non-summable ($\sum_{k=1}^\infty \alpha_k = \infty$), square-summable ($\sum_{k=1}^\infty \alpha_k < \infty$) stepsizes.
- Polyak's stepsize $\alpha_k = (g(\boldsymbol{\lambda}^k) - d^*)/\|\boldsymbol{h}\|_2^2$, where $\boldsymbol{h} \in \partial g(\boldsymbol{\lambda}^k)$.

Despite their similarities in implementation, the subgradient method is significantly slower in achieving convergence than the gradient method [Nes04]. Other methods for non-differentiable convex minimization such as the bundle method or the level method can achieve better convergence speed on small-dimensional problems, however, their convergence speed deteriorates with the dimension m, which is not the case for the subgradient method. The convex non-differentiable optimization problems studied in this dissertation typically present large m's, which is why, we restrict our attention to subgradient methods.

1.2.3 Stochastic subgradient method

Under certain circumstances it might useful to randomize otherwise deterministic algorithms in order to improve their performance. The subgradient method can be randomized by using a stochastic subgradient of $g(\boldsymbol{\lambda}^k)$, i.e. a random vector $\boldsymbol{H}$ such that

Algorithm 2 Projected stochastic subgradient method.

1: Initialize $\boldsymbol{\lambda}^1 := \boldsymbol{\lambda}_0$
2: **for** $k = 1, \ldots, N$ **do**
3: Sample a realization $\boldsymbol{h}$ of $\boldsymbol{H}|\boldsymbol{\lambda}^k$
4: $\boldsymbol{\lambda}^{k+1} := \mathcal{P}_\Lambda[\boldsymbol{\lambda}^k - \alpha_k \cdot \boldsymbol{h}]$
5: **end for**
6: Output: $\boldsymbol{\lambda}^{N+1}$, $g(\boldsymbol{\lambda}^{N+1})$

$$g(\boldsymbol{\lambda}) \geq g(\boldsymbol{\lambda}^k) + \mathbb{E}[\boldsymbol{H}|\boldsymbol{\lambda}^k]^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}^k) \quad \forall \boldsymbol{\lambda} \in \operatorname{dom} g,$$

instead of a subgradient as the update direction, leading to the projected stochastic subgradient method presented in Alg. 2. The stochastic subgradient method is guaranteed to converge in the sense that $\lim_{N\to\infty} \boldsymbol{\lambda}^N \in \Lambda^*$, where Λ^* denotes the set of optimal solutions, with probability 1 when using a non-summable, square summable stepsize [Erm83]. Whereas the convergence guarantees are weaker for the stochastic method than for the deterministic method, if obtaining a realization of the stochastic subgradient is much cheaper than evaluating g, then the iteration cost is significantly reduced by randomizing the method. At the same time, randomization might lead to improvements on the convergence speed and allow for parallelization of the method [NB01, NBB01].

1.2.4 Cutting-plane method

Cutting-plane methods are a family of methods among which we find Bender's decomposition for linear programs [Ben62] and Kelly method for non-linear programming [Nes04]. The basic idea of these methods is to progressively approximate an optimization problem, or a complicated part of it, using linear inequalities. Consider the following problem, representative of our application cases:

$$\max_{x \in X} f(\boldsymbol{x}) \tag{1.6}$$

$$\text{s.t.} \quad g(\boldsymbol{x}) \leq 0, \tag{1.7}$$

where f is a continuous concave function that is easy to evaluate, $X \subset \mathbb{R}^n$ is a bounded set that might contain binary restrictions, n is small (in the order of 10, at most) and g is a convex function that can only be accessed through a first-order oracle, i.e. a 'function' that when evaluated at the query point $\boldsymbol{x}$ returns the value of the function $g(\boldsymbol{x})$ and the slope of a supporting hyperplane $\boldsymbol{h} \in \partial g(\boldsymbol{x})$.

Cutting-plane methods collect the information returned by the first-order

Algorithm 3 Cutting-plane method.

1: Initialize $k := 0$, $\Delta := \infty$
2: **while** $\Delta > 0$ **do**
3: Solve (1.6), (1.8) $\rightarrow \boldsymbol{x}^{k+1}$
4: k++
5: Query first-order oracle of g at $\boldsymbol{x}^k \rightarrow g(\boldsymbol{x}^k), \boldsymbol{h}^k$
6: Let $\Delta := g(\boldsymbol{x}^k)$
7: **end while**
8: Output: $\boldsymbol{x}^k$

oracle until the current iteration k in order to construct the following approximations of $g(\boldsymbol{x}) \leq 0$:

$$g(\boldsymbol{x}^i) + (\boldsymbol{h}^i)^\top (\boldsymbol{x} - \boldsymbol{x}^i) \leq 0 \quad i = 1, \ldots, k. \tag{1.8}$$

Since (1.8) is a relaxation of (1.7), if an optimal solution to (1.6) subject to (1.8), $\tilde{\boldsymbol{x}}^*$, complies with $g(\tilde{\boldsymbol{x}}^*) \leq 0$, then $\tilde{\boldsymbol{x}}^*$ is also an optimal solution to (1.6) – (1.7). Putting these observations together we arrive at the cutting-plane method presented in Alg. 3, whose convergence can be ensured in the sense of obtaining an exact solution in finite time for piece-wise linear g [Ben62] or an ϵ-approximate solution in a certain number of iterations for general convex functions [Nes04].

1.2.5 Column-generation method

Column-generation methods work by considering at first a subset of the problem variables, while letting the rest of the variables at zero, and progressively taking into account more and more variables until reaching optimality. They were originally proposed by Dantzig and Wolfe [DW60] for solving large-scale linear problems with block-angular constraint matrices. For simplicity of exposition, we present the method for the following linear programming (LP) problem

$$\min_{x,y \geq 0} \ \boldsymbol{f}^\top \boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{y} \tag{1.9}$$

$$\text{s.t.} \quad A\boldsymbol{x} + B\boldsymbol{y} \geq \boldsymbol{c} \tag{1.10}$$

$$D\boldsymbol{y} \geq \boldsymbol{e}, \tag{1.11}$$

and we assume that $\{\boldsymbol{x} \in \mathbb{R}^n_+ | \, A\boldsymbol{x} \geq \boldsymbol{c}\}$ is a non-empty set, that $Y := \{\boldsymbol{y} \in \mathbb{R}^l_+ | \, D\boldsymbol{y} \geq \boldsymbol{e}\}$ is a non-empty bounded set and that optimizing a linear function over Y is much simpler than over the feasible set of (1.9) – (1.11) (which is the case, for instance, when D is a block-diagonal matrix).

Algorithm 4 Column-generation method.

1: Initialize $k := 0$, $\Delta := -\infty$
2: **while** $\Delta < 0$ **do**
3: Solve (1.12) – (1.14) $\rightarrow \boldsymbol{x}^*, \boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \gamma^*$
4: Let $\boldsymbol{v}^{k+1} := \arg\min_{y \in Y} (\boldsymbol{g} - B^\top \boldsymbol{\lambda}^*)^\top \boldsymbol{y}$, $\Delta := (\boldsymbol{g} - B^\top \boldsymbol{\lambda}^*)^\top \boldsymbol{v}^{k+1} - \gamma^*$
5: $k++$
6: **end while**
7: Output: $\boldsymbol{x}^*, \boldsymbol{y}^* = \sum_{i=1}^{k-1} \theta_k^* \boldsymbol{v}^i$

Column-generation is based on re-writing problem (1.9) – (1.11) as

$$\min_{x,\theta \geq 0} \ \boldsymbol{f}^\top \boldsymbol{x} + \boldsymbol{g}^\top \sum_{i=1}^{k} \theta_i \boldsymbol{v}^i \tag{1.12}$$

$$\text{s.t.} \ \ A\boldsymbol{x} + B \sum_{i=1}^{k} \theta_i \boldsymbol{v}^i \geq \boldsymbol{c} \qquad [\boldsymbol{\lambda}] \tag{1.13}$$

$$\sum_{i=1}^{k} \theta_k = 1 \qquad [\gamma] \tag{1.14}$$

where $\boldsymbol{v}^i$ are the extreme points of Y, in other words, we re-write y as a convex combination of the extreme points of Y. This alternate formulation cannot be solved directly for realistic cases because the number of extreme points of a polytope can be very large. Nevertheless, we can consider a small subset of the extreme points of Y in (1.12) – (1.14), leading to the so-called Restricted Master Problem (RMP), and add more extreme points into the problem only if they can reduce the objective. These extreme points can be detected by means of their reduced cost, i.e. if $(\boldsymbol{\lambda}^*, \gamma^*)$ is dual optimal for RMP and there exist an extreme point $\boldsymbol{y}$ of Y such that $(\boldsymbol{g} - B^\top \boldsymbol{\lambda}^*)^\top \boldsymbol{y} < \gamma^*$, then including $\boldsymbol{y}$ in the RMP would lead to a decrease in the objective function of the RMP. Following this line of reasoning, the column-generation method can be cast as Alg. 4.

Note that Alg. 4 uses the same principles than the simplex method uses for selecting variables entering the basis. The algorithm is guaranteed to converge to optimality in finite time because Y has only a finite number of extreme points.

1.3 Preliminaries on transmission-constrained electricity markets

Exchanges of energy in electricity markets are physically bounded by the constraints of the transmission grid. These constraints include thermal limits of transmission lines, voltage stability limits, frequency stability limits and security limits (i.e. reserve margins). Considering all of these constraints together can be cumbersome even for small systems, hence, in practice, most system operators use the DC approximation of the power flow equations alongside thermal limits and security limits in the scheduling and planning of power systems. This approximation has been found to be accurate for the high-voltage transmission grid under normal operation [SJA09] and we assume it to represent the real network constraints throughout this dissertation.

Beyond the approximations required for tractability, different electricity markets might chose to consider different representations of the grid, because of pricing, liquidity or political reasons. Electricity markets then work as auctions that are cleared respecting the constraints of these grid representations. This section introduces three different grid representations in electricity markets: the nodal representation and two different zonal representations. For simplicity of the exposition we assume that demand is fixed and that only producers bid in the market.

1.3.1 Nodal electricity markets

Nodal electricity markets use Locational Marginal Prices (LMPs), first proposed by Schweppe *et al.* [SCTB88], in order to price electricity at every node of the system, while accounting for transmission congestion in a DC approximation of the real network equations. Under our assumptions, the LMP market can be cleared by solving the optimization problem (1.15) – (1.17):

$$\min_{v\in[0,1],f,\theta} \quad \sum_{g\in G} P_g Q_g v_g \tag{1.15}$$

$$\text{s.t.} \quad \sum_{l\in L(n,\cdot)} f_l - \sum_{l\in L(\cdot,n)} f_l = \sum_{g\in G(n)} Q_g v_g - Q_n \quad \forall n \in N \quad [\rho_n] \tag{1.16}$$

$$-F_l \le f_l \le F_l,\ f_l = B_l\left(\theta_{m(l)} - \theta_{n(l)}\right) \quad \forall l \in L. \tag{1.17}$$

The notation in this model is as follows: Q_g, P_g correspond to the quantity and price bid by generator $g \in G$; $G(n)$ is the set of generators at node n; Q_n is the demand at node $n \in N$; $F_l, B_l, m(l), n(l)$ are the thermal limit, susceptance, and adjacent nodes (in the outgoing and incoming direction respectively) of line $l \in L$; $L(m,n)$ is the set of lines directed from node m to node n; v_g is the acceptance/rejection decision for the bid placed by generator g; f_l is the flow

through line l; and θ_n is the voltage angle at node n.

The objective function (1.15) corresponds to the total operation cost, constraints (1.16) enforce nodal power balance, and constraints (1.17) model the DC transmission constraints using the B-θ formulation[2]. Nodal prices ρ_n can be obtained as the dual multipliers of constraints (1.16).

The LMP policy implicitly allocates the capacity of all lines in the system while respecting the network constraints. This implies that the optimal acceptance/rejection decisions v^* can be implemented directly in the system without violating any technical constraint.

1.3.2 Zonal electricity markets with available-transfer-capacities

Zonal electricity markets schedule production and consumption in power systems using a simplified zonal representation of the underlying nodal electrical network. The zonal aggregation of the grid allows market participants to trade freely within each zone and to export/import energy to/from other zones up to certain technical limitations.

In zonal electricity markets with available-transfer-capacities (ATCs) these limitations are imposed on the bilateral exchanges between neighboring zones. In other words, zones correspond to vertices in a transportation network where every pair of zones connected by a transmission line in the real grid are connected by an edge, and the flow through that edge is limited by the ATCs [ES05].

A zonal market with ATCs can be cleared by solving problem (1.18) – (1.20):

$$\min_{v\in[0,1],e} \sum_{g\in G} P_g Q_g v_g \tag{1.18}$$

$$\text{s.t.} \quad \sum_{t\in T(z,\cdot)} e_t - \sum_{t\in T(\cdot,z)} e_t = \sum_{g\in G(z)} Q_g v_g - \sum_{n\in N(z)} Q_n \quad \forall z \in Z \quad [\rho_z] \tag{1.19}$$

$$-ATC_t^- \leq e_t \leq ATC_t^+ \quad \forall t \in T, \tag{1.20}$$

where Z is the set of zones, T is the set of interconnectors between zones, $G(z)$ and $N(z)$ are the set of generators and nodes in zone $z \in Z$, e_t corresponds to the bilateral exchange through interconnector $t \in T$, and ATC_t^-, ATC_t^+ are the backward and forward exchange limits. These limits are computed by system

[2] An equivalent formulation of constraints (1.17) can be obtained by eliminating variables θ_n, $\forall n \in N$, and expressing the power flows as linear functions of the nodal injections. This alternate formulation is known in the literature as the Power Transfer Distribution Factor ($PTDF$) formulation, where the $PTDF$ matrix maps nodal injections to line flows.

operators and they are supposed to reflect the maximum energy that can be transferred from one zone to the other.

The objective function (1.18) is identical to that of the nodal electricity market (1.15), whereas the balance constraints (1.19) are imposed over each zone instead of over each node, as in (1.16). Zonal prices ρ_z can be obtained as the dual multiplier of the zonal balance constraints (1.19). Note that, since problem (1.18) – (1.20) does not consider the real network, the optimal acceptance/rejection decisions v^* of (1.18) – (1.20) might be infeasible for the real network. Redispatch measures modifying v^* might then be required in order to recover implementable acceptance/rejection decisions.

1.3.3 Zonal electricity markets with a flow-based domain

Zonal electricity markets with a flow-based (FB) representation of the grid attempt to find a middle ground between nodal and zonal models [VBD16]. The core idea behind this representation is to approximate the flow on each line $l \in L$ as

$$f_l \approx f_l^0 + \sum_{z \in Z} PTDF_{l,z}(p_z - p_z^0),$$

where f_l^0 is the flow through line on a base case, p_z is the net position of zone z (i.e. exports − imports) and p_z^0 is the net position on the base case, and $PTDF_{l,z}$ are zone-to-line power-transfer-distribution-factors computed by system operators. Then, a zonal market with FB can be cleared by solving problem (1.21) – (1.24).

$$\min_{v \in [0,1], p} \sum_{g \in G} P_g Q_g v_g \tag{1.21}$$

$$\text{s.t.} \quad p_z = \sum_{g \in G(z)} Q_g v_g - \sum_{n \in N(z)} Q_n \quad \forall z \in Z \quad [\rho_z] \tag{1.22}$$

$$\sum_{z \in Z} p_z = 0 \tag{1.23}$$

$$-F_l \leq f_l^0 + \sum_{z \in Z} PTDF_{l,z}(p_z - p_z^0) \leq F_l \quad \forall l \in L \tag{1.24}$$

Constraints (1.23), (1.24) impose export/import limitations in the directly on the configuration of zonal net positions, defining the so-called flow-based domain. Zonal prices ρ_z can be computed as the dual multipliers of constraints (1.22) and, as with the zonal market with ATCs, the optimal acceptance/rejection decisions might be infeasible depending of the accuracy of the flow approximation, requiring redispatch measures to recover implementable acceptance/rejection decisions.

1.4 The European electricity market

The European electricity market has favored a zonal market design, known as Market Coupling (MC) [ABC+10a, HAA+17], over LMPs on the basis of simplicity and liquidity. There is a long-standing debate about the relative merits of the two designs, and the implementation of the market coupling design in Europe has generated considerable controversy (see [ES05] and references therein for a detailed discussion). European zonal markets are characterized by three features that differentiate them substantially from centralized nodal markets in the context of renewable energy integration: (*i*) the simplified representation of transmission at the day-ahead time stage, (*ii*) the sequential clearing of reserves and energy, and (*iii*) the limited real-time coordination among zones for relieving congestion and imbalances.

In this section we introduce the historical context where the European electricity market developed, we present a parallel between European and US electricity markets and we present a detailed instance of the Central Western European (CWE) system, which has taken the lead in implementing the market coupling design.

1.4.1 Historical context

The European zonal electricity market directly inherited some of its major characteristics from the Nordic system (Norway, Sweden, Finland and Denmark). Both are based on a decomposition of the market into zones connected by aggregated representations of the lines of the network. As in the European market, the decomposition into zones first reflects national borders. The day-ahead market is a pure energy market run by a Power Exchange (PX), which is seen as the spot market. The inevitable deviations between day ahead and real time are treated by a special mechanism (regulation market in the Nordic system), initially run by national TSOs that was only progressively integrated among them. Notwithstanding the integration of the energy market through a single PX, the management of the grid remains zonal (this time national). Redispatching is the main tool for dealing with congestion management in the Nordic system, but TSOs can also rely on the so called "market splitting" to deal with intra-zonal congestion. This property was only systematically used by Norway but the EU competition authorities forced it on Sweden when the later country had difficulties managing congestion on a line to Denmark.

Other regions of the world have also experienced zonal markets. Australia is a case in point that one might want to compare to Europe insofar as the markets of the different States each form an area connected to the neighboring States by transport capacities. The zones are fixed and determined by the State borders. Zonal markets also flourished in the US but soon gave way to nodal organizations. PECO was the first example of a zonal market: it was created

in 1997 but collapsed after a year due to dispatching difficulties. The zonal system was one of the causes of the meltdown of the first Californian restructured market, which was subsequently replaced by a nodal market. Finally ERCOT was the last attempt to install a zonal market in the United States. Here too, redispatching costs exploded compared to what was initially planned and ERCOT moved to the LMP model. Other restructured US systems immediately adopted the LMP framework.

The first trilateral version of the European market, coupling Belgium, France and the Netherlands, came live in November 2006 [ABP06] as a zonal system possibly inspired by the Nordic success and its careful mix of integration and remaining national identities. It has since developed as to encompass the whole continent including the Nordic countries. But the European zonal system did not resort to market splitting (except in the Nordic countries) but to a construction of cross border capacities (ATCs, see subsection 1.3.2), initally exposed in Regulation (EC) 1228/2003 [Eur03], later replaced at the CWE by the notion of flow-based domain (see subsection 1.3.3) following Regulation (EC) 714/2009 [Eur09]. The zones correspond to Member States (which can be as large areas as France and Germany or as small as Belgium). Congestion is fully managed by redispatch. While the zonal system is enshrined in legislation, including Regulations (EC) 1228/2003 [Eur03], 714/2009 [Eur09] and 1222/2015 [Eur15a], one cannot exclude that the evolution may end up being more in line with what happened in the US. The reasons are technical: well before restructuring a common Nordic argument was that their transmission system had been designed for north-south transport (at the time Danish coal in the south and hydro in the north). The argument remains valid today but in a different form (Danish renewable in the south and storage in the north). Unscheduled flows were thus not a main issue, given the largely radial grid structure. But neither the EU grid not the US system were designed for transport with the consequence that unscheduled flows and congestion may turn out to be more important. Also while the Nordic system had that zone splitting effectively used in Norway to relieve the recourse to redispatching, the EU has so far always resisted resorting to smaller zones that could help when a single price zone contains a congested "critical infrastructure". Recalling ERCOT experience when redispatching costs had been dismissed as irrelevant before they blew up a few years later, one can only note that these costs that were deemed to remain below 45 M€/year in Germany (2007, maximum prior to market coupling) now amount to 1 161 M€/year in that country (2017) [Eur15b].

1.4.2 European vs. US electricity markets

Setting aside the zonal/nodal difference between current European and US electricity markets, there are other subtle differences and multiple shared characteristics between them. For a start, while several US electricity markets have

adopted LMP and centralized operations under a single ISO, zonal prices continue to be used for billing loads in certain markets such as NYISO [New16a]. Nevertheless, in contrast to European market coupling design, these zonal prices are computed after scheduling production with a nodal model [New16b], hence the use of zonal prices does not create unscheduled flows on the transmission network.

Another difference between US and European electricity markets is that day-ahead markets in the US are considered as forward markets, several of them accepting virtual bids [Cal18]. European day-ahead markets, on the other hand, are treated as the spot electricity market, explicitly forbidding virtual bidding [EPE15], and considering most actions following the clearing of the day-ahead market as out-of-market corrections.

The market coupling design is, to some extent, a counterpart to the bilateral market-to-market operations in US power systems. Market coupling uses uniform pricing within each zone and clears market-to-market interchanges within the European wide day-ahead energy market, implicitly allocating transmission capacity between zones [ABC+10a]. US markets use LMP within each balancing authority area [Nor12], however market-to-market interchanges are usually arranged prior to clearing of the day-ahead market in a bilateral fashion and they must be approved by every affected ISO [Nor12, Sou15]. Approved interchanges are then considered firm and unscheduled flows caused by them are taken into account in day-ahead market operations [Cal15].

Similar real-time coordination mechanisms are used both in market coupling and wide US interconnections. In the market coupling design, balance responsible parties are entitled to maintain their scheduled net positions in real time [ENT14]. Likewise, in US markets each balancing authority (ISO) must maintain its area control error (a measure consisting of the difference between scheduled and real area net position, and the area frequency deviation) below certain established limits [Nor].

These similarities make certain operational lessons learned in one market relevant for the other, particularly with regards to the challenges of integrating renewables, where Europe has taken the lead, exhibiting today many of the problems expected in the US markets within the coming years [ESS+10, MDE+10].

1.4.3 A detailed instance of the Central Western European system

At the time of starting this dissertation a number of studies had focused on developing representative models for the European grid. Leuthold [LWv12] and Hutcheon and Bialek [HB13] model the European transmission grid based on study models and maps published by ENTSO-E [Eurc] and by national TSOs, while Egerer *et al.* [EGI+14] document the existing available information on

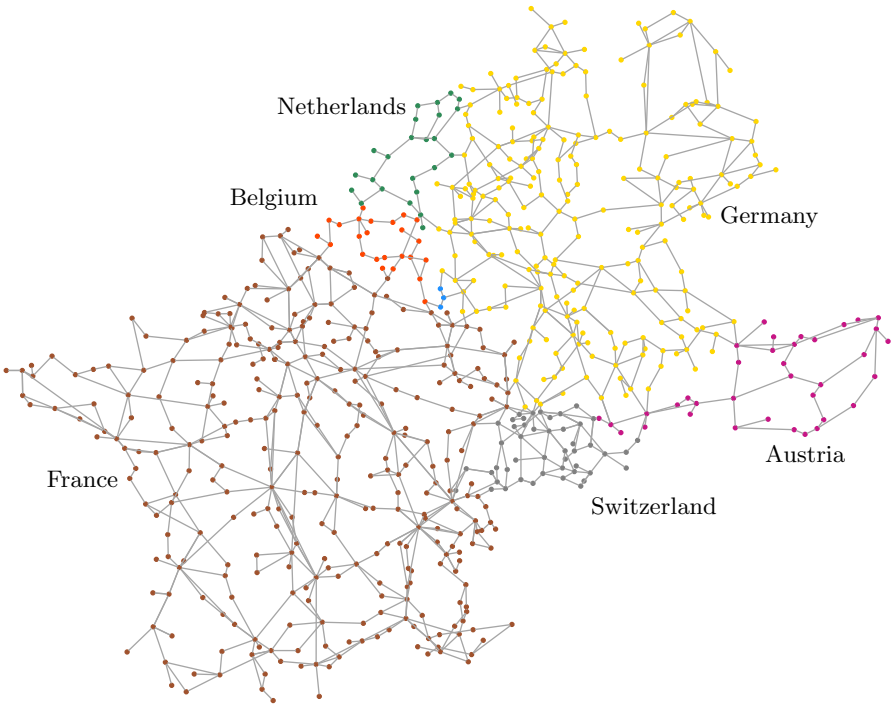


Figure 1.1: Nodal model of the CWE network [HB13]. The model comprises the high voltage networks of the 7 countries in the CWE system, it includes 679 nodes and 1073 lines.

transmission, generation and demand in Europe. These models and studies, however detailed, did not encompass the required information for simulating unit commitment on a European scale under different system conditions. This motivated us to take upon the task of putting together a detailed instance of the CWE system, which is constructed as follows.

We use the transmission network model of Hutcheon and Bialek [HB13] for the CWE system, presented in Fig. 1.1. Thermal ratings for cross-border lines were updated to their current values, as published in [EE14]. Thermal ratings for internal lines within the Netherlands were established as published in [Ten09]. Thermal ratings for internal lines within other countries were estimated through an iterative process of simulating system operations and correcting internal capacities, with the objective of approximating the congestion management costs for the year 2015 [Eurc].

The network was populated using an industrial database of thermal generators, provided by ENGIE, which includes technical and economic characteristics of 656 generating units. Thermal generators were assigned to network buses ac-

cording to their approximate geographical location [DCDN14]. These units are classified into four groups: 87 nuclear units (85G W), 144 combined heat and power (CHP) units (40 GW), 398 conventional thermal units (neither nuclear nor CHP, totaling 113 GW) and 27 aggregated small generators (10 GW).

The capacity of thermal generators within Germany, France and Belgium was reduced in order to account for scheduled maintenance and large outages. A different outage de-rating factor was computed for each generator and each season based on the outage information published by national TSOs [Res, ELI] and PXs [Eurb] for the year 2014.

Zonal reserve targets were obtained from [HAE+14] for Belgium, Germany and the Netherlands, and from national TSOs for other countries [Res, Aus, Swi15].

Historical 15-minute demand profiles for 2014 were collected from national TSOs for Austria [Aus], France [Res], Belgium [ELI] and the Netherlands [Ten], and hourly demand profiles for Germany and Switzerland were collected from ENTSO-E [Eurc]. Demand profiles were distributed across the buses of the network within the relevant area using the participation factors included in [HB13]. Exchanges between CWE countries and non-CWE countries are collected from [Eurc] and are modeled as fixed flows of power at the corresponding borders.

Regional 15-minute production profiles and day-ahead forecasts for wind and solar PV for years 2013-2014 were also collected from national TSOs and power exchanges. The spatial resolution of renewable production data varies from country to country. There are 4 geographical regions in Germany, 21 regions in France, 2 regions in Belgium and 1 region in Austria. Profiles for the Netherlands and Switzerland were estimated by averaging data of neighboring regions. Offshore wind power profiles were associated to offshore wind connection buses in the transmission system. Onshore wind and solar PV capacity within each administrative region of each country (12 states of Germany, 21 regions of France, 2 regions of Belgium, Austria, Switzerland and the Netherlands) was distributed uniformly among generation and load buses within that region. Each bus with renewable capacity was then assigned a wind and a solar PV production profile according to its location, therefore the spatial information of renewable resource dispersion is preserved in our data set.

Hydro power resources are modeled as fixed injections or withdrawals from the transmission system. Hydro production profiles for seasonal storage, pumped storage and run of river were collected from RTE [Res] for each power plant in France. Hydro plants in other countries were assigned production profiles based on profiles of French hydro plants and by taking into account the characteristics of these plants (technology, size and location).

We used clustering to select 8 representative day types of load for 2014, corresponding to one weekday and one weekend day for each season. We used the forecast errors of 2013-2014 to generate samples of real-time renewable

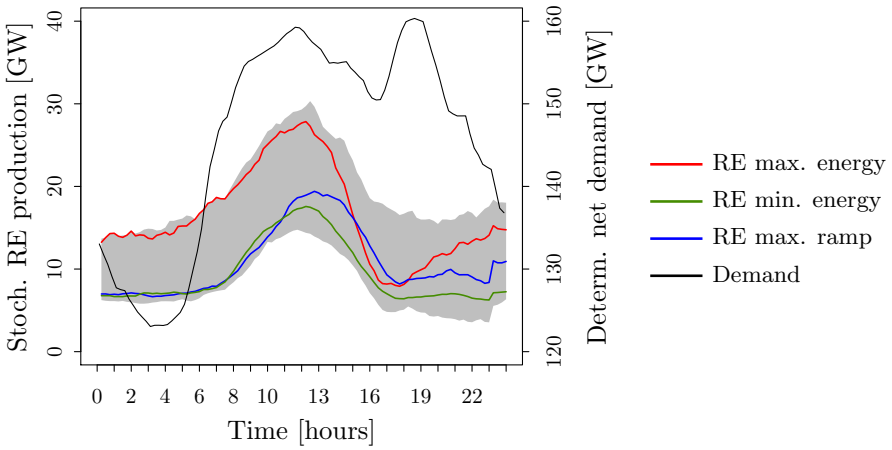


Figure 1.2: Stochastic renewable energy production (wind and solar) and deterministic net demand (demand minus hydro power) for a typical autumn weekday. The shaded gray area, included in the background, shows the variation range of renewable energy production.

energy production. Fig. 1.2 presents the resulting deterministic net demand and uncertainty faced by the system in the day ahead. The net load forecast error can span more than 15 GW and can exhibit ramps of up to 3 GW in 15 minutes. Renewable production ramps of this magnitude already occur in Germany, for instance in May 11, 2014, between 17:45 and 18:00 [Eurb].

1.5 Structure of the dissertation

The main contributions of this dissertation are organized in three chapters, which are summarized in the following.

1.5.1 Chapter 2

In Chapter 2, we present an asynchronous algorithm for solving the stochastic unit commitment (SUC) problem using Lagrangian decomposition. The algorithm is motivated by large differences in run times observed among subproblems, which can result in inefficient use of distributed computing resources by synchronous parallel algorithms. Dual iterations are performed asynchronously using a block-coordinate subgradient descent method which allows performing block-coordinate updates using delayed information, while candidate primal solutions are recovered from the solutions of scenario subproblems using heuristics. The asynchronous algorithm is implemented in a high performance computing cluster and we conduct numerical experiments for two-stage SUC

instances of the Western Electricity Coordinating Council (WECC) system with up to 1000 scenarios [PO13] and of the CWE system with up to 120 scenarios, presented in section 1.4.3. The algorithm provides solutions to all problems within 2% of optimality in at most 23 minutes for WECC and 98 minutes for CWE, and solutions within 1% of optimality in at most 63 minutes for WECC and 133 minutes for CWE. Moreover, we find that an equivalent synchronous parallel algorithm would leave processors idle up to 80.4% of the time, an observation which underscores the need for designing asynchronous optimization schemes in order to fully exploit distributed computing on real world applications.

1.5.2 Chapter 3

Chapter 3 proposes a novel framework for modeling zonal electricity markets, based on projecting the constraints of the nodal network onto the space of the zonal aggregation of the network. The framework avoids circular definitions and discretionary parameters, which are recurrent in the implementation and study of zonal markets. Using this framework, we model and analyze the two zonal market designs currently present in Europe: FBMC and ATCMC. We develop cutting-plane algorithms for simulating FBMC and ATCMC while accounting for robustness of imports/exports to single element outages, and we conduct numerical simulations of FBMC and ATCMC using single-period models for the CWE instance of section 1.4.3 under 768 000 different operating conditions. We find that FBMC and ATCMC are unable to anticipate congestion of branches interconnecting zones and branches within zones, and that both zonal designs achieve similar overall cost efficiencies (0.5% difference in favor of FBMC), while a nodal market design largely outperforms both of them (5.9% better than FBMC). These findings raise the question of whether it is worth for more European countries to switch from ATCMC to FBMC, instead of advancing directly towards a nodal design.

1.5.3 Chapter 4

Chapter 4 investigates the impact of zonal network management at an operational level for power systems with significant levels of renewable energy integration. The study is inspired by the state of the European energy market at the beginning of 2015, and we focus on a case study of the CWE system. We use a hierarchy of models that account for unit commitment, the separation of energy and reserves, and the simplified representation of transmission constraints in a zonal market, in order to examine the impact of these factors on efficiency in a regime of large-scale renewable energy integration. We simulate operations of the CWE system under the zonal market design using the instance of section 1.4.3 and we compare zonal market operations against

deterministic unit commitment and SUC. We find that market design can have an influence on cost efficiency which far exceeds the benefits of stochastic unit commitment relative to deterministic unit commitment. We conduct a detailed analysis of the numerical results in order to explain the relative performance of the different models.

Chapter 2

Solving stochastic unit commitment through asynchronous parallel optimization

2.1 Introduction

The unit commitment problem is a classical problem in the short-term scheduling of electric power systems. Unit commitment deals with deciding which generating units will supply energy to a power system over a certain time horizon, so as to minimize the operation cost while respecting the technical constraints of the power system. The problem is usually formulated as a mixed integer linear program (MILP) and it is solved on a daily basis by power systems operators worldwide. Stochastic unit commitment (SUC) is a widely studied approach to incorporate uncertainty in this scheduling problem. SUC can be formulated as the two-stage stochastic mixed integer program (SMIP) (2.1) – (2.4),

$$\max_{u,v,w} \sum_{i=1}^{N} (\boldsymbol{c}_i^T \boldsymbol{v}_i + \boldsymbol{d}_i^T \boldsymbol{w}_i) \tag{2.1}$$

$$\text{s.t.} \quad \boldsymbol{u} \in \mathcal{U}, \tag{2.2}$$

$$(\boldsymbol{v}_i, \boldsymbol{w}_i) \in \mathcal{D}_i, \quad i = 1, \ldots, N \tag{2.3}$$

$$\boldsymbol{v}_i - \boldsymbol{u} = \boldsymbol{0}, \quad i = 1, \ldots, N \tag{2.4}$$

where i indexes scenarios[1], $\boldsymbol{u}$ is the vector of non-anticipative decision variables, $\mathcal{U} \subset \mathbb{R}^n$ is a bounded convex set[2], $\boldsymbol{v}_i$ are local copies of non-anticipative variables at each scenario, $\boldsymbol{w}_i$ are recourse variables of each scenario, $\mathcal{D}_i \subset \mathbb{R}^{n+m_i}$ $i = 1, \ldots, N$ are bounded non-convex sets and, constraints (2.4) models non-anticipativity constraints. We assume that the problem has relatively complete recourse, that is $\mathcal{P}_{\boldsymbol{v}_i}(\mathcal{D}_i) = \mathcal{P}_{\boldsymbol{v}_j}(\mathcal{D}_j)$ for any $i, j \in \{1, \ldots, N\}$ where $\mathcal{P}_{\boldsymbol{v}_i}$ is the projection of $\mathcal{D}_i$ on the coordinates of $\boldsymbol{v}_i$. Typically, $\boldsymbol{v}_i$ corresponds to commitment variables of thermal generators (binary decisions), but it can also include production variables of inflexible generators (continuous decisions), and $\boldsymbol{w}_i$ includes commitment variables of fast generators, production variables of all generators and flows over the network (mixed integer decisions). $\mathcal{D}_i$ describes the feasible operation domain for scenario i in terms of production constraints (minimum stable level, maximum capacity, maximum ramp rates, minimum up/down times) and power grid constraints (power balance, power flow equations, flow limits). $\mathcal{U}$ corresponds to a convex relaxation of the production constraints for variables included in $\boldsymbol{u}$, $\mathcal{P}_{\boldsymbol{v}_i}(\mathcal{D}_i) \subseteq \mathcal{U}$ for any $i, j \in \{1, \ldots, N\}$. See [POR15] for a detailed description of the SUC model.

Our aim is to solve problem (2.1) – (2.4) for real power systems, within the time limits imposed by daily operations. This differentiates SUC from other applications of stochastic programming in that the typical scale of realistic SUC instances (see [POR15, CGSM+15, vAM16]), as measured by the number of variables and the number of constraints required to describe $\mathcal{U}$ and $\mathcal{D}_i$, is orders of magnitude larger than common test cases for SMIP, such as `sslp` or `dcap` [AGK+15]. Furthermore, the computational effort required for optimizing over $\mathcal{D}_i$ can vary significantly from one scenario to another, as well as for the same scenario with slight modifications. Therefore, a parallel implementation of a serial decomposition scheme may perform inefficiently in practice.

In order to overcome these challenges, in this chapter we present an asynchronous distributed algorithm for solving (2.1) – (2.4) and we present a high performance computing implementation of the algorithm which is used for solving SUC instances of two industrial-scale systems.

2.1.1 Literature review

Stochastic unit commitment was initially proposed in the seminal work of Takriti *et al.* [TBL96] and Carpentier *et al.* [CGCR96] as a methodology for coping with demand uncertainty in power systems. The scope for application of SUC has, since then, been extended to renewable energy forecast uncertainty

[1] For conciseness, we include the probabilities of each scenario within the coefficients $\boldsymbol{c}_i$ and $\boldsymbol{d}_i$ in (2.1).

[2] Note that this is not a restrictive assumption on the first stage variables or constraints of problem (2.1)–(2.4). All non-convexities can be considered within sets $\mathcal{D}_i$, $i = 1, \ldots, N$, and they are enforced on the first stage variables through the non-anticipativity constraints (2.4).

and component failures (referred to as contingencies in the power engineering literature), among other sources of uncertainty in power systems operations.

SUC studies commonly use decomposition methods for handling the scale of the problem, which increases linearly with the number of scenarios. Takriti *et al.* [TBL96] use Lagrange relaxation of non-anticipativity constraints (i.e. scenario decomposition), and a progressive hedging heuristic to obtain non-anticipative solutions. Carpentier *et al.* [CGCR96] relax the problem over generators using an augmented Lagrangian. They maximize the dual function with a proximal point method and recover primal solutions by allowing demand shedding at a quadratic penalty. Shiina and Birge [SB04] use a column generation approach over production schedules, decomposing over generators and solving the slave production scheduling problems using dynamic programming. Cerisola *et al.* [CBFL$^+$09] compare scenario decomposition with variants of Benders decomposition and stress the importance of finding good initial solutions. Additional decomposition approaches for SUC can be found in [TvAFL15].

SUC instances in the literature have often been limited to test systems which fall short of industrial scale instances. In order to advance towards more realistic instances, several recent studies exploit distributed computing alongside decomposition methods. Papavasiliou *et al.* [POR15] implement scenario decomposition in an HPC cluster to solve instances of the WECC system with up to 1000 scenarios in at most 24 hours within an optimality gap of 1% – 2.5%. Cheung *et al.* [CGSM$^+$15] decompose the problem by scenarios and use progressive hedging on a multi-processor workstation and on an HPC cluster to solve instances of the WECC system with up to 100 scenarios in at most 25 minutes within an optimality gap of 1.5% – 2.5%. Kim and Zavala [KZ15] propose an interior point cutting-plane algorithm to handle dual iterations in scenario decomposition and solve SUC instances based on the IEEE 118-bus test system with up to 32 scenarios, using an HPC cluster, in 6 hours within an optimality gap of 0.01%.

Other recent methods, which do not exploit distributed computing explicitly, have also being used to solve large SUC instances. Among them, Schulze *et al.* [SGM15] use a stabilized Dantzig-Wolfe decomposition to solve multi-stage SUC instances of the British system with up to 50 scenarios in 2 hours within an optimality gap of 0.1%. van Ackooij and Malick [vAM16] use primal-dual decomposition along with bundle methods to solve SUC instances of the French system with up to 250 scenarios within an optimality gap of 1%, however no solution times are reported.

As our research is focused on making it practical to solve industrial-scale SUC instances, (*i*) we present a method capable of solving SUC instances faster than the state-of-the-art, (*ii*) we release all the industrial-scale test instances used in this study and (*iii*) we provide the time it takes to solve them with the proposed method. These three elements are not found simultaneously in any

of the aforementioned studies.

In the broader class of general stochastic mixed integer programs, to which SUC belongs, scenario decomposition has inspired several different decomposition algorithms. Carøe and Schultz [CS99] propose a scenario decomposition method where non-anticipativity constraints are gradually enforced through a branch-and-bound algorithm. Lubin *et al.* [LMPS13] extend the previous method to a parallel computing setting by solving the dual problems at each node of the branch-and-bound tree using a structure-exploiting interior point solver. Oliveira *et al.* [OSS11], and references therein, use scenario decomposition and solve the dual problem using bundle methods. Ahmed [Ahm13, Ahm15] proposes an alternative approach for stochastic integer programs where solutions to scenario subproblems are directly used for primal recovery while, at the same time, these solutions are separated from subproblems in order to improve the bound of the relaxation. Ryan *et al.* [RRA16] develop several improvements over Ahmed's original algorithm, including a distributed asynchronous implementation.

Distributed computing has also found applications in stochastic programming outside the realm of scenario decomposition. Lubin *et al.* [LHPA13] propose a distributed memory simplex algorithm for stochastic linear programs. Munguía *et al.* [MOR15] propose a branch-and-bound algorithm for stochastic mixed integer programs on which each node of the tree is solved using Lubin's method. Moritchs *et al.* [MPS01] propose a nested asynchronous decomposition algorithm for multistage stochastic linear programs. Chaturapruek *et al.* [CDR15] propose and prove optimal convergence for asynchronous stochastic gradient descent methods on unconstrained stochastic programs with strongly convex and differentiable objective functions.

A crucial aspect to scenario decomposition is the method used to optimize the dual function, which is separable, convex and non-differentiable. Certain specialized methods allow exploiting the separable structure of the dual function in a distributed computing infrastructure. Nedić *et al.* [NB01, NBB01, Ned02] analyze incremental subgradient algorithms, on which each update is made along the direction of the subgradient of a part of the objective function. Coordinate descent methods are a different approach to exploit the structure of minimization problems for which it is cheaper to compute the gradient with respect to a subset of variables (coordinates) than it is to compute the full gradient of the objective. Wright [Wri15] provides a recent survey on coordinate descent methods. Nesterov [Nes12] provides worst-case complexity results of randomized coordinate descent methods for smooth optimization. Fercoq and Richtárik [FR13] propose a parallel synchronous coordinate descent method for minimizing non-differentiable simple composite functions. The authors use Nesterov's smoothing technique [Nes05] to obtain a smooth approximation of the non-decomposable part of the objective and perform a line search separately on each coordinate of each iteration, as proposed by Tseng [Tse01]. Fisher

and Helmberg [FH14] propose an asynchronous distributed bundle method for non-differentiable convex optimization, where at each step a processor greedily selects a subset of variables, blocks them from being accessed by the other processors and performs a proximal bundle iteration on the selected variables.

2.1.2 Contributions

Stochastic unit commitment, despite its attractiveness, has failed to become an industry standard due to several reasons, including the difficulty of solving the mathematical programs in an operationally acceptable time frame. In the present chapter, we aim at overcoming this challenge by developing an asynchronous scenario decomposition scheme based on distributed computing. The main innovation of the developed scheme is that we optimize the dual function using an asynchronous block-coordinated subgradient algorithm. Our algorithm for optimizing the dual function does not require differentiability or strong convexity assumptions commonly used in the literature [Nes12, Wri15, LWR$^+$15, CDR15]; it differs from the algorithm of Fercoq and Richtárik [FR13] in that we perform iterations asynchronously and without the need for line search, which would be prohibitive in our context; and it requires less serial coordination overhead than the asynchronous bundle method [FH14]. We provide convergence guarantees for the proposed algorithm for optimizing the dual function based on previous results for stochastic subgradient methods [Erm83] and incremental subgradient methods [NB01, NBB01, Ned02].

We also perform primal recovery asynchronously and in parallel to the dual iterations, either by recovering solutions from scenario subproblems, as proposed in [Ahm13, Ahm15], or by using recombination heuristics, similar to those proposed in [CS99]. The proposed asynchronous algorithm allows us to solve limited-size SUC instances faster than the state-of-the-art [CGSM$^+$15] and to solve SUC instances for systems larger than the state-of-the-art [vAM16] within operationally acceptable time frames.

Even though the proposed algorithm is inspired and tailored to SUC, the proposed framework for asynchronous dual decomposition can be applied in other contexts, such as temporal or spatial decomposition of unit commitment, as well as other two-stage SMIP problems.

2.1.3 Notation and chapter organization

We use boldface to denote vectors, lowercase letters to denote variables and uppercase letters to denote parameters or sets. Additionally, we use partial indexation of vectors to keep notation simple, i.e. $\boldsymbol{x} = [\boldsymbol{x}_1^T \ \dots \ \boldsymbol{x}_N^T]^T$ and $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$.

The rest of the chapter is organized as follows. Section 2.2 introduces the stochastic unit commitment problem and its scenario decomposition in

a stylized fashion. Section 2.3 presents the asynchronous distributed block-coordinate subgradient method and provides convergence results for the dual iterations. Section 2.4 describes our primal recovery heuristics. Section 2.5 describes the HPC implementation of the dual algorithm, and we cover aspects of communications and load balancing. Section 2.6 presents the numerical results for the WECC and CWE systems. Finally, section 2.7 presents conclusions and points to directions of future research.

2.2 Scenario decomposition in stochastic unit commitment

Problem (2.1) – (2.2) is a general formulation for two-stage mixed integer stochastic programs with finitely many scenarios and bounded feasible sets [LS04].

Following [TBL96] we relax problem (2.1)–(2.4) by associating multipliers $\boldsymbol{x}_i$ to non-anticipativity constraints (2.4), obtaining the dual problem (2.5), where f_0 and f_i are defined according to (2.6) and (2.7), respectively.

$$\min_{x \in \mathbb{R}^m} \quad f_0(\boldsymbol{x}) + \sum_{i=1}^{N} f_i(\boldsymbol{x}_i) \tag{2.5}$$

$$f_0(\boldsymbol{x}) := \sup_{u \in \mathcal{U}} \left(-\sum_{i=1}^{N} \boldsymbol{x}_i^T \right) \boldsymbol{u} \tag{2.6}$$

$$f_i(\boldsymbol{x}_i) := \sup_{(v,w) \in \mathcal{D}_i} \left((\boldsymbol{c}_i^T + \boldsymbol{x}_i^T)\boldsymbol{v} + \boldsymbol{d}_i^T \boldsymbol{w} \right) \quad i = 1, \ldots, N \tag{2.7}$$

Scenario decomposition schemes for solving (2.1)–(2.2) work by solving the dual problem (2.5) and generating primal solutions based on the solution to subproblems (2.6) and (2.7). The objective of problem (2.5) has a separable structure. Moreover, by evaluating the component functions f_0 and $f_i,\ i = 1, \ldots, N$ at a certain $\bar{\boldsymbol{x}}$, we obtain a subgradient of the objective at $\bar{\boldsymbol{x}}$. These two properties motivate the use of subgradient algorithms for solving (2.5) and evaluating the component functions in parallel in order to speed up the algorithm [POR15]. A third important property of problem (2.5) should be carefully considered, namely, the differences in evaluation times between component functions.

In our context, the evaluation of f_i at a certain $\bar{\boldsymbol{x}}_i$, $i = 1, \ldots, N$, requires the solution of a large mixed integer linear problem, while the evaluation of f_0 at a certain $\bar{\boldsymbol{x}}$ requires solving a medium-size linear program, hence the evaluation of f_0 requires only a small fraction of the time that it takes to evaluate f_i for any i[3].

[3]This differs from the usual scope of application of coordinate descent methods, where

In addition, for a given $\bar{\boldsymbol{x}}$ and two component functions f_i and f_j, the evaluation times of f_i and f_j can be dramatically different (we have observed differences of more than 7500%). These differences in evaluation time can also arise for the same component function evaluated at two different iterates. Altogether, these differences can render synchronous parallel algorithms ineffective because the time between iterations is limited by the component function which requires the greatest time to be evaluated. The main aim of the present work is to overcome this limitation.

2.3 Asynchronous distributed block-coordinate subgradient method

In this section we present a minimization method that exploits the special structure of (2.5) in order to effectively harness distributed computing. The proposed method has been inspired by previous work on incremental subgradient algorithms by Nedić *et al.* [NB01, NBB01, Ned02].

For reasons that we explain later, we replace the non-decomposable component function f_0 by a smooth approximation f_0^μ, defined in equation (2.8), as done in [FR13]. Recall that $\mu > 0$ and certain $\boldsymbol{u}_0 \in \mathcal{U}$, f_0^μ is a convex differentiable function and its gradient has Lipschitz constant L_0^μ [Nes05]. We focus then on solving problem (2.9), where X is a convex set, onto which it is easy to project.

$$f_0^\mu(\boldsymbol{x}) := \sup_{u \in \mathcal{U}} \left(\left(-\sum_{i=1}^{N} \boldsymbol{x}_i^T \right) \boldsymbol{u} - \frac{1}{2}\mu \|\boldsymbol{u} - \boldsymbol{u}_0\|_2^2 \right) \tag{2.8}$$

$$\min_{x \in X} f(\boldsymbol{x}) = f_0^\mu(\boldsymbol{x}) + \sum_{i=1}^{N} f_i(\boldsymbol{x}_i) \tag{2.9}$$

The solution of problem (2.9) will be an approximated solution to problem (2.5) [Nes05]. Note that problem (2.9), as well as problem (2.5), are non-differentiable convex optimization problems.

For ease of presentation, we first introduce a serial variant of the proposed method and then we generalize it an asynchronous distributed setting. Proofs for all results presented in this section are provided in section 2.A, at the end of this chapter.

the decomposable part of the objective is easier to evaluate than the non-decomposable part [Nes12, FR13, Wri15, LWR+15].

2.3.1 Serial method

Let us consider a block version of the randomized coordinate descent method [Nes12] for minimizing f. At iteration k, with current iterate $\boldsymbol{x}^k$, we select uniformly at random a block $j(k)$ from $\{1,\dots,N\}$ and compute the next iterate $\boldsymbol{x}^{k+1}$ using the following update rule:

$$\boldsymbol{x}^{k+1} := \mathcal{P}_X\left[\boldsymbol{x}^k - \lambda_k \cdot I_{j(k)}^T\Big(I_{j(k)}\nabla f_0^\mu\big(\boldsymbol{x}^k\big) + g\big(j(k), \boldsymbol{x}^k_{j(k)}\big)\Big)\right] \tag{2.10}$$

where λ_k is the step size, $g(j,\boldsymbol{x}) \in \partial f_j(\boldsymbol{x}_j)$ and I_j is a matrix that maps ∇f_0^μ to its components relevant to block j, i.e.

$$I_j = \left[\mathbf{0}_{n_j\times\left(\sum_{i=1}^{j-1} n_i\right)}\ \mathbf{1}_{n_j\times n_j}\ \mathbf{0}_{n_j\times\left(\sum_{i=j+1}^{N} n_i\right)}\right]$$

with $\mathbf{1}_{n_j\times n_j}$ corresponding to the identity matrix. Simply put, rule (2.10) updates the multipliers by moving them in the opposite direction to a subgradient of the objective function, restricted to the coordinates related to scenario $j(k)$.

Although f is a non-differentiable function, due to the smoothness of f_0^μ, the update rule (2.10) is equivalent to the update rule of the stochastic subgradient method, as in the following proposition.

Proposition 1. *Let J be a discrete uniform random variable on the set $\{1,\dots,N\}$. The expected direction of update rule* (2.10)*, $\mathbb{E}[I_J^T(I_J\nabla f_0^\mu\big(\boldsymbol{x}^k\big) + g\big(J,\boldsymbol{x}^k_J\big))|\ \boldsymbol{x}^k]$, coincides with the direction of a subgradient of f at $\boldsymbol{x}^k$.*

The stochastic subgradient method and its convergence have been well studied in the literature, see for instance [Erm83, Thm. 2]. By extension, with an appropriate selection of the step size λ_k (diminishing, nonsummable, square summable), the method defined by the update rule (2.10) will converge to an optimal solution with probability 1.

Note that the update rule (2.10) requires computing the gradient of f_0^μ and a subgradient of $f_{j(k)}$ at each iteration k, which makes it more expensive than the subgradient method when we consider an entire pass over all scenarios, i.e. N iterations of the method, which is the minimum amount of iterations required to update every block of variables. Nevertheless, since the cost of computing the gradient of f_0^μ is almost negligible compared to the cost of computing a subgradient of f_i for any i, the extra cost in practice would not be noticeable.

2.3.2 Asynchronous distributed method

The idea behind the asynchronous distributed method is essentially the same as in the serial method presented in the previous subsection, with the exception that subgradients of component functions are computed in parallel to the updates. Specifically, following [NBB01], we use the computation model presented in Fig. 2.1. In this computation model only the *Updating system* can

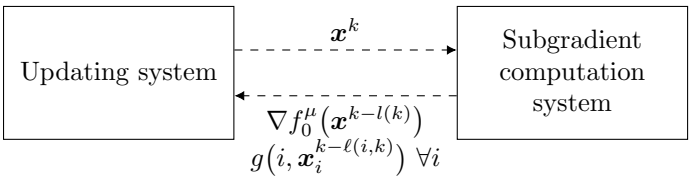


Figure 2.1: Computation model for the asynchronous distributed method. At each iteration k the *Updating system* communicates the current iterate $\boldsymbol{x}^k$ to the *Subgradient computation system*, which in turn, communicates back the last computed subgradient for each component function. The parallelism of this scheme resides within the *Subgradient computation system*.

increase iteration counters and updates are performed serially using the information provided by the *Subgradient computation system*. The serial nature of the *Updating system* allows us to describe the algorithm using a single iteration counter k.

Note that subgradients communicated by the *Subgradient computation system* to the *Updating system* might have been computed at previous iterates, in other words, the subgradient information available to the *Updating system* might have delays. We denote these delays by $l(k)$, the total number of updates since the last evaluation of the gradient of f_0^μ, at iteration k; and $\ell(j,k)$, the number of updates to block j since the last computation of its subgradient, at iteration k. Considering delays, we propose update rule (2.11) for the randomized block-coordinate subgradient method, where $j(k)$ is selected uniformly at random from $\{1,\dots,N\}$. This update rule is an extension of (2.10) where we allow delays in the subgradient information.

$$\boldsymbol{x}^{k+1} := \mathcal{P}_X\left[\boldsymbol{x}^k - \lambda_k \cdot I_{j(k)}^T\Big(I_{j(k)}\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big) + g\big(j(k), \boldsymbol{x}_j^{k-\ell(j(k),k)}\big)\Big)\right] \tag{2.11}$$

The presence of delays implies that Proposition 1 is no longer valid for update rule (2.11). Nevertheless, we can use essentially the same idea in order to prove convergence of the method defined by rule (2.11) as the one we used in the previous subsection, that is, to show that the expected update direction coincides with the direction of the approximate subgradient of the objective function and that the error in the approximate subgradient vanishes as the iterations advance. Our analysis is based on the Supermartingale Convergence Theorem (SCT)) [BT96, Prop. 4.2] and the following assumptions:

Assumption 1 (Subgradient boundedness)**.** *The subgradients of component functions are bounded above by some positive constants. In particular, there exist positive constants C and D such that*

$$\sup_{\substack{j\in\{1,\cdots,N\}\\ x,y\in X}} \left\|I_j\nabla f_0^{\mu}(\boldsymbol{x})+g\big(j,\boldsymbol{y}_j\big)\right\|_2 \le C \quad \textit{and} \quad \sup_{\substack{j\in\{1,\cdots,N\}\\ x\in X}} \left\|g\big(j,\boldsymbol{x}_j\big)\right\|_2 \le D.$$

Assumption 2 (Delay boundedness)**.** *There exist a positive integer L (possibly unknown) such that $l(k) \le L, \forall k = 1,\dots,\infty$ and $\ell(j,k) \le L, \forall j = 1,\dots,N, \forall k = 1,\dots,\infty$.*

Assumption 3 (Diminishing-bounded stepsize)**.** *The stepsize λ_k might be a function of $\boldsymbol{x}^k, \boldsymbol{x}^{k-1}, \dots, \boldsymbol{x}^0$, but not of the block coordinate to be updated $j(k)$. Further, the sequence $\{\lambda_k\}$ is bounded above and below by a deterministic sequence $\{\gamma_k\}$, such that*

$$\check{G}\gamma_k \le \lambda_k \le \hat{G}\gamma_k,\ \gamma_k = \frac{1}{(1+rk)^q} \quad \forall k, \quad \sum_{k=0}^{\infty}\gamma_k = \infty, \quad \sum_{k=0}^{\infty}\gamma_k^2 < \infty,$$

where $\check{G}, \hat{G}, r, q$ are positive constants.

For SUC instances, Assumption 1 is ensured by the boundedness of the sets $\mathcal{U}$ and $\mathcal{D}_i, i = 1,\dots,N$ of the primal problem (2.1) – (2.2), while Assumption 2 will hold as long as the evaluation time of all component functions is finite. The selection of a stepsize which is consistent with Assumption 3 is discussed in subsection 2.3.3.

The following lemma conveys the key idea of our analysis and the rest of the proof follows almost directly from it.

Lemma 1. *Let Assumptions 1 and 2 hold. Additionally, let J be a discrete uniform random variable on the set $\{1,\dots,N\}$ and $\mathcal{F}_k = \{\boldsymbol{x}^k, \boldsymbol{x}^{k-1}, \dots, \boldsymbol{x}^0\}$. Then, the expected direction of update rule* (2.11)*,*

$$\mathbb{E}[I_J^T(I_J\nabla f_0^{\mu}(\boldsymbol{x}^{k-l(k)}) + g(J, \boldsymbol{x}_J^{k-\ell(J,k)}))|\,\mathcal{F}_k],$$

coincides with the direction of an approximate subgradient of f at $\boldsymbol{x}^k$.

Lemma 1 shows that the update direction of rule (2.11) should, on average, be close to a subgradient of the objective and that the error in the subgradient is bounded by the sum of the last L stepsizes, hence by choosing a stepsize consistent with Assumption 3 this error will vanish as k grows. In the following, Proposition 2 gives an estimate of the progress of the asynchronous method at each iteration, based on the result of Lemma 1, while Proposition 3 presents a straightforward consequence of Assumption 3.

Proposition 2. *Let Assumptions 1, 2 and 3 hold and let* $\mathcal{F}_k = \{\boldsymbol{x}^k, \boldsymbol{x}^{k-1}, \ldots, \boldsymbol{x}^0\}$. *Then, for the sequence* $\{\boldsymbol{x}^k\}$ *generated by the update rule* (2.11), *we have that*

$$\mathbb{E}\big[\|\boldsymbol{x}^{k+1} - \boldsymbol{y}\|_2^2 | \mathcal{F}_k\big] \leq \|\boldsymbol{x}^k - \boldsymbol{y}\|_2^2 - 2\frac{\lambda_k}{N}\big(f(\boldsymbol{x}^k) - f(\boldsymbol{y})\big) + \lambda_k^2 C^2 + 2\frac{C^2 L_0^\mu}{N}\lambda_k \sum_{m=k-L}^{k-1} \lambda_m^2 + 4CD\lambda_k \sum_{m=k-L}^{k-1} \lambda_m. \tag{2.12}$$

Proposition 3. *Let Assumption 3 hold. Then, we have*

$$\sum_{k=0}^{\infty} \lambda_k = \infty, \quad \sum_{k=0}^{\infty} \lambda_k^2 < \infty, \quad \sum_{k=0}^{\infty} \lambda_k \sum_{m=k-L}^{k-1} \lambda_m < \infty, \quad \sum_{k=0}^{\infty} \lambda_k \sum_{m=k-L}^{k-1} \lambda_m^2 < \infty,$$

where for notational convenience we let $\lambda_{-l} = \lambda_0 \ \forall l \in \mathbb{N}$.

Finally, in the following theorem, we apply the SCT to the result of Proposition 2 in order to prove the convergence of the asynchronous method to an optimal solution.

Theorem 1. *Let Assumptions 1, 2 and 3 hold, and assume further that the optimal solution set* X^* *is nonempty. Then the sequence* $\{x_k\}$ *generated by the randomized method converges to some optimal solution with probability 1.*

The result presented in Theorem 1 extends the state-of-the-art by providing a convergence guarantee for the asynchronous block-coordinate descent method for non-differentiable optimization problems with the structure of problem (2.9) without the need for a line search on $\boldsymbol{x}_{j(k)}$ at each iteration [Tse01, FR13].

An important remark is that the asynchronous incremental method proposed in [NBB01] is guaranteed to converge to an optimal solution for both problem (2.5) and problem (2.9), while also exploiting their decomposable structure to a certain extent. We utilize block-coordinate descent instead of incremental methods for two reasons:

- The choice between incremental and block-coordinate subgradient methods can have significant implications on the magnitude of delays whenever we are minimizing a problem with the structure of problem (2.9), where a gradient of f_0^μ is much easier to evaluate than a subgradient of f_i for any $i = 1, \ldots, N$, and $X = \mathcal{R}^n$ (unconstrained optimization).

  The incremental subgradient method at iteration k selects at random a component function, $j(k) \in \{0, 1, \ldots, N\}$, and performs an update following the direction of the computed subgradient for the selected component

function. Whenever $j(k) \geq 1$, the algorithm will update block-coordinate $j(k)$, which will cause the current gradient of f_0^μ and the subgradient $f_{j(k)}$ to gain one unit of delay. On the other hand, every time $j(k) = 0$, the algorithm will update the entire vector $\boldsymbol{x}$, adding one unit of delay to all the subgradient information available to the *Updating system*. This effect, which is unavoidable for the problem structure analyzed in [NBB01], is undesirable because errors on the update direction depend directly on the magnitude of the delays.

The block-coordinate subgradient method at iteration k updates only the coordinates of block $j(k) \in \{1, \ldots, N\}$, causing the available gradient of f_0^μ and subgradient $f_{j(k)}$ to gain a unit of delay, but leaving unaffected the rest of the subgradient information available to the *Updating system*. Moreover, new gradients for f_0^μ can be computed very fast, even continuously as iterations advance, which allows us to maintain small delays throughout the solution process.

- For SUC instances and, in general, for Lagrangian relaxation of constraints linking duplicated variables, as the dual multiplier $\boldsymbol{x}$ approaches an optimal value, $\boldsymbol{u}$ and $\boldsymbol{v}_i$ start becoming similar for all i. As a consequence, the gradient of f_0^μ with respect to block i, i.e. $-\boldsymbol{u}^*(\boldsymbol{x})$ (the optimal solution of (2.8)), will tend to point in an opposite direction to the subgradient of f_i, i.e. $\boldsymbol{v} \in \partial f_i(\boldsymbol{x}_i) \equiv V_i^*(\boldsymbol{x}_i)$ (an optimal solution to (2.7)), for any $i = 1, \ldots, N$, thereby causing the incremental method to be susceptible to oscillations in $\boldsymbol{x}$ (incremental updates using the gradient of f_0^μ would move the iterates in the opposite direction to incremental updates using the subgradients of $f_i, i = 1, \ldots, N$).

2.3.3 Stepsize selection and function value estimation

Although Assumption 3 might seem to restrict the stepsize to a diminishing series of the type $1/k^q$, it also allows us to use a stepsize similar to the dynamic stepsize proposed by Polyak for the subgradient method [Pol69],

$$\lambda_k = p\frac{f(\boldsymbol{x}^k) - f^*}{\|g(\boldsymbol{x}^k)\|_2^2}, \quad g(\boldsymbol{x}^k) \in \partial f(\boldsymbol{x}^k), 0 < p < 2.$$

The original Polyak stepsize requires knowledge of the objective value at the current iterate $f(\boldsymbol{x}^k)$, the optimal value f^* and the norm of the subgradient at the current iterate $\|g(\boldsymbol{x}^k)\|_2$, none of which are available for the asynchronous method. Instead, we use estimates for each of the aforementioned quantities. An estimate of the current objective, which is also an upper bound on the objective of the primal problem (2.1)–(2.2), can be obtained at the cost of

evaluating f_0 as follows

$$f(\boldsymbol{x}^k) \approx UB_k := f_0\Big(\big[(\boldsymbol{x}_1^{k-\ell(1,k)})^T \dots (\boldsymbol{x}_N^{k-\ell(N,k)})^T\big]^T\Big) + \sum_{j=1}^{N} f_j(\boldsymbol{x}_j^{k-\ell(j,k)}), \tag{2.13}$$

while an estimate of the subgradient norm $\bar{g}_k$ can be computed using the last known subgradients for the component functions,

$$\|g(\boldsymbol{x}^k)\|_2 \approx \bar{g}_k := \max\left\{\sigma, \left\|\nabla f_0^{\mu}(\boldsymbol{x}^{k-l(k)}) + \sum_{j=1}^{N} I_j g(j, \boldsymbol{x}_j^{k-\ell(j,k)})\right\|_2\right\}$$

where σ is a small positive constant intended to prevent that $\bar{g}_k = 0$ (note that due to the delays, $\bar{g}_k = 0$ does not imply that $\boldsymbol{x}^k$ is optimal). An underestimate for the optimal value LB_k can be obtained from a feasible solution to the primal problem, computed as described in section 2.4. We assume that this is a strict underestimate, i.e. $\theta \leq f^* - LB_k$ for some $\theta > 0$, in other words, we assume that strong duality is never attained for realistic SUC instances.

Using these estimates, we propose the following dynamic stepsize,

$$\lambda_k = \frac{p}{(1+rk)^q} \cdot \frac{\min\{\xi,\, UB_k - LB_k\}}{\bar{g}_k^2}, \tag{2.14}$$

where p, r, ξ are positive constants and $1/2 < q \leq 1$. The goal of ξ is to prevent the method from taking long steps whenever the underestimate of the optimal value is loose. The proposed stepsize λ_k, as defined in equation (2.14), agrees with Assumption 3, as can be seen from the following inequality:

$$p\frac{\theta}{C} \cdot \gamma_k \leq \lambda_k \leq p\frac{\xi}{\sigma} \cdot \gamma_k.$$

Therefore, by Theorem 1, the asynchronous algorithm using the proposed stepsize will converge with probability 1 to an optimal solution.

A diminishing stepsize of type $1/k^q$ it is also guaranteed to achieve convergence. However, for the method to work effectively, it is necessary to determine a 'good' initial stepsize in absolute terms. The process of selecting a 'good' initial stepsize can require several trial-and-error runs of the algorithm for every instances to be solved, which would not be possible in an industrial implementation with a strict time limitation.

By contrast, in order to set the parameters for the proposed stepsize (2.14) (p, r and q) we only need to decide the proportion of the Polyak stepsize that we would like to have at two different iteration counts and the rate at which we would like to decrease this proportion (for instance, 50% at the beginning, 25% after $50N$ iterations, and decreasing with $1/k$, respectively). Given the

relative nature of these parameters, there is no need for trial-and-error runs when using this stepsize in industrial implementations.

2.4 Primal recovery

Primal recovery is an essential component of any Lagrangian relaxation aiming at solving the original problem. Although there exist exact methods for recovering primal solutions in the case of linear programs [AW09], these methods do not extend to the mixed integer case and primal recovery relies, generally, on heuristics in the latter case.

In the case of stochastic unit commitment instances, we exploit the relatively complete recourse of the problem, i.e. that given $\bar{\boldsymbol{v}}$ such that $(\bar{\boldsymbol{v}}, \boldsymbol{w}_i) \in \mathcal{D}_i$ for certain i and $\boldsymbol{w}_i$, then $\bar{\boldsymbol{v}} \in \mathcal{U}$ and $(\bar{\boldsymbol{v}}, \boldsymbol{w}_j) \in \mathcal{D}_j$ for some $\boldsymbol{w}_j$ for any $j = 1, \ldots, N$. In other words, any solution to a scenario subproblem $\bar{\boldsymbol{v}}$ can be used as a candidate non-anticipative first stage solution. We can then compute the second stage cost $h_i(\bar{\boldsymbol{v}})$ by solving second stage problems with fixed $\boldsymbol{v}_i = \bar{\boldsymbol{v}}$,

$$h_i(\bar{\boldsymbol{v}}) = \boldsymbol{c}_i^T \bar{\boldsymbol{v}} + \min_{w|(\bar{v},w)\in\mathcal{D}_i} \boldsymbol{d}_i^T \boldsymbol{w}$$

obtaining a complete primal solution to (2.1) – (2.4) and a lower bound on the objective of (2.5) [Ahm13, Ahm15].

Recovering one feasible non-anticipative solution at every dual iteration would require the solution of N^2 second stage MILPs for every dual pass over data. For medium to large scenario sets, the computational requirements of primal recovery can easily become larger than the requirements of the dual algorithm. If both dual iterations and primal recovery are performed concurrently, primal candidates would need to enter into a queue for evaluation, which will typically grow as dual iterations advance (assuming similar resources are allocated for dual iterations and primal recovery). Within this context, we test three rules for determining the order of evaluation of primal candidates in the queue:

- First-in-first-out (FIFO).
- Random order (RND), motivated by the possibility that a good solution might appear anywhere in the sequence of solutions to scenario subproblems.
- Last-in-first-out (LIFO), an approach that takes into account that as dual iterations advance, solutions to scenario subproblems tend to be almost non-anticipative ($\boldsymbol{v}_i = \boldsymbol{v}_j$, $i, j = 1, \ldots, N$), therefore scenario subproblem solutions in later iterations could have better overall performance.

A different approach towards recovering primal solutions is to create primal candidates as they are required, combining the solutions to different scenario subproblems. Carøe and Schultz [CS99] propose creating primal candidates by first averaging the solutions to all scenario subproblems and rounding the result using a heuristic. We use a variant of this idea combined with importance sampling (IS) to create primal solutions. Our recovery heuristic proceeds as follows:

1. Associate to each scenario a probability proportional to its estimated importance, e.g. $p_i \propto f_i(\boldsymbol{x}_i^{k-\ell(i,k)}), i = 1, \ldots, N$
2. Pick a sample of the scenarios of size $M < N$, using p_i as the probability of sampling scenario i
3. Average the current scenario subproblem solution $\bar{\boldsymbol{v}}_i$ associated to the sampled scenarios $\{i(m), m = 1, \ldots, M\}$ in order to generate an average $\tilde{\boldsymbol{v}}$
$$\tilde{\boldsymbol{v}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{v}_{i(m)}$$
4. Generate a primal candidate $\bar{\boldsymbol{u}}$ by projecting the average of step 3 onto $\mathcal{V} = \{\boldsymbol{v} \in \mathcal{U} \mid \exists w_i, (v, w_i) \in \mathcal{D}_i, i = 1, \ldots, N\}$,
$$\bar{\boldsymbol{u}} = \operatorname*{arg\,min}_{u \in \mathcal{V}} \ \|\boldsymbol{u} - \tilde{\boldsymbol{v}}\|_2^2 \,. \tag{2.15}$$

This method allows us to create primal candidates that combine the characteristics that make a solution optimal for a representative subset of the scenarios while, at the same time, generating different candidates after each dual iteration if necessary. The latter would not be possible if we were to follow [CS99], since the average of the solution to all scenario subproblems does not change significantly from one coordinate descent iteration to the next.

2.5 High performance computing implementation

Section 2.3 provides convergence guarantees for the asynchronous dual optimization algorithm, based on the conceptual distributed computation model of Fig. 2.1, and section 2.4 two parallelizable primal recovery schemes. This section specifies the actual implementation of the algorithm, that is, the different processes running in parallel and the information to be exchanged between them.

We implement the algorithm using the *Master/Slave* design presented in Fig. 2.2. The *Master* coordinates the work of all processes, dynamically assigning tasks (solving optimization problems) to *Slaves* as the algorithm progresses.

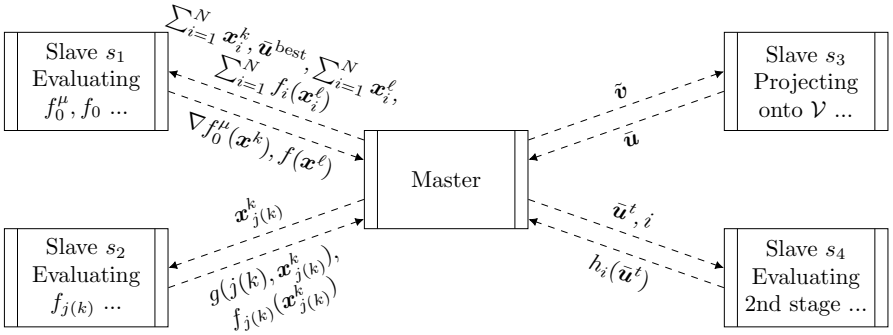


Figure 2.2: Execution snapshot of the asynchronous distributed algorithm for stochastic unit commitment. Each box correspond to a process and each dashed line correspond to information exchanged between processes. The *Master* dynamically assigns tasks to each of the *Slaves*. Not all types of tasks need to be present at all times and there might be several *Slaves* engaged on the same task but over different data.

Slaves, on the other hand, limit themselves to perform the tasks demanded by the *Master*, without having view of the global progress of the algorithm.

In contrast to the conceptual computation model presented in Fig. 2.1, in the actual implementation there is no clear separation between the *Updating system* and the *Subgradient computation system*. The *Updating system* is contained within the *Master* process, while the *Subgradient computation system* is split between *Master* and the *Slaves* currently evaluating f_0^μ or $f_i, i = 1, \ldots .N$, shown to the left of the *Master* in Fig. 2.2.

Primal recovery is performed concurrently with dual iterations using a portion of the *Slaves*, shown to the right of the *Master* in Fig. 2.2. Primal recovery evaluates the second-stage cost of primal candidates and, if using the IS heuristic, it also creates new candidates by projecting averaged first-stage solutions onto the first-stage feasible set.

Performing dual iterations alongside primal recovery enables the algorithm to continuously compute upper bounds (dual function evaluations) and lower bounds (primal recovery) on the optimal value. This allows us to establish a natural termination criterion, $UB - LB \leq \epsilon$, as well as to terminate the algorithm early at a certain wall time or number of dual iterations, returning the incumbent solution and the current optimality gap, as MILP solvers do.

In the following we detail the internal layout of the *Master* and *Slave* processes, and how they interact with each other. The implementation is based on the SMPS file format for stochastic programs [GS01]. In particular, it uses the concepts of CORE problem, time indexation of variables and constraints (TIME file), and the specification of scenarios by their differences to the CORE problem (STOCH file). In order to maintain a small memory footprint, which is

critical for solving large SUC instances, only the CORE problem and the TIME indexes are maintained in memory, while the information in the STOCH file is loaded from the hard drive as needed and purged after it has been used.

2.5.1 Slave

The *Slave* process, presented in Fig. 2.3, starts by partially reading the instance to be solved (steps 1-2, Fig. 2.3): it loads the CORE problem, loads all the information in the TIME file (time stages of variables and constraints) and gathers information about the organization of the STOCH file (metadata), e.g. a list of the scenarios and where in the STOCH file are they located.

The reading process respects the classification of constraints within the CORE file, in particular, it differentiates between normal constraints, *delayed constraints* (i.e. constraints that are necessary for feasibility but are unlikely to be binding, also known as *lazy constraints*) and *model constraints* (i.e. constraints redundant at the optimal MIP solution, also known as *user cuts*)[4]. Current commercial MILP solvers can take advantage of this classification of constraints to speed up the solution process. Note that, in order to read files with this constraint differentiation, the TIME file must explicitly declare the time index of rows and columns [GS01].

After every process finishes the reading step (step 3), the control flow is organized around a loop within which the *Slave* receives a task from the *Master*, executes it and communicates back the result. Subproblems in all tasks are formulated either by modifying the CORE problem, as done in steps 6, 10 and 22, or by taking a subset of the constraints of the CORE problem, as done in steps 14 and 18. The transformation of steps 6, 10 and 22 uses the metadata to avoid parsing unnecessary parts of the STOCH file.

There are 5 types of tasks, as well as 1 termination signal, that the *Slave* can receive from the *Master*. Among these, *dual scenario* corresponds to evaluating a certain component $j \in \{1, \ldots, N\}$ of the dual objective for certain multipliers, *primal projection* corresponds to projecting an averaged candidate onto the feasible set of first stage decisions (see equation (2.15)) and *second stage scenario* corresponds to solving a recourse problem for a given first stage decision.

The *dual* f_0 task involves the following two actions: computing the gradient of f_0^μ and computing an upper bound. These tasks are merged because they involve solving very similar mathematical programs, (2.6) and (2.8), none of which has a strict requirement on the frequency with which it must be solved (contrary to the case of *primal projection*, which must be solved whenever we need a new primal candidate). The task uses four pieces of data. The first two, the sum of the current multipliers $\sum_{i=1}^{N} \boldsymbol{x}_i$ and a center $\boldsymbol{u}_0$, are used to compute

[4] *delayed constraints* and *model constraints* correspond to the terminology used by Xpress, while *lazy constraints* and *user cuts* is the terminology used by Cplex.

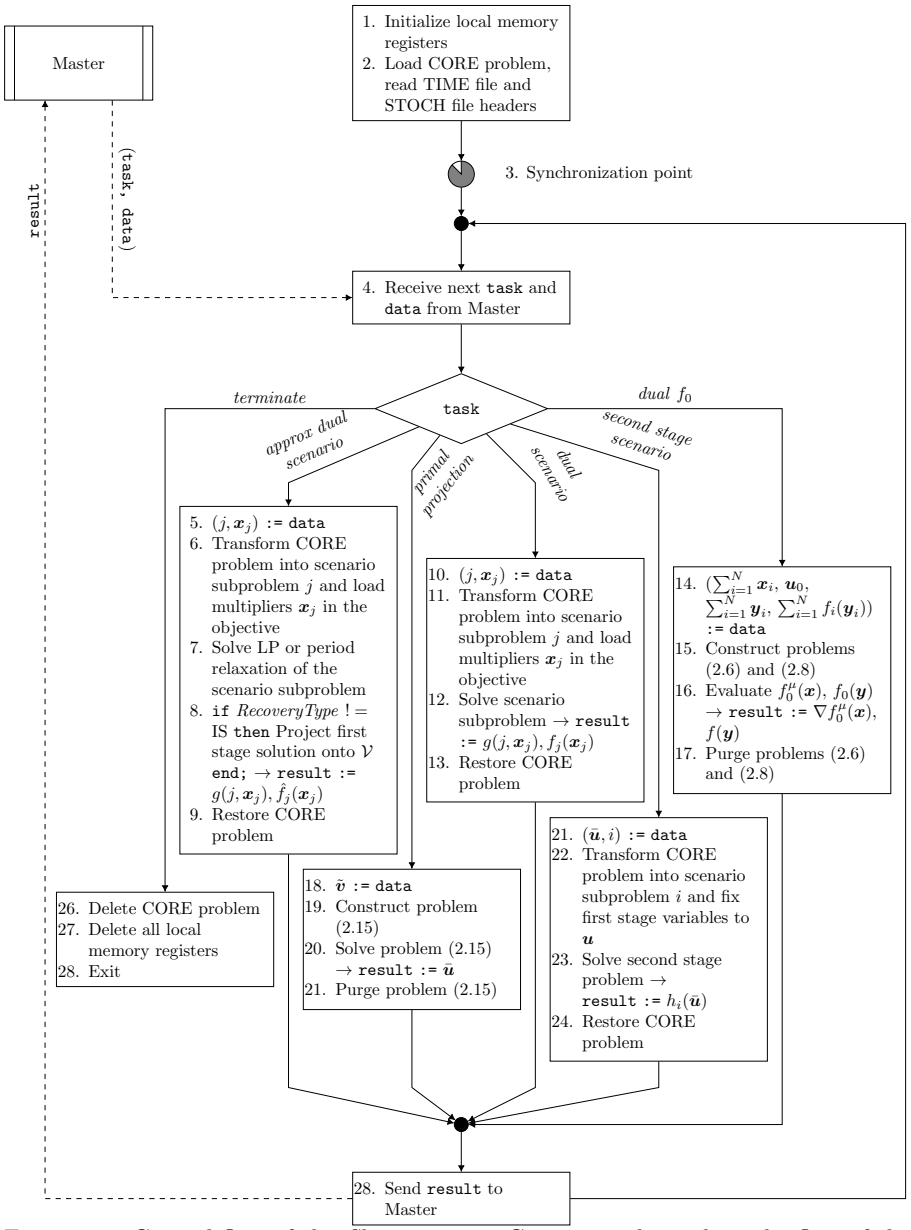


Figure 2.3: Control flow of the *Slave* process. Continuous lines show the flow of the program, while dashed lines indicate exchange of information with other processes. Italics denote constants and parameters. The iteration counter k and the candidate index t are dropped because they are not relevant within the *Slave*.

the gradient of f_0^μ at the current iterate. The other two, the sum of certain multipliers $\boldsymbol{y}$, $\sum_{i=1}^{N} \boldsymbol{y}_i$, and the sum of the scenario component functions of the dual evaluated at $\boldsymbol{y}$, $\sum_{i=1}^{N} f_i(\boldsymbol{y}_i)$, are used to obtain an upper bound on the optimal value of the original program by evaluating $f_0(\boldsymbol{y})$, as indicated in equation (2.13).

The *approx dual scenario* task has the same objective as the *dual scenario* task, with the difference that the former solves only a relaxation of the scenario subproblem. Two types of relaxation are considered, the linear programming (LP) relaxation and the period relaxation, which solves each period of the scenario subproblem independently. The period subproblem is constructed using the indexation of variables and constraints present in the TIME file. These relaxations provide cheap subgradient estimates and upper bounds on component functions. They are used in the initialization procedure described in subsection 2.5.3.

2.5.2 Master

The *Master* process, presented in Fig. 2.4 and Fig. 2.5, starts in the same manner as the *Slave*, by allocating memory to variables and reading the necessary information from the CORE, TIME and STOCH files, steps 1-2 in Fig. 2.4. Note that the *Master* does not require further information regarding the subproblems because this is not used directly for dual iterations or primal recovery.

After reaching the synchronization point, step 3, the *Master* uses the *Slaves* to perform the initialization procedure, step 4, which provides with initial values for the subgradients and upper bounds. This step is followed by the launching of the initial batch of tasks of the algorithm, in steps 5 and 6. We update as many blocks as possible, launching the corresponding subgradient evaluation tasks. If there are free *Slaves* after updating all blocks, we use them to perform primal recovery tasks, so that all *Slaves* are assigned a task. The *Master* keeps track of the task assigned to each *Slave*.

The *Master* then enters its main loop, which receives the result of a task from *Slave s* (step 7), process it (steps 8-16) and assigns a new task to the *Slave s* (steps 21-28). The processing procedure depends on the type of task. For *dual* f_0, we simply overwrite the gradient of f_0 and update the upper bound, while for *primal projection* we add the new candidate to the list of primal candidates.

The processing of *dual scenario* tasks requires checking whether the received result contains new information relative to what is already available to the master before overriding it (step 11). This is necessary, because as the updates of $\boldsymbol{x}$ are performed at random, there might be two or more *Slaves* evaluating the same component function i, each for a different $\boldsymbol{x}_i$. If the evaluation for an older $\boldsymbol{x}_i$ finishes later, it should not overwrite the subgradient information available to the `Master`, since this would only introduce more delays in the

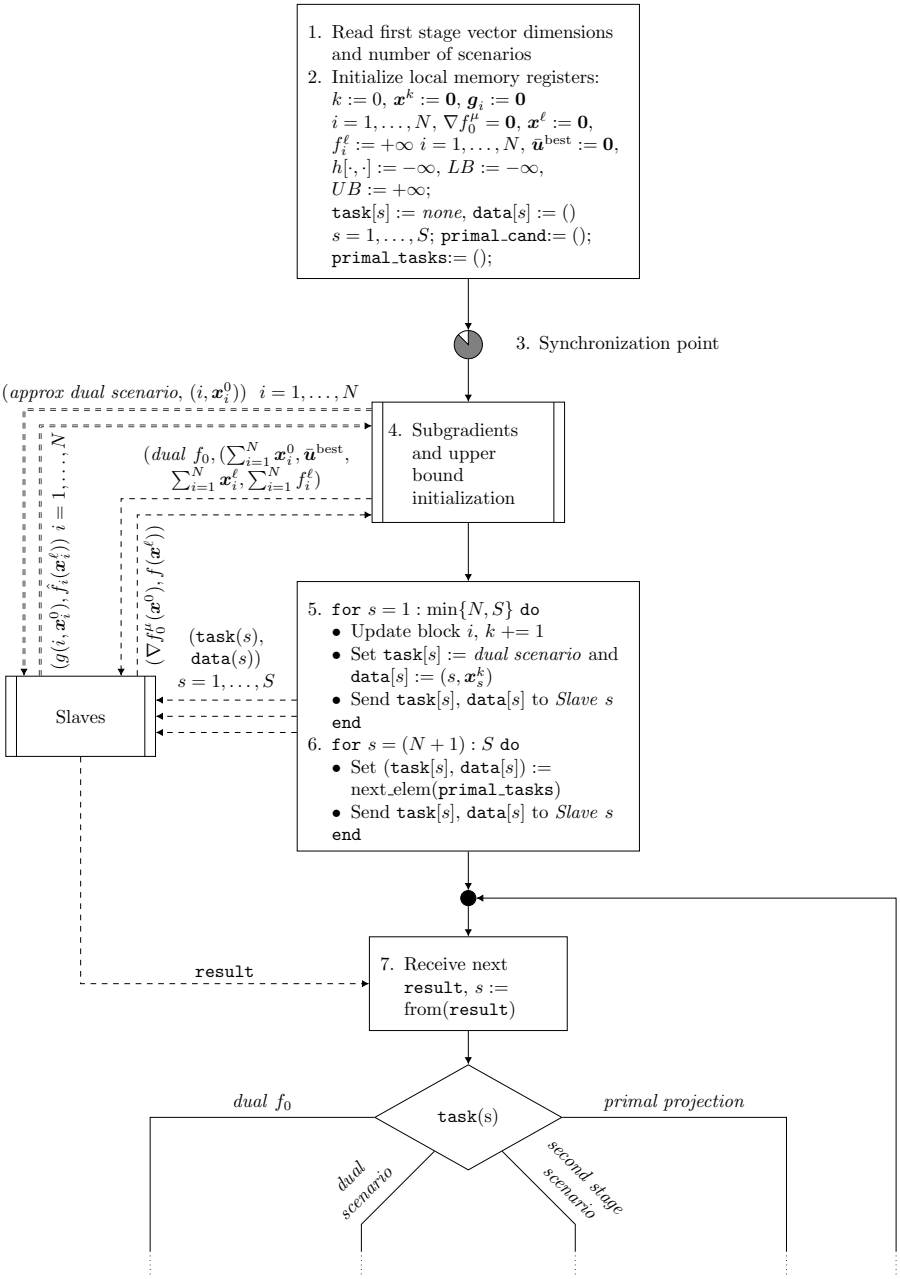


Figure 2.4: Control flow of the *Master* process.

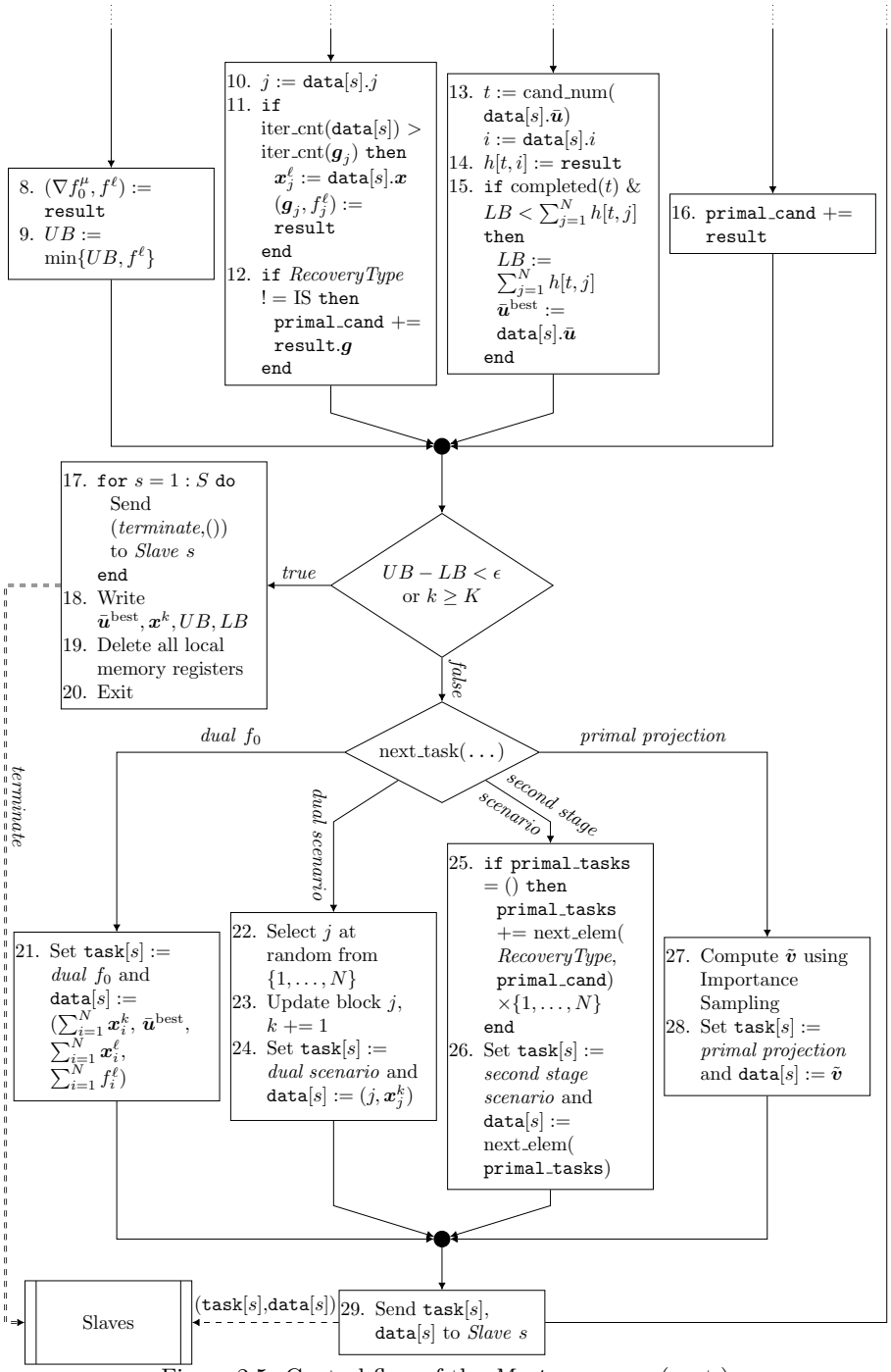


Figure 2.5: Control flow of the *Master* process (cont.).

subgradient information. Irrespective of its delay, the result is used in step 12 as a primal candidate as long as the primal recovery heuristic is not set to importance sampling.

second stage scenario tasks, on the other hand, return the second stage cost of a given first stage candidate solution at a certain scenario (step 14). If the first stage solution has been evaluated for all scenarios, it is used in step 15 to update the lower bound on the primal objective. Note that we do not assume any order in the evaluation of first-stage candidate solutions, allowing for primal recovery to be performed asynchronously on several candidates at the same time.

Once the processing phase is completed, at the top of Fig. 2.5, the algorithm checks whether the termination criterion is met and terminates the execution of the *Master* and *Slaves* if that is the case. If the termination criterion is not met, the algorithm will proceed to decide which task to assign to *Slave s*. This decision is made according to the following criteria, in the same order of importance as they are presented:

1. Use at most N concurrent processes for *dual scenario* tasks.
2. Maintain the proportion of *Slaves* engaged in dual iterations as close as possible to the value of the configuration parameter *DualShare*.
3. A *dual* f_0 task must be executed every certain number of *dual scenario* tasks.
4. If using the IS primal recovery heuristic, a *primal recovery* task must be executed whenever the number of elements left in `primal_tasks` is deemed small[5].

Before sending the chosen task to *Slave s*, certain preprocessing steps are required. In the case of a *dual scenario* task, a block coordinate descent update is performed on a random scenario j and, only then, a *dual scenario* task for scenario j with the new multipliers is assigned to *Slave s*. This action (steps 22-23) correspond to the updating system of Fig. 2.2.

Before assigning a *second stage scenario* task, on the other hand, it is necessary to check whether there are pending primal tasks to be executed. If this is not the case, then, in step 25, new tasks are created using the next candidate, which is chosen according to the order specified by the *RecoveryType* parameter (FIFO, RND, LIFO, IS) from the list of pending candidates, `primal_cand`, and removed from it. If the *RecoveryType* configuration parameter is set to IS, then the most recent candidate in the list is chosen.

The assigned task is then sent to *Slave s* and the *Master* returns to the beginning of its main loop, where it will wait for the next result.

[5]The definition of small here depends upon the number of *Slaves*, the time between received results and the time it takes to complete a *primal recovery* task.

2.5.3 Initialization

The main objective of the initialization subroutine, executed by the *Master* at step 4 (Fig. 2.4), is to obtain cheap estimates of the subgradients of component functions, upper bounds on the component function values at the initial multipliers $\boldsymbol{x}^0$ and, optionally, an initial set of primal candidates.

The initialization proceeds as follows. First, subgradient estimates and upper bounds are computed by executing *approx dual scenario* using the *Slaves*, for all scenarios with the initial multipliers. Once the results for all scenarios have been collected, a *dual* f_0 task is executed to obtain a gradient for f_0^μ and a valid upper bound on the primal objective.

If the selected primal recovery heuristic is not IS, subgradient estimates can be used as primal candidates (because of the projection onto $\mathcal{V}$, step 8, Fig. 2.3) and evaluated at step 6 of Fig. 2.4 (if $S > N$). On the other hand, if the selected primal recovery heuristic is IS, then at step 6 of Fig. 2.4 the next primal tasks would correspond to *primal projection*, with averaged candidates generated by the *Master* on the fly.

Without these subgradient estimates, the first round of updates (step 5, Fig. 2.4) would not modify the multipliers, and the computation of the first upper bound would be delayed until all component functions have been evaluated, that is, at least until the slowest of all component functions is accurately evaluated. Considering that differences in evaluation times of component functions observed in real instances can be as high as 7500%, the lack of an initial upper bound can significantly delay termination of the algorithm, even when a good primal candidate and lower bound are already available.

Note that primal recovery also benefits from the initialization, since without the initial set of candidates, primal recovery would be delayed until results from accurate evaluation of component functions f_i are returned to the *Master*.

2.6 Numerical results

We implement the proposed asynchronous algorithm as described in the previous section in C, using Xpress (through its C API) [Fai16] for solving all mathematical programs and MPI [For15] for handling communications between processes. We configure Xpress to solve the root node in all subproblems using the barrier algorithm and we set the termination gap of subproblems to 1%.

We test the proposed algorithm on instances of the WECC [POR15] and CWE (section 1.4.3) systems. Sizes and solution times of scenario subproblems are presented in Table 2.1, where it can be observed that the instances used in the present study are at least as large, or present scenario subproblems as difficult to solve (in terms of solution time) as the models in the literature. Numerical experiments were run on the Cab cluster of the Lawrence Livermore National Laboratory (LLNL). Each node of the Cab cluster is equipped with

Table 2.1: Scenario subproblem sizes and solution times for different instances used in the present study (highlighted in boldface) and in the literature. Subproblem solution times for WECC [POR15] correspond to winter weekend instance with 100 scenarios, while subproblem solution times for CWE correspond to spring weekday instance with 120 scenarios.

Instance	Rows	Columns	Non-zeros	Integers	Subproblem solution time [s], avg. (max.)	
WECC [CGSM+15]	69 447	28 943	240 724	4 080	9.4	(25.7)
WECC [POR15]	34 441	23 090	139 394	3 074	8.3	(67.9)
EDF [vAM16]	812 906	73 562	–	26 122	–	–
CWE	609 589	390 075	1 941 270	9 753	3 383.2	(7 851.8)

two Intel Xeon E5-2670 processors (16 cores per node) and 32GB of RAM memory.

2.6.1 Western Electricity Coordinating Council system instances

The WECC system instance [POR15] is composed of 130 thermal generators, 182 nodes and 319 lines. It features multiarea renewable production with hourly resolution over a 24 hour horizon for 8 representative day types, one weekday and one weekend day per season. The number of scenarios ranges from 10 to 1000, and each scenario is associated with different renewable production profiles and contingencies.

The solution times of these instances are summarized in Table 2.2 for different configurations of the asynchronous algorithm. Regarding stepsizes, 'Dim. 1/k' corresponds to a stepsize of the type $1/k$ and 'Polyak' corresponds to the Polyak stepsize defined in equation (2.14), where the diminishing part is set to decrease from 0.5 to 0.25 in $50N$ iterations with $q = 1$. Primal solution recovery methods correspond to the four methods described in section 2.4. We use 1 node for solving the 10-scenario instances, 10 nodes for solving the 100-scenario instances and 16 nodes for solving the 1000-scenario instances. For all WECC instances we use 1 core per process, so that $S = \#\text{Cores} - 1$, and we limit the run time of the algorithm to 2 hours. The modification of the *DualShare* parameter for 1000-scenario instances is motivated by the findings covered in subsection 2.6.3.

From the perspective of dual optimization, Polyak stepsizes outperform $1/k$ stepsizes when considering a 1% termination criterion. In all our instances and test runs, Polyak stepsizes provided better results without the need for tuning.

For primal recovery, we observe that the three methods that recover solutions directly from solutions to scenario subproblems (FIFO, RND, LIFO) exhibit similar performance for the 10-scenario instances. RND and LIFO outperform FIFO on instances with 100 scenarios. The IS heuristic outperforms

Table 2.2: Solution time statistics for WECC instances, over 8 representative day types. Statistics for configurations that failed to achieve the target optimality gaps, within the limit wall time, for one or more day types are not reported and are denoted with a dash.

N	Step size	*Recovery Type*	# Cores	*Dual Share*	Solution time [s], avg. (max.)			
					2% optimality		1% optimality	
10	Dim. 1/k	FIFO	16	0.5	228.1	(792.9)	–	
	Dim. 1/k	RND	16	0.5	229.4	(856.1)	–	
	Dim. 1/k	LIFO	16	0.5	200.6	(739.7)	–	
	Dim. 1/k	IS	16	0.5	178.0	(638.0)	–	
	Polyak	FIFO	16	0.5	148.2	(469.3)	424.4	(1 361.1)
	Polyak	RND	16	0.5	117.8	(392.6)	–	
	Polyak	LIFO	16	0.5	131.2	(446.2)	384.0	(1 326.9)
	Polyak	**IS**	**16**	**0.5**	**118.4**	**(441.4)**	**364.7**	**(1 291.5)**
100	Polyak	FIFO	160	0.5	267.6	(325.1)	–	
	Polyak	RND	160	0.5	113.2	(345.6)	534.2	(1 134.1)
	Polyak	LIFO	160	0.5	99.4	(268.3)	508.9	(1 152.4)
	Polyak	**IS**	**160**	**0.5**	**95.5**	**(289.8)**	**517.9**	**(1 126.1)**
1000	Polyak	LIFO	256	0.75	723.9	(2 155.2)	–	
	Polyak	**IS**	**256**	**0.75**	**411.7**	**(1 354.5)**	**2 535.0**	**(6 427.0)**

its counterparts in all instances.

The best configuration of Table 2.2, which is highlighted in boldface, outperforms the run time reported in [POR15], in which 1000 scenario instances (23.1 million variables, 35.9 million constraints and 3.1 million integers) were solved in up to 24 hours using 1000 processors. Instead, the best performing algorithm in this chapter solves the same instance in less than 2 hours using 256 processors. Similar speedups are observed for instances with fewer scenarios with respect to [POR15]. The method of Cheung *et al.* [CGSM$^+$15] solves instances of the WECC with up to 100 scenarios within 1.5–2.5% optimality within 25 minutes. In comparison, the proposed algorithm solves instances of similar difficulty in terms of solution time of subproblems (see Table 2.1) to 1% optimality in at most 18.7 minutes and to 2% suboptimality in less than 5 minutes. Detailed solution statistics per day type for the best configuration are reported in Table 2.6, in section 2.B at the end of this chapter.

2.6.2 Central Western European system instances

The CWE system instance corresponds to one described in section 1.4.3. We model SUC using a hybrid time resolution: hourly commitment decisions and quarterly dispatch decisions. We consider the commitment of nuclear and slow thermal units (conventional generators with a minimum up and down time

Table 2.3: Solution time statistics for CWE instances, over 8 representative day types. All instances use the Polyak stepsize the IS primal recovery heuristic and a *Dual Share* parameter setting of 0.75.

N	# Cores	Solution time [s], avg. (max.)			
		2% optimality		1% optimality	
30	96	2 580.3	(5 908.2)	3 806.2	(9 279.1)
60	192	2 563.7	(5 593.3)	3 774.2	(8 323.4)
120	384	2 696.5	(5 973.0)	3 876.2	(7 952.6)

greater than 3 hours), and the set point for CHP power plants[6] (continuous variables) as first-stage decisions of SUC. The second stage includes the commitment of fast thermal generators (with minimum up and down time less than or equal to 3 hours) along with all continuous dispatch decisions. Each instance covers a 24-hour horizon (96 quarters). We refer the reader to section 4.C, in the final chapter of this dissertation, for a detailed description of this hybrid-resolution SUC model.

The problem is solved for 8 representative day types and using 30, 60 and 120 scenarios of renewable production. For all CWE instances, ramp rate constraints within each hour are declared as delayed constraints. Each scenario subproblem of the CWE instances is almost one order of magnitude larger, in terms of matrix size, than the scenario subproblems of the largest instance considered in the SUC literature [vAM16], see also Table 2.1.

Table 2.3 presents the summarized solution statistics for the proposed algorithm on the CWE instances. We use 6 nodes for the 30-scenario instances, 12 nodes for the 60-scenario instances and 24 nodes for the 120-scenario instances. Each process uses 2 cores, we set the MIP time limit to 1 hour and 30 minutes and the maximum run time of the algorithm to 6 hours. Detailed solution statistics per day type are reported in Table 2.7 of appendix 2.B.

Solution times are larger that those observed for the WECC, nevertheless they remain within operationally acceptable time frames for day-ahead scheduling (at most 2 hours and 34 minutes are required for obtaining a solution within 1% optimality). This increase in overall solution time with respect to the WECC instances results mostly from the time required for solving dual scenario subproblems which, as shown in Table 2.1, is two orders of magnitude larger than for WECC.

[6]Combined Heat and Power (CHP) plants can vary their production of electricity only within a limited range for a given production of heat [DA15].

Table 2.4: Variation of solution time with the *Dual Share* parameter setting, which vary smoothly between its start value ($k = 0$) and its value after 200 dual passes over data ($k = 200N$). Results in the table correspond to WECC, spring weekdays, 100 scenario instance, solved using 8 nodes (96 cores) and a Polyak stepsize. Statistics are obtained over 4 runs for each configuration.

Primal	*Dual Share*		Solution time [s], avg. (max.)			
Recovery	$k = 0$	$k = 200N$	2% optimality		1% optimality	
IS	0.1	0.1	150.5	(168.7)	1731.9	(1821.0)
	0.25	0.25	78.9	(82.8)	785.7	(807.8)
	0.5	0.5	52.5	(55.6)	441.4	(467.6)
	0.75	0.75	67.2	(78.0)	307.4	(333.3)
	0.9	0.9	58.1	(72.3)	291.0	(294.2)
LIFO	0.5	0.5	97.5	(120.0)	529.9	(598.3)
	0.75	0.75	92.3	(104.5)	479.9	(624.1)
IS	0.75	0.25	50.9	(59.0)	313.9	(334.5)
	0.9	0.1	66.8	(77.5)	280.3	(320.2)

2.6.3 Sensitivity of solution times to the allocation of resources

The *Dual Share* configuration parameter determines how distributed computing resources are allocated between tasks related to dual iterations or to primal solution recovery, which can significantly impact the overall performance of the algorithm, particularly when parallel computing resources are limited.

Table 2.4 shows how varying the *Dual Share*, while keeping the rest of the parameters constant, affects run time. We select the WECC spring weekday instance with 100 scenarios to perform this test because it corresponds to a medium-size instance and has the median solution time among WECC instances with 100 scenarios.

The configuration using the IS primal recovery heuristic (first 5 rows) exhibits an important improvement in solution time as we increase the *Dual Share*. For the LIFO heuristic (6th and 7th rows), on the other hand, we observe only a minor improvement in solution time with the increase of *Dual Share*. This shows that averaging scenario subproblem solutions can indeed generate better candidates than simply using scenario subproblem solutions, avoiding the use of computing power in evaluating the performance of low-quality primal candidates. This effect was not observed in the 10 and 100-scenario instances of Table 2.2 because we used almost twice as many processors as scenarios, allowing FIFO, LIFO and RND to carry out a very large number of candidate evaluations.

Our implementation gives us the freedom to change the *Dual Share* during the solution of an instance, allocating fewer or more resources to dual tasks in

earlier iterations. This can be beneficial as the subgradient method achieves fast improvements during the first iterations, but it becomes slow as it approaches the optimal solution. Therefore, by starting with a high *Dual Share* and gradually decreasing it, we can take advantage of the rapid bound improvement during the first iterations and use more computing power in later stages in order to recover better primal solutions. Rows 8th and 9th present the results of applying this idea. The improvements with respect to maintaining a fixed resource allocation, as done in all other rows of the table, are modest, at best, because of a combination of factors. During the first iterations, the Polyak stepsize cannot be computed accurately due to the lack of a good lower bound (see equation (2.14)) and the updates tend to overshoot. On the other hand, during the last iterations, the IS heuristic has difficulties to find new primal candidates, because the frequency at which we obtain solutions to scenario subproblems decreases with the *Dual Share*.

2.6.4 Parallel computing performance

The performance of a parallel algorithm can be measured against various metrics. We focus on (*i*) how the proposed algorithm compares to a synchronous algorithm, i.e. an algorithm that uses a synchronous method to carry out dual iterations, and (*ii*) how effectively the proposed algorithm scales up with the number of processors.

Solving the instances used in this work with a fully synchronous algorithm may require excessive use of computational resources [PO13]. In order to avoid using additional computing time for this comparison, we compare the proposed algorithm to a synchronous algorithm in terms of idle time of processors. This can be estimated using the solution times of subproblems, information which is already available to us from the previous sections. We consider a synchronous algorithm that solves the dual problem with the subgradient method using $\min\{N, \textit{Dual Share} \times S\}$ slaves (rounded to the closest integer, if necessary), and that performs primal recovery synchronously using the remaining slaves.

Table 2.5 presents the estimated idle times for the described synchronous algorithm. Note that idle times generally increase with the number of processors, since with fewer processors tasks can be stacked. In order to obtain a measure of the variation of idle times with the number of processors, Table 2.5 includes estimates for the same number of processors as used by the asynchronous algorithm, half the processors used by the asynchronous algorithm, and $2N + 1$ processes, i.e. the *Master*, N *Slaves* dedicated to dual iterations and N *Slaves* dedicated to primal recovery.

It can be directly observed that synchronous schemes tend to underutilize parallel computing infrastructure, leaving processors idle up to 80.4% of the time. For the same problem, the synchronous algorithm achieves almost zero idle time (the percentage of time dedicated to solving mathematical programs

Table 2.5: Estimated idle time of processors when solving SUC using a parallel synchronous algorithm. Average and maximum over 8 day types.

System	N	# Cores	*Dual Share*	Synchronous idle time [%], avg. (max.)					
				Same conditions		Half # cores		$1 + 2N$ processes	
WECC	10	16	0.5	48.0	(53.2)	28.3	(34.5)	54.2	(59.2)
	100	160	0.5	70.6	(80.4)	52.9	(65.0)	74.3	(83.8)
	1000	256	0.75	37.6	(57.0)	21.1	(36.9)	85.4	(92.4)
CWE	30	96	0.75	36.4	(47.0)	27.1	(33.9)	46.5	(53.9)
	60	192	0.75	43.6	(62.2)	33.4	(47.3)	53.3	(67.2)
	120	384	0.75	46.9	(61.2)	36.3	(46.6)	54.7	(65.4)

never falls below 97%, see Table 2.6 and 2.7 for details). Decreasing or increasing the number of processors does not change this observation for the synchronous scheme.

In order to determine how well the proposed algorithm scales with respect to the number of processors, we numerically estimate its parallel efficiency (i.e. speedup divided by the number of processors), presented in Fig. 2.6. Our baseline is the serial execution of the asynchronous algorithm, i.e. the method of section 2.3.1, with delays in the gradient of the smooth part of the objective and interleaving subgradient evaluation with primal recovery.

We observe three main effects limiting the parallel performance of our algorithm:

1. The existence of a *Master* process, that does not perform any other work than coordination, limits the parallel performance to $1 - 1/(S + 1)$ in expectation.

2. The asynchronous execution of our algorithm can cause delays on the subgradients used to update the dual multipliers, which can led to errors in the update direction[7]. As the magnitude of the delays tends to increase with the number of processors, this effect deteriorates the parallel efficiency of the dual method progressively with the number of processors.

3. Our primal recovery heuristic IS scale efficiently with the number of processors. The candidates proposed by the IS heuristic tend to be of very good quality, but the quality of these solutions does not improve drastically by recovering more candidates, because the heuristic is bounded by the candidates provided by the evaluation of dual scenario subproblems.

These three effects explain the parallel efficiency curves presented in Fig. 2.6 as follows. In all cases, the parallel efficiency for a small number of processes

[7]This problem is common to all asynchronous gradient-type of methods. It is known in the literature as *gradient staleness* and it is currently under study in the machine learning community [ZMW+17].

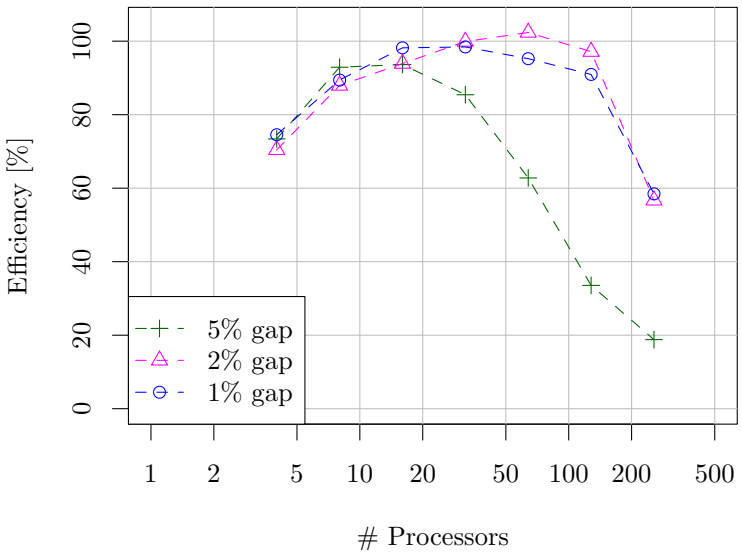


Figure 2.6: Parallel efficiency plot of the asynchronous algorithm for different termination optimality gaps. Plot drawn using WECC, spring weekdays, 100 scenarios, *Dual Share* 0.75, Polyak stepsize and IS primal recovery with 4, 8, 16, 32, 64, 128 and 256 processes. Wall times are obtained by averaging 4 runs for each processor count presented in the plot.

(smaller than $0.2N$) is limited by effect 1. The efficiency curve for 5% optimality quickly drops after $0.2N$ mainly due to effect 3: it is very easy for the IS heuristic to obtain a solution that is less than 5% suboptimal. The efficiency curves for 2% and 1% optimality exhibit a similar behavior: the curves drop first because of effect 2 and then due to a combination of effects 2 and 3. The difference between these two curves is explained by the extent to which delays affect dual convergence for different levels of accuracy, i.e. as delays increase we can still terminate with 2% optimality, but it becomes more challenging to terminate with 1% optimality because of subgradient errors. This causes the parallel efficiency for 1% optimality to start decreasing at a smaller number of processors than the curve for 2% optimality.

2.7 Conclusions

We propose an asynchronous dual decomposition algorithm for stochastic unit commitment, in which dual iterations are performed using a block-coordinate subgradient method, for which we provide convergence guarantees. We also propose primal recovery heuristics and present a high performance computing implementation of the algorithm. The algorithm is able to solve all instances of WECC and CWE within operationally acceptable time frames and exhibits parallel efficiency above 90% when using between $0.1N$ and $1.3N$ processors for termination gaps of 2% or below.

We find that synchronous algorithms dramatically underutilize high performance computing infrastructure, resulting in processor idles time of up to 80.4%, which stresses the need for designing asynchronous algorithms in order to tackle industrial scale unit commitment problems.

Future extensions of the present work will focus on the application of the developed asynchronous decomposition framework for tackling (i) multi-stage stochastic unit commitment and (ii) detailed deterministic unit commitment problems over large interconnected power systems, integrating the optimization of transmission and distribution systems through convex relaxations of AC power flow constraints [CNH$^+$16].

Development history

An early version of the algorithm was published in [AP15]. In it we assumed that differences in solution times among scenario subproblems were not systematic and that, therefore, unbiased randomization was obtained as a byproduct of the differences in solution times. This turned out not to be the case in general, and very large biases were observed among subproblems in subsequent studies, preventing convergence to the dual optimum. Motivated by this observation, we developed a second version of the algorithm included in [APP17],

where we scaled the stepsizes applied to different scenarios in direct proportion to their average solution time. While this approach is guaranteed to converge and it worked well in practice for the CWE instances, it was not successful when applied to the WECC instances, due to the practical difficulty of estimating the average solution times when they are very volatile across iterations. Finally, in order to remove all dependencies from solution times characteristics, we adopted the randomization scheme presented in this chapter.

Alongside improvements in the dual part of the algorithm, significant improvements were made to primal recovery, initialization and implementation since the first and second versions, including: from using only LIFO [AP15, APP17] to implement RND, LIFO and develop IS; initialization schemes based on linear relaxation and time decoupling [APP17]; and moving from a Mosel implementation limited to a single machine [AP15], to a Mosel implementation on multiple machines [APP17], to arrive at the current MPI implementation in C.

The algorithm as presented in this chapter is currently under review for Mathematical Programming Computation.

Appendix

2.A Proofs

Proof of Prop. 1. By inspection we have that,

$$\mathbb{E}\left[I_J^T\Big(I_J\nabla f_0^\mu(\boldsymbol{x}^k)+g\big(J,\boldsymbol{x}_J^k\big)\Big)\Big|\,\boldsymbol{x}^k\right]=\frac{1}{N}\nabla f_0^\mu(\boldsymbol{x}^k)+\frac{1}{N}\sum_{j=1}^N I_j^T g\big(j,\boldsymbol{x}^k\big)$$
$$=\frac{1}{N}g\big(\boldsymbol{x}^k\big),\ g\big(\boldsymbol{x}^k\big)\in\partial f\big(\boldsymbol{x}^k\big).$$

□

Proof of Lemma 1. Recall that a vector $\boldsymbol{g}$ is said to be an approximate or ϵ-subgradient of a convex function f at $\boldsymbol{x}$ if and on

$$(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{g}\geq f(\boldsymbol{x})-f(\boldsymbol{y})-\epsilon,\quad\forall\boldsymbol{z}\in\mathrm{Dom}(f).$$

On the other hand, for the expected direction of update rule (2.11) we have

$$N\cdot(\boldsymbol{x}^k-\boldsymbol{y})^T\mathbb{E}\left[I_J^T\Big(I_J\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)+g\big(J,\boldsymbol{x}_J^{k-\ell(J,k)}\big)\Big)\Big|\,\mathcal{F}_k\right]=$$
$$(\boldsymbol{x}^k-\boldsymbol{y})^T\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)+\sum_{j=1}^N(\boldsymbol{x}_j^k-\boldsymbol{y}_j)^T g\big(j,\boldsymbol{x}_j^{k-\ell(j,k)}\big),\quad\forall\boldsymbol{y}\in X. \tag{2.16}$$

The first term in the right-hand side of (2.16) can be expanded as follows

$$\begin{aligned}\big(\boldsymbol{x}^k-\boldsymbol{y}\big)^T\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)&=\big(\boldsymbol{x}^{k-l(k)}-\boldsymbol{y}\big)^T\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)+\\&\quad\big(\boldsymbol{x}^k-\boldsymbol{x}^{k-l(k)}\big)^T\nabla f_0^\mu\big(\boldsymbol{x}^k\big)-\\&\quad\big(\boldsymbol{x}^k-\boldsymbol{x}^{k-l(k)}\big)^T\Big(\nabla f_0^\mu\big(\boldsymbol{x}^k\big)-\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)\Big)\\&\geq f_0^\mu\big(\boldsymbol{x}^k\big)-f_0^\mu(\boldsymbol{y})-\\&\quad\big(\boldsymbol{x}^k-\boldsymbol{x}^{k-l(k)}\big)^T\Big(\nabla f_0^\mu\big(\boldsymbol{x}^k\big)-\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)\Big)\\&\geq f_0^\mu\big(\boldsymbol{x}^k\big)-f_0^\mu(\boldsymbol{y})-\\&\quad\big\|\boldsymbol{x}^k-\boldsymbol{x}^{k-l(k)}\big\|_2\Big\|\nabla f_0^\mu\big(\boldsymbol{x}^k\big)-\nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big)\Big\|_2\\&\geq f_0^\mu\big(\boldsymbol{x}^k\big)-f_0^\mu(\boldsymbol{y})-L_0^\mu\big\|\boldsymbol{x}^k-\boldsymbol{x}^{k-l(k)}\big\|_2^2,\end{aligned}$$

where in the second line we use the convexity of f_0^μ ($(\boldsymbol{x}-\boldsymbol{y})^T\nabla f_0^\mu(\boldsymbol{x})\geq f_0^\mu(\boldsymbol{x})-f_0^\mu(\boldsymbol{y})$), in the third line we use the Cauchy-Schwarz inequality, and in the

fourth line we use the definition of the Lipschitz constant of the gradient of f_0^μ. Following a similar reasoning, each of the terms under the sum on the right hand side of (2.16) can be expanded as follows

$$\begin{aligned}(\boldsymbol{x}_j^k - \boldsymbol{y}_j)^T g\big(j, \boldsymbol{x}_j^{k-\ell(j,k)}\big) &\geq f_j\big(\boldsymbol{x}_j^k\big) - f_j\big(\boldsymbol{y}_j\big) - \\ &\qquad \big\|\boldsymbol{x}^k - \boldsymbol{x}^{k-\ell(j,k)}\big\|_2 \Big\|g\big(j, \boldsymbol{x}_j^k\big) - g\big(j, \boldsymbol{x}_j^{k-\ell(j,k)}\big)\Big\|_2 \\ &\geq f_j\big(\boldsymbol{x}_j^k\big) - f_j\big(\boldsymbol{y}_j\big) - 2D\big\|\boldsymbol{x}^k - \boldsymbol{x}^{k-\ell(j,k)}\big\|_2,\end{aligned}$$

where in the second line we use Assumption 1. Furthermore, Assumptions 1 and 2 allow us to bound the difference between current and delayed iterates using the stepsize as

$$\begin{aligned}\big\|\boldsymbol{x}^k - \boldsymbol{x}^{k-\ell(j,k)}\big\|_2 &\leq C \sum_{m=k-\ell(j,k)}^{k-1} \lambda_m \leq C \sum_{m=k-L}^{k-1} \lambda_m \\ \big\|\boldsymbol{x}^k - \boldsymbol{x}^{k-l(k)}\big\|_2^2 &\leq C^2 \sum_{m=k-L}^{k-1} \lambda_m^2,\end{aligned}$$

which along with the previous expressions lead us to the relation (2.17), concluding the proof.

$$\begin{aligned}N \cdot (\boldsymbol{x}^k - \boldsymbol{y})^T \mathbb{E}\left[I_J^T\Big(I_J \nabla f_0^\mu\big(\boldsymbol{x}^{k-l(k)}\big) + g\big(J, \boldsymbol{x}_J^{k-\ell(J,k)}\big)\Big)\Big|\, \mathcal{F}_k\right] \geq \\ f(\boldsymbol{x}^k) - f(\boldsymbol{y}) - \left(C^2 L_0^\mu \sum_{m=k-L}^{k-1} \lambda_m^2 + 2CDN \sum_{m=k-L}^{k-1} \lambda_m\right)\end{aligned} \tag{2.17}$$

□

Proof of Prop. 2. Let $\boldsymbol{h}_j = I_j^T(I_j \nabla f_0^\mu(\boldsymbol{x}^{k-l(k)}) + g(j, \boldsymbol{x}_j^{k-\ell(j,k)}))$ and J be a discrete uniform random variable on the set $\{1, \ldots, N\}$. Then for all $k = 1, \ldots, \infty$ and $\boldsymbol{y} \in X$, we have

$$\begin{aligned}\|\boldsymbol{x}^{k+1} - \boldsymbol{y}\|_2^2 &= \big\|\mathcal{P}_X\big[\boldsymbol{x}^k - \lambda_k \boldsymbol{h}_{j(k)}\big] - \boldsymbol{y}\big\|_2^2 \\ &\leq \big\|\boldsymbol{x}^k - \lambda_k \boldsymbol{h}_{j(k)} - \boldsymbol{y}\big\|_2^2 \\ &\leq \|\boldsymbol{x}^k - \boldsymbol{y}\|_2^2 - 2\lambda_k (\boldsymbol{x}^k - \boldsymbol{y})^T \boldsymbol{h}_{j(k)} + \lambda_k^2 C_2^2,\end{aligned}$$

where in the second line we use the nonexpansive property of the projection. Therefore, for the expectation conditioned on $\mathcal{F}_k$ it holds that

$$\mathbb{E}\big[\|\boldsymbol{x}^{k+1} - \boldsymbol{y}\|_2^2 \big| \mathcal{F}_k\big] \leq \|\boldsymbol{x}^k - \boldsymbol{y}\|_2^2 - 2\lambda_k (\boldsymbol{x}^k - \boldsymbol{y})^T \mathbb{E}\big[\boldsymbol{h}_J \big| \mathcal{F}_k\big] + \lambda_k^2 C^2$$

Finally, by substituting $\mathbb{E}\big[\boldsymbol{h}_J\big|\mathcal{F}_k\big]$ in the last relation using (2.17) (Lemma 1) we obtain the desired inequality.

□

Proof of Prop. 3. The first two inequalities follow directly from Assumption 3. For the third inequality we have

$$\sum_{k=0}^{\infty}\lambda_k\sum_{m=k-L}^{k-1}\lambda_m \;\leq\; \hat{G}^2\sum_{k=0}^{\infty}\gamma_k\sum_{m=k-L}^{k-1}\gamma_m \;\leq\; \hat{G}^2L\sum_{k=0}^{\infty}\gamma_{k-L}^2 \;<\; \infty,$$

where for the last inequality we use the fact that $\{\gamma_k\}$ is non-increasing and, therefore, $\sum_{m=k-L}^{k-1}\gamma_m \leq L\gamma_{k-L}\ \forall k$. The same reasoning applies to the fourth inequality of the present proposition.

□

Theorem 2 (Supermartingale Convergence Theorem [BT96, Prop. 4.2])**.** *Let X_t, Y_t and Z_t, $t = 0, 1, 2, \ldots$, be three sequences of random variables and let $\mathcal{F}_t$, $t = 0, 1, 2, \ldots$, be sets of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all t. Suppose that:*

(a) The random variables X_t, Y_t and Z_t are nonnegative, and are functions of the random variables in $\mathcal{F}_t$.

(b) For each t, we have $\mathbb{E}[X_{t+1}|\mathcal{F}_t] \leq X_t - Y_t + Z_t$.

(c) There holds $\sum_{t=0}^{\infty} Z_t < \infty$.

Then, we have $\sum_{t=0}^{\infty} Y_t < \infty$, and the sequence X_t converges to a non-negative random variable X with probability 1.

Proof of Theorem 1. From Proposition 2, with $\boldsymbol{y} = \boldsymbol{x}^*$, we obtain

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|_2^2\big|\mathcal{F}_k\big] \leq &\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 - 2\frac{\lambda_k}{N}\big(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)\big) + \\ &\lambda_k^2\hat{C}^2 + 2\frac{C^2L_0^{\mu}}{N}\lambda_k\sum_{m=k-L}^{k-1}\lambda_m^2 + 4CD\lambda_k\sum_{m=k-L}^{k-1}\lambda_m.\end{aligned}$$

Using Proposition 3 and by the Supermartingale Convergence Theorem (Theorem 2), with probability 1 and for each $\boldsymbol{x}^* \in X^*$, we have

$$\sum_{k=0}^{\infty}\lambda_k\big(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)\big) < \infty, \tag{2.18}$$

and with probability 1 the sequence $\big\{\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2\big\}$ converges to a random variable. The rest of the proof follows exactly the proof of [Ned02, Prop. 3.4]

(see also [Erm83, Thm. 2]) and it is repeated here only for the sake of being self contained.
For each $\boldsymbol{x}^* \in X^*$, let Ω_{x^*} denote the set of all sample paths for which equation (2.18) holds and $\left\{\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2\right\}$ converges. By convexity of f, the set X^* is convex, so there exist vectors $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_p \in X^*$ that span the smallest affine set containing X^*, and are such that $\boldsymbol{v}_j - \boldsymbol{v}_0$, $j = 1, \ldots, p$, are linearly independent.
The intersection $\Omega = \cap_{j=1}^{p} \Omega_{v_j}$ has probability 1, and for each sample path in Ω, the sequences $\left\{\|\boldsymbol{x}^k - \boldsymbol{v}_j\|_2\right\}$, $j = 0, \ldots, p$, converge. Thus, with probability 1, $\{\boldsymbol{x}^k\}$ is bounded, and therefore it has limit points. Furthermore, for each sample path in Ω, by equation (2.18) and the relation $\sum_{k=0}^{\infty} \lambda_k = \infty$, it follows that

$$\liminf_{k\to\infty} f(x^k) = f^*,$$

implying that $\{\boldsymbol{x}^k\}$ has at least one limit point that belongs to X^* by continuity of f. For any sample path in Ω, let $\bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{x}}$ be two limit points of $\{\boldsymbol{x}^k\}$ such that $\bar{\boldsymbol{x}} \in X^*$. Because the sequences $\left\{\|\boldsymbol{x}^k - \boldsymbol{v}_j\|_2\right\}$, $j = 0, \ldots, p$, converge, we must have

$$\|\bar{\boldsymbol{x}} - \boldsymbol{v}_j\|_2 = \|\hat{\boldsymbol{x}} - \boldsymbol{v}_j\|_2, \quad \forall j = 0, 1, \ldots, p.$$

Moreover, since $\bar{\boldsymbol{x}} \in X^*$, the preceding relation can hold only for $\bar{\boldsymbol{x}} = \hat{\boldsymbol{x}}$ by convexity of X^* and the choice of vectors $\boldsymbol{v}_j$. Hence, for each sample path in Ω, the sequence $\{\boldsymbol{x}^k\}$ has a unique limit point in X^*, implying that $\{\boldsymbol{x}^k\}$ converges to some optimal solution with probability 1.

□

2.B Detailed results

This section presents solution statistics per instance for the best configurations found. Table 2.6 presents solution statistics for WECC instances using the Cab supercomputer at LLNL. Table 2.7 presents solution statistics for CWE instances using Cab.

Table 2.6: Solution statistics for WECC. Results are obtained using the best configuration in Table 2.2: Polyak stepsize and IS primal recovery.

N	Season	Day	LB [MM$]	UB [MM$]	Sol. time 1% [s]	Av. *approx dual scenario* time [s]	Av. *dual* f_0 time [s]	Av. *primal projection* time [s]	*dual scenario* time [s] Av.	Sd.	*second stage scenario* time [s] Av.	Sd.	Math. Prog. time [%]
10	Autumn	WD	-7.161	-7.090	33.53	1.25	0.25	5.29	3.22	0.67	1.74	0.45	97.4
	Autumn	WE	-5.152	-5.100	75.52	1.06	0.23	1.78	2.59	1.11	1.41	1.54	99.1
	Spring	WD	-5.940	-5.881	126.25	1.07	0.23	1.32	2.99	1.15	1.38	0.29	99.1
	Spring	WE	-2.875	-2.847	1150.14	1.21	0.21	0.92	9.20	6.69	1.42	1.32	99.9
	Summer	WD	-11.301	-11.190	92.25	1.17	0.24	3.36	4.00	1.84	1.83	0.37	99.2
	Summer	WE	-7.166	-7.095	93.97	1.27	0.24	8.21	3.29	0.74	1.72	0.32	99.2
	Winter	WD	-7.007	-6.937	54.11	1.10	0.24	3.06	3.41	0.83	1.63	0.30	98.5
	Winter	WE	-3.438	-3.405	1291.48	1.17	0.20	0.58	10.98	6.61	1.49	0.53	99.9
100	Autumn	WD	-8.256	-8.174	38.05	1.14	0.19	9.76	3.10	0.57	1.64	0.28	97.0
	Autumn	WE	-5.104	-5.054	141.11	1.07	0.18	5.36	2.98	1.89	1.34	0.33	99.1
	Spring	WD	-5.798	-5.740	237.72	1.00	0.19	2.90	3.43	1.62	1.37	0.21	99.5
	Spring	WE	-3.309	-3.276	892.53	1.12	0.18	7.52	6.17	4.86	1.38	0.75	99.9
	Summer	WD	-11.380	-11.267	755.19	1.34	0.19	9.08	4.39	2.17	1.87	0.54	99.8
	Summer	WE	-8.460	-8.376	175.44	1.26	0.19	9.65	3.12	0.60	1.75	0.70	99.4
	Winter	WD	-6.060	-5.999	776.76	1.09	0.18	0.72	3.90	1.58	1.62	0.23	99.8
	Winter	WE	-3.556	-3.522	1126.09	1.26	0.18	4.98	7.99	4.51	1.51	0.45	99.9
1000	Autumn	WD	-8.106	-8.026	143.18	1.20	0.20	18.57	3.06	0.66	1.59	0.45	98.5
	Autumn	WE	-5.391	-5.337	383.42	1.07	0.19	12.20	2.72	1.95	1.34	0.47	98.3
	Spring	WD	-5.891	-5.832	868.55	1.04	0.19	1.39	3.26	1.34	1.42	0.53	98.7
	Spring	WE	-3.384	-3.350	5134.57	1.17	0.18	19.53	7.78	6.96	1.36	0.69	99.6
	Summer	WD	-11.418	-11.304	3080.87	3.28	0.19	19.22	3.98	2.19	1.84	0.56	99.1
	Summer	WE	-8.296	-8.214	433.10	1.17	0.19	20.10	3.19	0.74	1.66	0.26	98.6
	Winter	WD	-5.792	-5.737	6427.02	1.07	0.18	0.14	3.96	1.38	1.70	0.78	98.5
	Winter	WE	-3.621	-3.585	3809.37	1.18	0.19	4.13	8.32	6.16	1.49	0.48	99.7

Table 2.7: Solution statistics for CWE. Results are obtained using the same configuration as in Table 2.3: Polyak stepsize and IS primal recovery.

N	Season	Day	LB [MM€]	UB [MM€]	Sol. time 1% [s]	Av. *approx dual scenario* time [s]	Av. *dual* f_0 time [s]	Av. *primal projection* time [s]	*dual scenario* time [s] Av.	*dual scenario* time [s] Sd.	*second stage scenario* time [s] Av.	*second stage scenario* time [s] Sd.	Math. Prog. time [%]
30	Autumn	WD	-49.685	-49.212	3429.7	58.3	2.2	8.8	849.0	379.8	89.3	8.0	99.0
	Autumn	WE	-37.099	-36.779	1589.2	60.3	2.3	15.6	879.9	331.2	86.7	12.7	97.6
	Spring	WD	-33.152	-32.839	9279.1	79.4	2.2	1.3	3295.8	1408.6	88.6	27.2	99.4
	Spring	WE	-24.436	-24.216	6028.2	76.3	2.0	1.3	4473.0	1452.4	81.1	37.7	99.2
	Summer	WD	-35.805	-35.470	2493.1	68.5	2.2	1.3	1996.8	212.3	81.3	13.2	98.4
	Summer	WE	-23.385	-23.175	3046.2	72.7	2.2	1.4	1086.3	204.7	75.0	20.4	98.1
	Winter	WD	-30.462	-30.165	2136.9	62.4	2.2	1.3	849.1	363.7	76.7	12.4	97.6
	Winter	WE	-25.782	-25.543	2447.3	62.4	2.0	1.3	820.6	294.1	92.9	32.6	98.5
60	Autumn	WD	-49.792	-49.342	3538.7	58.5	2.2	12.3	889.5	508.5	89.1	9.1	99.1
	Autumn	WE	-37.144	-36.811	2178.6	60.0	2.3	2.5	989.9	374.8	85.7	10.6	98.2
	Spring	WD	-33.130	-32.821	8323.4	79.1	2.2	1.3	3295.1	1476.3	89.7	30.2	99.3
	Spring	WE	-24.397	-24.170	6116.6	76.7	2.0	1.4	4361.6	1524.8	79.3	25.4	99.2
	Summer	WD	-35.742	-35.500	2547.8	68.6	2.2	1.3	2016.4	240.5	82.6	15.0	98.4
	Summer	WE	-23.390	-23.185	2253.3	74.3	2.3	1.4	1176.8	282.2	75.1	18.9	97.4
	Winter	WD	-30.442	-30.143	2211.8	62.1	2.2	1.4	860.6	401.9	76.8	12.6	97.5
	Winter	WE	-25.754	-25.504	3023.3	62.8	2.0	1.4	815.6	351.6	96.6	42.4	98.5
120	Autumn	WD	-49.614	-49.120	2366.0	58.6	2.2	16.5	784.0	295.4	89.2	9.3	98.5
	Autumn	WE	-37.010	-36.642	1877.3	59.6	2.2	2.1	874.2	375.2	86.2	10.5	97.9
	Spring	WD	-33.187	-32.857	7952.6	78.3	2.1	1.3	3383.2	1408.9	88.9	27.0	99.2
	Spring	WE	-24.436	-24.214	5973.0	76.1	2.0	1.4	4135.9	1638.7	78.3	25.7	99.2
	Summer	WD	-35.806	-35.468	2497.3	68.4	2.1	1.5	2012.9	226.3	81.5	14.5	98.4
	Summer	WE	-23.377	-23.147	3506.5	74.2	2.2	1.5	1258.2	385.4	75.6	20.3	98.4
	Winter	WD	-30.431	-30.168	3932.2	61.6	2.1	1.4	1020.4	486.0	76.7	12.6	98.5
	Winter	WE	-25.780	-25.522	2904.4	62.2	2.1	1.5	793.6	387.5	93.2	32.2	98.5

Chapter 3

Transmission capacity allocation in zonal electricity markets

3.1 Introduction

Zonal electricity markets allow market participants to trade freely within each zone and to export/import energy to/from other zones up to certain technical limitations. Two approaches toward zonal market design are the focus of this chapter:

1. Available-Transfer-Capacity Market Coupling (ATCMC) [ABC+10b], which imposes limitations on the export/import between pairs of zones (see section 1.3.2), and

2. Flow-Based Market Coupling (FBMC) [HAA+17], which imposes limitations on the configuration of zonal net positions (see section 1.3.3).

Both approaches currently coexist in the European electricity market and they present significant conceptual and operative differences. FBMC has the ability to allocate the export/import capacities of each zone implicitly, potentially capturing inter-dependencies between bilateral exchanges. For instance, in a system with three zones A, B and C, all connected, the ability of A for exporting energy to B increase whenever A is not exporting to C at the same time. Such inter-dependencies are ignored in ATCMC. Additionally, FBMC can handle a larger variety of constraints on inter-zonal exchanges than ATCMC, thereby allowing TSOs to include transmission constraints in a more transparent and explicit manner for day-ahead market clearing. Consequently, under

FBMC, TSOs may not need to consider large security margins on bilateral exchanges. The expected benefits of FBMC rendered it as the preferred method for linking markets in the European Union [Eur15a].

Despite their differences, FBMC and ATCMC are both zonal electricity markets and, as such, they can only approximate the inter- and intra-zonal power flows of the real grid to a limited extent, allocating transmission capacity in an inaccurate fashion. This often causes market-clearing schedules that would result in overloaded transmission equipment under both FBMC and ATCMC. Congestion management measures are then required after the clearing of the electricity market, in order to operate the system within its security limits in real time. The costs of these remedial measures can be very important, however they are commonly ignored in market analyses. For instance, the parallel run between FBMC and ATCMC for the CWE system found potential welfare gains of FBMC over ATCMC in the order of 95 M€ for 2013 [AAC+15] while ignoring congestion management costs, which would amount to 945 M€ in 2015 [Eur15b] (the first year of operations under FBMC). Therefore, by virtue of the magnitude of congestion management costs, the effect of the different policies on these remedial actions could have affected the conclusions of the parallel run.

The models presented in this chapter are intended to provide an instrument to perform comprehensive policy analyses, accounting for the overall performance of zonal market designs, and to understand the possible future evolution of the European zonal electricity market. The model avoids simplifying assumptions often made in economic analysis and aims at being as realistic as possible in its encompassing of current and future legal obligations. The day-ahead market follows the organization described in Regulation (EC) 714/2009 [Eur09]. Real-time operation, which is only currently defined in Regulations (EC) 1222/2015 [Eur15a] and (EU) 2195/2017 [Eur17a] stating legal objectives, assumes that these objectives are effectively realized. The price to pay for this lack of simplification is computational both in terms of algorithmic tools and machine resources.

3.1.1 Literature review

ATCMC has been the standard zonal electricity market model analyzed in the literature because of its past presence in US electricity markets and its current presence in European electricity markets. Studies using small examples and realistic systems, under various assumptions and modeling choices, have all concluded that the performance of ATCMC is significantly worse than that of a nodal system (see section 4.1.1, in the following chapter, for a survey), even in the case where ATCs can be optimized so as to reduce real operation costs, as done by Jensen *et al.* [JKP17].

Academic studies on FBMC, on the other hand, are scarce as FBMC is a

relatively new capacity allocation methodology. Early studies were performed before the *go-live* of FBMC at the CWE. Waniek *et al.* [WRH10] study the day-ahead market performance and the accuracy of power flow approximations of ATCMC, FBMC and LMP. For ATCMC and FBMC, the authors disaggregate zonal injections into nodal injections in proportion to the injections in a base case (similar to the procedure followed by RTE [HAA$^+$17]). The factors that are used for disaggregating zonal injections into nodal injections are known as Generation Shift Keys (GSKs) in the literature. The study finds that FBMC outperforms ATCMC in terms of both performance and accuracy, while LMP exhibits superior performance relative to both FBMC and ATCMC. Following the *go-live*, in an effort towards understanding the new capacity allocation mechanism, Van den Bergh *et al.* [VBD16] summarize the concepts and methodology used in FBMC.

Significant attention has been dedicated towards understanding how discretionary parameters determined by TSOs (see sections 1.3.2 and 1.3.3) affect the day-ahead outcome of FBMC. Marien *et at.* [MLTW13] study the effect of the configuration of bidding zones, and the determination of Flow Reliability Margins (FRM) and GSKs on exchanges and prices. The authors find that different choices for these parameters for the same system can lead to very different market outcomes. In the same vein, Dierstein [Die17] analyses different strategies used by CWE TSOs to compute GSKs and how they affect the outcome of FBMC and congestion management for cross-border lines. The author finds that dynamic GSK strategies (i.e. where GSKs vary from one hour to the next) outperform static GSK strategies, the latter being currently used by all TSOs except RTE.

3.1.2 Contributions and chapter organization

The contributions of the present chapter are threefold. In terms of modeling, we propose a framework for modeling zonal electricity markets that avoids the discretionary parameters and circular definitions present in the current practice and in the literature. This is achieved in the proposed FBMC and ATCMC models by projecting the actual network constraints onto the space of zonal net positions and bilateral exchanges, respectively. The computational contribution of the chapter is the development of cutting-plane algorithms for clearing the day-ahead market under each zonal policy, while endogenously enforcing robustness of the import/export decisions against any single element failure (i.e. satisfying the N-1 security criterion [HAA$^+$17,ABC$^+$10b,RTE06]). These clearing problems correspond to adjustable robust optimization problems [BTEN09]. The policy contribution of the chapter is the detailed simulation of a realistic-scale instance of the CWE system (section 1.4.3) against detailed models of LMP, FBMC and ATCMC. These simulations account for the clearing of energy and reserves, renewable supply forecasts errors and the outage of components,

the pricing of non-convex operating costs and constraints, and the two-stage nature of market operations whereby day-ahead market clearing is followed by congestion management and balancing. Numerical results over 768 000 different operating conditions demonstrate that FBMC and ATCMC attain very similar performance. The major policy message of the chapter is to challenge whether is it worth for more European countries to switch from ATCMC to FBMC, instead of advancing directly to a nodal design.

The rest of the chapter is organized as follows. Section 3.2 introduces our modeling framework for transmission capacity allocation in zonal markets using linear programs that capture the most important differences in the alternative market designs. These simplified models permit an analysis of the differences with respect to LMP and the FBMC methodology implemented in the CWE. Section 3.3 develops cutting-plane algorithms for simulating FBMC and ATCMC under N-1 security, starting from the simplified models of section 3.2 to ease the exposition. Section 3.4 introduces our two-settlement market models (day-ahead, real-time) with commitment decisions. Section 3.5 presents the simulation setup, and section 3.6 presents the main numerical results and discusses the implications of these results for zonal electricity markets. Finally, section 3.7 concludes the chapter and outlines directions for future research.

3.2 Transmission capacity allocation in electricity markets

Transmission capacity allocation mechanisms include (*i*) forward contracts, for instance, the long term auctions carried out by the Joint Allocation Office (JAO) [Joi15], and (*ii*) different types of implicit allocation, typically used in day-ahead electricity markets [SCTB88]. In the following, we focus on day-ahead electricity markets and describe each policy for transmission capacity allocation (LMP, FBMC and ATCMC) in its simplest form in order to better understand the differences between them. We assume that demand is fixed and only producers bid in the market. We further assume that all market participants act as price takers (i.e. that they bid their true cost to the market) and that all energy is traded in the day-ahead auction (i.e. we ignore long-term contracts and bilateral trades).

3.2.1 Nodal electricity markets

The nodal day-ahead electricity market can be cleared by solving the following optimization problem (already introduced in section 1.3.1, repeated here for

convenience):

$$\begin{aligned}\min_{v\in[0,1],f,\theta} \quad & \sum_{g\in G} P_g Q_g v_g \\ \text{s.t.} \quad & \sum_{g\in G(n)} Q_g v_g - \sum_{l\in L(n,\cdot)} f_l + \sum_{l\in L(\cdot,n)} f_l = Q_n \quad \forall n \in N \quad [\rho_n] \\ & -F_l \le f_l \le F_l,\ f_l = B_l\left(\theta_{m(l)} - \theta_{n(l)}\right) \quad \forall l \in L.\end{aligned}$$

The notation in this model is as follows: Q_g, P_g correspond to the quantity and price bid by generator $g \in G$; $G(n)$ is the set of generators at node n; Q_n is the forecast demand at node $n \in N$; $F_l, B_l, m(l), n(l)$ are the thermal limit, susceptance, and adjacent nodes (in the outgoing and incoming direction respectively) of line $l \in L$; $L(m,n)$ is the set of lines directed from node m to node n; v_g is the acceptance/rejection decision for the bid placed by generator g; f_l is the flow through line l; θ_n is the voltage angle at node n; and ρ_n denotes the LMP at node n.

The LMP policy implicitly allocates the capacity of all lines in the system, without any distinction between zones, and while respecting the network constraints.

3.2.2 Zonal electricity markets

In zonal electricity markets, electricity is priced at a zonal level, and nodal level information is discarded [ES05]. Bids are associated to zones instead of nodes, and transmission constraints can only be imposed at a zonal level. A zonal market can be cleared by solving problem (3.1) – (3.3):

$$\min_{v\in[0,1],p} \quad \sum_{g\in G} P_g Q_g v_g \tag{3.1}$$

$$\text{s.t.} \quad \sum_{g\in G(z)} Q_g v_g - p_z = \sum_{n\in N(z)} Q_n \quad \forall z \in Z \quad [\rho_z] \tag{3.2}$$

$$p \in \mathcal{P}, \tag{3.3}$$

where p_z corresponds to the net position of zone $z \in Z$ and $G(z), N(z)$ correspond to the sets of generators and nodes in zone z. $\mathcal{P}$ corresponds to the feasible set of net positions, a convex polyhedron defined differently according to the inter-zonal capacity allocation mechanism employed, but always demanding that system balance is respected, i.e. $\sum_z p_z = 0$. In this model, constraints (3.2) define the zonal net positions and constraint (3.3) enforce that the zonal net positions belong in $\mathcal{P}$.

European regulations set forth guidelines for defining $\mathcal{P}$. Annex I of Regulation (EC) 714/2009 [Eur09] establishes that "... *TSOs shall endeavour to accept all commercial transactions, including those involving cross-border-trade* ..." (Article 1.1) and that "... *TSOs shall not limit interconnection capacity*

in order to solve congestion inside their own control area, save for the above-mentioned reasons and reasons of operational security ..." (Article 1.7). In the spirit of this Regulation, $\mathcal{P}$ should include all net position configurations p that are feasible with respect to the real grid and only exclude those that can be proven to lead to insecure operating conditions. We describe three methodologies for defining $\mathcal{P}$: FB with GSKs, FB with exact projection, and ATC.

3.2.2.1 Flow-based methodology with Generation Shift Keys

This methodology is currently used in the implementation of FBMC in the CWE [HAA+17]. The first step in the method is to determine GSKs for each generator within each zone. GSKs quantify how a change in zonal net position would be achieved by changing the output of generators, with respect to a base case dispatch, i.e. $GSK_g = Q_g \Delta v_g / \Delta p_{z(g)}$. GSKs, along with the node-to-line PTDF matrix, are used to compute zone-to-line PTDFs as $PTDF_{l,z} = \sum_{g \in G(z)} GSK_g \cdot PTDF_{l,n(g)},\ \forall l \in L, z \in Z$. Then, using the base case dispatch (denoted with a zero superscript), the flow across each line is approximated as $f_l \approx f_l^0 + \sum_{z \in Z} PTDF_{l,z} \cdot (p_z - p_z^0),\ \forall l \in L$.

If, for a certain $l \in L$ and any $z \in Z$, $PTDF_{l,z}$ is larger than 5%, then l is considered as a Critical Branch (CB). For each critical branch, TSOs determine its corresponding Remaining Available Margin (RAM) starting from the thermal capacity of the branch minus the flow on the base case ($F_l - f_l^0$), subtracting the FRM, and adding the Final Adjustment Value (FAV). Finally, the set of feasible net positions with GSKs, $\mathcal{P}^{FB-GSK}$, is described by

$$\mathcal{P}^{FB-GSK} = \left\{ p \in \mathbb{R}^{|Z|} \;\middle|\; \sum_{z \in Z} p_z = 0,\ \sum_{z \in Z} PTDF_{cb,z} \cdot p_z \leq RAM_{cb}\ \forall cb \in CB \right\}. \tag{3.4}$$

The methodology has several points where discretionary decisions are required by TSOs. These include the selection of a base case, the determination of GSKs, the selection of CB, and the determination of FRM and FAV, all of which are subject to the discretion of the TSOs, and indeed different TSOs use different criteria for certain choices [HAA+17, CRE17]. Additionally, $\mathcal{P}^{FB-GSK}$ will only be a good representation of reality whenever the flows on critical branches are approximated accurately. Unfortunately, this cannot be guaranteed [WRH10]. Furthermore, a circular problem arises: the better TSOs can anticipate the outcome of the market, the closer the base case would be to reality and, consequently, power flows would be approximated more accurately. However, the outcome of the market depends on the parameters decided by TSOs. These problems, among others, have placed the FBMC-GSK methodology under scrutiny by National Regulatory Authorities (NRAs) [EN17,CRE17].

3.2.2.2 Flow-based methodology with exact projection

Instead of resorting to assumptions for approximating power flows using zonal net positions, we can construct $\mathcal{P}$ based solely on the physics of the actual grid and the requirements of the European regulation [Eur09, Annex I] on $\mathcal{P}$: (i) $\mathcal{P}$ should include all feasible net position configurations and (ii) $\mathcal{P}$ should exclude all net position configurations that can be proven infeasible.

Note that there exists a unique set $\mathcal{P}$ respecting these requirements, because the set of all feasible net positions in uniquely defined for any particular grid. We propose constructing that set as the projection of the grid constraints onto the space of net positions:

$$\begin{aligned}\mathcal{P}^{FB-EP} = \Bigg\{ p \in \mathbb{R}^{|Z|} \;\Bigg|\; & \exists (\bar{v}, f, \theta) \in [0,1]^{|G|} \times \mathbb{R}^{|L|} \times \mathbb{R}^{|N|} : \\ & \sum_{g \in G(z)} Q_g \bar{v}_g - p_z = \sum_{n \in N(z)} Q_n \quad \forall z \in Z, \\ & \sum_{g \in G(n)} Q_g \bar{v}_g - \sum_{l \in L(n,\cdot)} f_l + \sum_{l \in L(\cdot,n)} f_l = Q_n \quad \forall n \in N, \\ & -F_l \leq f_l \leq F_l, \; f_l = B_l \left(\theta_{m(l)} - \theta_{n(l)} \right) \quad \forall l \in L \Bigg\}.\end{aligned} \tag{3.5}$$

It is easy to verify that $\mathcal{P}^{FB-EP}$ complies with the aforementioned restrictions. Further, by our previous observation, $\mathcal{P}^{FB-EP}$ is the unique definition of a flow-based domain in compliance with the current European regulation.

Using $\mathcal{P}^{FB-EP}$ in (3.3) has also the advantages that it does not require any approximation assumptions or arbitrary parameters and that it is implementable[1], making the proposed model an objective tool for studying the operational implications of flow-based market coupling and of the European Target Model.

3.2.2.3 Available-transfer-capacity methodology

The ATC methodology defines a set of interconnectors T between neighboring zones, each $t \in T$ comprising cross-border lines $L(t) \subseteq L$, and assigns a maximum capacity in the forward (ATC_t^+) and backward directions (ATC_t^-) for each interconnector. These capacities are determined in a series of steps, analogous to those presented in subsection 3.2.2.1, aiming at computing simultaneous limits on bilateral exchanges, see [ABC+10b]. We abstract from these arbitrary considerations and, following the principle that $\mathcal{P}$ should include the

[1]The model is implementable in practice because it only requires information about the grid, the installed generation capacity and the forecast demand at each node, all of which are already available to system operators and market participants. No information about the price or attribution of the bids to nodes is necessary.

largest possible subset of feasible net position configurations, compute ATCs by solving the optimization problem (3.6) – (3.8):

$$\max_{ATC} \prod_{t\in T} \left(ATC_t^- + ATC_t^+\right) \tag{3.6}$$

$$\text{s.t.} \quad -ATC_t^- \leq ATC_t^+ \quad \forall t \in T \tag{3.7}$$

$$[-ATC^-, ATC^+] \subseteq \mathcal{E}^{EP}, \tag{3.8}$$

where $[-ATC^-, ATC^+] \subseteq \mathbb{R}^{|T|}$ is the rectangle with lower vertex $-ATC^-$ and upper vertex ATC^+, and $\mathcal{E}^{EP}$, defined in (3.9), is the feasible domain of commercial exchanges:

$$\begin{aligned}\mathcal{E}^{EP} = \Bigg\{ e \in \mathbb{R}^{|T|} \Bigg| & -\sum_{l\in L(t)} F_l \leq e_t \leq \sum_{l\in L(t)} F_l \quad \forall t \in T, \\ & \exists p \in \mathcal{P}^{FB-EP} : p_z = \sum_{t\in T(z,\cdot)} e_t - \sum_{t\in T(\cdot,z)} e_t \quad \forall z \in Z \Bigg\}.\end{aligned} \tag{3.9}$$

Problem (3.6) – (3.8) seeks to maximize the volume of the rectangle formed by the ATC values of all interconnectors, while ensuring that (i) the bilateral exchange between each pair of zones is bounded by their total interconnection capacity, and (ii) the net position configuration at each exchange configuration within the ATC rectangle is feasible with respect to the real network constraints [Ten14].

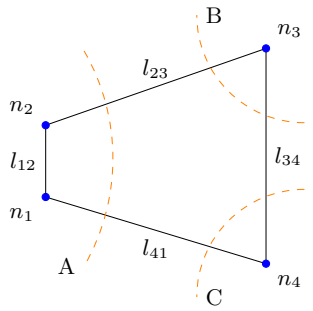
Once the ATC values are available from solving problem (3.6) – (3.8), the feasible net position domain under the ATC methodology, $\mathcal{P}^{ATC}$, can be defined as:

$$\begin{aligned}\mathcal{P}^{ATC} = \Bigg\{ p \in \mathbb{R}^{|Z|} \Bigg| \; \exists e \in \mathbb{R}^{|T|} : \; & p_z = \sum_{t\in T(z,\cdot)} e_t - \sum_{t\in T(\cdot,z)} e_t \quad \forall z \in Z, \\ & -ATC_t^- \leq e_t \leq ATC_t^+ \quad \forall t \in T \Bigg\}.\end{aligned} \tag{3.10}$$

In contrast with $\mathcal{P}^{FB-GSK}$, defined in (3.4), and $\mathcal{P}^{FB-EP}$, defined in (3.5), where we enforce zonal balance explicitly, $\mathcal{P}^{ATC}$ enforces zonal power balance implicitly, because the interconnectors define a transportation network between the zones.

3.2.3 Feasible domain comparison

The feasible set of zonal net positions differs among policies. Denote by $\mathcal{P}^{LMP}$ the feasible set of zonal net positions of the LMP policy. Then, by construction, we have that $\mathcal{P}^{LMP} = \mathcal{P}^{FB-EP} \supseteq \mathcal{P}^{ATC}$, while $\mathcal{P}^{FB-GSK}$ is not comparable to the previous sets. Thus, while FBMC-EP allows all possible inter-zonal exchange schedules, ATCMC might not allow some of them, but FBMC-EP



Generators

g	$n(g)$	Q_g [MW]	P_g [\$/MWh]
1	n_1	500	8
2	n_2	200	45
3	n_3	300	18
4	n_4	500	200

Consumers

n	Q_n [MW]
n_2	300
n_4	300

Figure 3.1: 4-node, 3-zone network data of section 3.2.4. All lines have equal impedance.

and ATCMC will not allow clearing with an infeasible cross-border exchange schedule. FBMC-GSK, on the other hand, does not offer any guarantees for allowing feasible or rejecting infeasible cross-border exchange schedules.

The feasible set of acceptance/rejection of bids is another interesting point of comparison. Let us define the feasible set of acceptance/rejection of each policy as:

$$\begin{aligned}
\mathcal{V}^{LMP} &= \{v \in [0,1]^{|G|} \mid \exists (f,\theta),(v,\theta,f) \text{ respecting (1.16) – (1.17)}\},\\
\mathcal{V}^{FB-GSK} &= \{v \in [0,1]^{|G|} \mid \exists p,(v,p)\\
&\qquad \text{respecting (3.2) – (3.3) with } \mathcal{P} := \mathcal{P}^{FB-GSK}\},\\
\mathcal{V}^{FB-EP} &= \{v \in [0,1]^{|G|} \mid \exists p,(v,p)\\
&\qquad \text{respecting (3.2) – (3.3) with } \mathcal{P} := \mathcal{P}^{FB-EP}\} \text{ and}\\
\mathcal{V}^{ATC} &= \{v \in [0,1]^{|G|} \mid \exists p,(v,p) \text{ respecting (3.2) – (3.3) with } \mathcal{P} := \mathcal{P}^{ATC}\}.
\end{aligned}$$

Then, we have that $\mathcal{V}^{LMP} \subseteq \mathcal{V}^{FB-EP} \supseteq \mathcal{V}^{ATC}$, while $\mathcal{V}^{FB-GSK}$ is not comparable with the previous sets. This implies that, although FBMC-EP and ATCMC are guaranteed to clear with a feasible cross-border exchange schedule $p_z, z \in Z$, the acceptance/rejection decisions for bids taken by these models might not be feasible for the real network.

3.2.4 Policy comparison using a small instance

We compare LMP, FBMC-GSK, FBMC-EP and ATCMC using the 4-node, 3-zone network presented in Fig. 3.1. We investigate both cases of inter- as well as intra-zonal scarce transmission capacity.

Table 3.1: Summary of clearing quantities and prices for a case of inter-zonal congestion (l_{41} limited to 100 MW). Flow approximation absolute error refers to the sum over all lines of the difference between model flows (i.e. flows inside the definitions of $\mathcal{P}^{FB-GSK}$ in (3.4) and of $\mathcal{P}^{FB-EP}$ in (3.5)) and implied flows (i.e. flows that would transit over the lines in the network if the optimal decisions of each policy v^* were implemented).

Policy	Total cost [$]	ρ [$/MWh] Min.	Max.	Abs. error flow approx. [MW]	Overload l_{41} [MW]
LMP	15 200	8	119	0	0
FBMC-GSK	7 217	8	200	475	79
FBMC-EP	7 800	8	23	300	50
ATCMC	23 208	8	200	–	50

3.2.4.1 Inter-zonal scarce transmission capacity.

In order to study the behavior of the different policies under inter-zonal scarce transmission capacity, we clear the market for the system of Fig. 3.1 with the thermal capacity of line l_{41} assumed equal to 100 MW. All other lines are assumed to have unlimited capacity. We use the GSK strategy of Elia, whereby we disaggregate changes in net position in proportion to the installed capacity of each generator [HAA+17].

Table 3.1 presents a summary of the clearing results. We find that all zonal policies clear with acceptance/rejection decisions that result in infeasible flows for the real network (6^{th} column), and that FBMC-GSK leads to a wider price range (3^{rd} and 4^{th} columns) and a larger approximation error than FBMC-EP (5_{th} column). We present the feasible domains of net positions for each model in Fig. 3.2. It is interesting to observe that LMP and FBMC-EP clear with the same net positions. Nevertheless, the LMP acceptance/rejection decisions are feasible for the real network, while the decisions of FBMC-EP are not. This occurs because of the difference in price and location of bids within zone A. Zonal policies will accept these bids following the merit order, in the sense of accepting the bid at n_1 fully before accepting any part of the bid at n_2. This decision, however, leads to an overloading of line l_{41}. According to FBMC-EP, this decision is feasible because there exist at least one $\bar{v}$ with the same net positions (for instance, the optimal decision of the LMP policy).

FBMC-GSK clears with a cross-border exchange schedule that is infeasible for the real network. Note also that the set $\mathcal{P}^{FB-GSK}$ does not include a slice of $\mathcal{P}^{FB-EP}$ (containing feasible net positions for the real network) to the left of the graph of Fig. 3.2. In other words, FBMC-GSK fails to accurately account for cross-border exchanges, and distorts the market outcome due to discretionary parameters used for approximating power flows on lines.

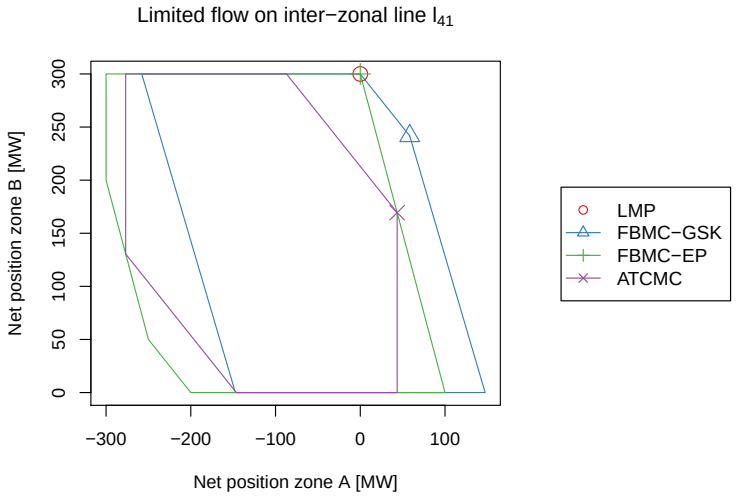


Figure 3.2: Set of feasible net positions on plane A-B for a case of inter-zonal congestion (l_{41} limited to 100 MW). The net position of zone C is implied by the net positions of A and B because of energy balance. The points indicate the zonal net positions at the optimal solution for each policy.

3.2.4.2 Intra-zonal scarce transmission capacity.

In order to study the effect of intra-zonal congestion, we use the same system of Fig. 3.1, where we now constrain the thermal capacity of line l_{12} to 100 MW and assume an unlimited capacity for all other lines. We employ the same GSKs as in the previous subsection.

We can observe in table 3.2 that all zonal policies clear with acceptance/rejection decisions that are infeasible for the real network. As in the case of inter-zonal congestion, the flows estimated by FBMC-GSK are a less accurate approximation than those estimated by FBMC-EP. Interestingly, the estimated flows in FBMC-GSK can be in the opposite direction of the flows implied by the DC power flow equations. As shown in Fig. 3.3, for this case $\mathcal{P}^{FB-GSK}$ turns out to be a relaxation of $\mathcal{P}^{FB-EP}$.

In summary, zonal markets fail to properly allocate scarce transmission capacity both when inter-zonal and intra-zonal congestion arises. FBMC-EP outperforms FBMC-GSK in accuracy while not introducing market distortions. For this reason, in what follows, we will only consider FBMC-EP and will refer to it simply as FBMC.

Table 3.2: Summary of clearing quantities and prices for a case of intra-zonal congestion (l_{12} limited to 100 MW).

Policy	Total cost [$]	ρ [$/MWh] Min.	Max.	Abs. error flow approx. [MW]	Overload l_{12} [MW]
LMP	10 267	8	45	0	0
FBMC-GSK	5 800	18	18	536	150
FBMC-EP	5 800	18	18	300	150
ATCMC	9 750	8	200	–	108

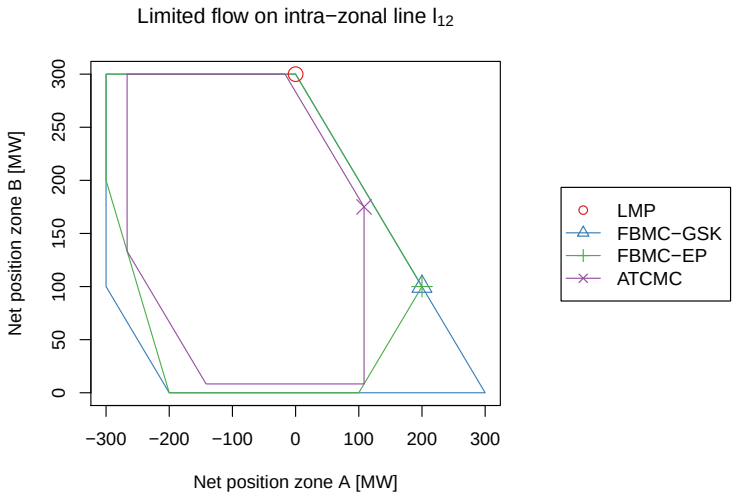


Figure 3.3: Set of feasible net positions on plane A-B for a case of intra-zonal congestion (l_{12} limited to 100MW). The points indicate the zonal net positions at the optimal solution for each policy.

3.3 Cutting-plane algorithms for zonal market clearing

In order to obtain numerical results in subsection 3.2.4, we solve (3.1) – (3.3) by substituting $\mathcal{P}^{FB-EP}$ in constraint (3.3) with its definition (3.5) using auxiliary variables $\bar{v}, f, \theta$. Similarly, in order to solve (3.6) – (3.8) we explicitly model all vertices $\nu \in \mathcal{V}$ of the rectangle $[-ATC^-, ATC^+]$, and we enforce that the corresponding limit exchange e^ν of each vertex $\nu \in \mathcal{V}$ is a feasible commercial exchange configuration, i.e. $e^\nu \in \mathcal{E}^{EP}$. Furthermore, writing the latter constraints requires auxiliary variables $p^\nu, \bar{v}^\nu, f^\nu, \theta^\nu$ for describing $\mathcal{E}^{EP}$ in terms of its definition (3.9), for each vertex $\nu \in \mathcal{V}$. These approaches, while useful for understanding the behavior of these models on a small example, are not suitable for realistic systems under the N-1 security criterion on export/imports, which we impose in accordance with the technical documentation on FBMC [HAA+17] and ATCMC [ABC+10b]. We define N-1 security as the ability of a system to maintain its net position configuration while supplying all the demand, for any outage of a single generating unit or a single transmission line in the system [RTE06, Chapter 7, Article 7.1]. Describing $\mathcal{P}^{FB-EP}_{N-1}$ and $\mathcal{E}^{EP}_{N-1}$ (i.e. the sets of feasible net positions and feasible commercial exchanges, robust to any single element contingency) using auxiliary variables would lead to mathematical programs with millions of variables and constraints. Writing $\mathcal{P}^{FB-EP}_{N-1}$ and $\mathcal{E}^{EP}_{N-1}$ explicitly in terms of net positions and exchanges, on the other hand, is neither a performant option since this would require the projection of the descriptions with auxiliary variables onto lower dimensional spaces, leading to an exponentially larger number of constraints [Mon10, Lemma 2].

In this section, we propose cutting-plane algorithms that progressively refine outer-approximations of $\mathcal{P}^{FB-EP}_{N-1}$ and $\mathcal{E}^{EP}_{N-1}$ as needed, allowing us to simulate zonal market clearing for the real-scale instance presented in section 3.5. Note that our setting is ideal for cutting-plane algorithms due to the small number of dimensions of $\mathcal{P}^{FB-EP}_{N-1}$ and $\mathcal{E}^{EP}_{N-1}$ ($|Z|$ and $|T|$, respectively). These algorithms have been inspired by the algorithm proposed by Street *et al.* [SMA14] for security-constrained unit commitment, where the authors use a cutting-plane procedure to progressively improve an under approximation of the maximum demand curtailment after a contingency, for a given commitment decision. In contrast, our algorithms generate descriptions of the inclusion constraints (3.3) and (3.8) based on distance functions that become zero if and only if the inclusion constraint is respected. In what follows, subsection 3.3.1 presents a cutting-plane algorithm for solving (3.1) – (3.3) and subsection 3.3.2 presents a cutting-plane algorithm for solving (3.6) – (3.8). These algorithms can be easily adapted to be applied to the detailed day-ahead market models of section 3.4. Proofs for the propositions presented in this section are provided

Algorithm 5 Cutting-plane algorithm for solving FBMC.

Initialize $V := 0_{1,|Z|}, W := 0$, `inclusion` := `FALSE`
while !`inclusion` **do**
  Call $MCO(V, W) \to p$
  if `status` = `INFEASIBLE` **then**
    Terminate: FBMC clearing problem (3.31) – (3.38) is infeasible.
  end if
  Call $NPO(p) \to$ `inclusion`, (v, w)
  $V := [V^\top \ v]^\top, \ W := [W^\top \ w]^\top$
end while
Terminate: inner model of $MCO(V, W)$ gives the optimal clearing.

in section 3.A at the end of this chapter.

3.3.1 Decomposition algorithm for flow-based market coupling with N-1 security

The main idea behind the decomposition algorithm for FBMC under the N-1 security criterion, i.e. (3.1) – (3.3) with $\mathcal{P} \equiv \mathcal{P}_{N-1}^{FB-EP}$, is to replace the inclusion condition $p \in \mathcal{P}_{N-1}^{FB-EP}$ by a polyhedral outer approximation, $\sum_{z \in Z} V_{m,z} p_z \leq W_m \quad \forall m = 1, \ldots, M$, which is tight at the optimal solution p^*. Following this reasoning, we propose clearing the market under FBMC using Alg. 5, which is based on repeatedly calling two oracles:

- A market-clearing oracle $MCO(V, W)$ which, for a given $V \in \mathbb{R}^{M \times |Z|}$ and $W \in \mathbb{R}^M$, solves the FBMC clearing problem using $Vp \leq W$ as a substitute for $p \in \mathcal{P}_{N-1}^{FB-EP}$, and returns a vector of optimal net positions p^*.

- A net position oracle $NPO(p)$ which, for a given vector of net positions p, either certifies that $p \in \mathcal{P}_{N-1}^{FB-EP}$, or returns a hyperplane that separates p from $\mathcal{P}_{N-1}^{FB-EP}$.

The effectiveness of Alg. 5 in handling realistic instances depends on the specific net position oracle used. For example, an oracle that produces a deep separating hyperplane but relies on checking all N-1 contingencies one-by-one would not be effective in practice.. In order to design an effective oracle, we first define $\mathcal{P}_{N-1}^{FB-EP}$ formally as $\mathcal{P}_{N-1}^{FB-EP} = \cap_{\substack{u \in \{0,1\}^{|G|+|L|} \\ \|u\|_1 \leq 1}} \mathcal{P}^{FB-EP}(u)$, where

$\mathcal{P}^{FB-EP}(u)$ corresponds to

$$\mathcal{P}^{FB-EP}(u) = \Bigg\{ p \in \mathbb{R}^{|Z|} \;\Bigg|\; \exists (\bar{v}, f, \theta) \in [0,1]^{|G|} \times \mathbb{R}^{|N|} \times \mathbb{R}^{|L|} : \\ \sum_{g \in G(z)} Q_g (1-u_g)\bar{v}_g - p_z = \sum_{n \in N(z)} Q_n \;\forall z \in Z, \\ \sum_{i \in G(n)} Q_g (1-u_g)\bar{v}_g - \sum_{l \in L(n,\cdot)} f_l + \sum_{l \in L(\cdot,n)} f_l = Q_n \;\forall n \in N, \\ -F_l \le f_l \le F_l,\; f_l = B_l(1-u_l)\left(\theta_{m(l)} - \theta_{n(l)}\right) \;\forall l \in L \Bigg\}.$$

Further, note that we can express the inclusion condition equivalently in terms of point-to-set distance, i.e. $\bar{p} \in \mathcal{P}_{N-1}^{FB-EP}$ if and only if $d(p, \mathcal{P}_{N-1}^{FB-EP}) = 0$ for any distance function d. Particularly, given our definition of $\mathcal{P}_{N-1}^{FB-EP}$, we consider the following distance function:

$$d(\bar{p}, \mathcal{P}_{N-1}^{FB-EP}) = \max_{\substack{u \in \{0,1\}^{|G|+|L|} \\ \|u\|_1 \le 1}} \; \min_{p \in \mathcal{P}^{FB-EP}(u)} \|\bar{p} - p\|_1,$$

which is defined using a bi-level mathematical program. The inner problem can be cast as a linear program which we can dualize and derive an alternative definition of $d(\bar{p}, \mathcal{P}_{N-1}^{FB-EP})$ as a bi-linear program:

$$\begin{aligned} d(\bar{p}, \mathcal{P}_{N-1}^{FB-EP}) = \\ \max_{\substack{u,\psi,\sigma, \\ \gamma,\phi,\rho}} \; & \sum_{z \in Z} \bar{p}_z \psi_z + \sum_{n \in N} Q_n \rho_n + \sum_{z \in Z} \psi_z \cdot \sum_{n \in N(z)} Q_n - \sum_{g \in G} \sigma_g - \sum_{l \in L} F_l(\gamma_l^- + \gamma_l^+) \\ \text{s.t.} \quad & -1 \le \psi_z \le 1 \quad \forall z \in Z \\ & \sigma_g \ge Q_g(1-u_g)(\rho_{n(g)} + \psi_{z(g)}) \quad \forall g \in G \\ & -\gamma_l^- + \gamma_l^+ + \phi_l - \rho_{n(l)} + \rho_{m(l)} = 0 \quad \forall l \in L \\ & \sum_{l \in L(n,\cdot)} B_l(1-u_l)\phi_l + \sum_{l \in L(\cdot,n)} B_l(1-u_l)\phi_l = 0 \quad \forall n \in N \\ & \sigma \ge 0,\; \gamma \ge 0,\; u \in \{0,1\}^{|G|+|L|},\; \|u\|_1 \le 1 \end{aligned} \tag{3.11}$$

Then, we employ a net position oracle based on distance $NPOD$ in the implementation of Alg. 5, which performs the following operations at every query point $\bar{p}$:

1. Compute $\tilde{w} := d(\bar{p}, \mathcal{P}_{N-1}^{FB-EP})$ and obtain a subgradient $v \in \partial_p d(\bar{p}, \mathcal{P}_{N-1}^{FB-EP})$ ($v := \psi^*$).
2. If $\tilde{w} = 0$, then return `TRUE`, $(0_{|Z|}, 0)$.
3. Else return `FALSE`, $(v, -\tilde{w} + v^T \bar{p})$.

Using $NPOD$ we can prove the following (proof presented in section 3.A):

Proposition 4. *Algorithm 5 terminates with an optimal solution in a finite number of iterations when using $NPOD$ as net position oracle.*

In addition to the previous result guaranteeing finite termination, the proposed approach attains practical performance because of the following reasons. (i) We do not solve (3.11) explicitly when calling $NPOD$, but a mixed integer linear reformulation where the products between variables are relaxed using their McCormick envelopes [McC76]. Since u is binary, this is an exact relaxation. (ii) $\mathcal{P}_{N-1}^{FB-EP}$ is a low dimensional set even for realistic systems (the dimension of $\mathcal{P}_{N-1}^{FB-EP}$ is at most $|Z|-1$, and $|Z|=4$ for the CWE system). This allows the cutting-plane algorithm to converge in very few iterations (less than 10 for the CWE system).

3.3.2 Decomposition algorithm for computing maximum volume available-transfer-capacities with N-1 security

Our approach for computing the maximum volume available-transfer-capacities is analogous to one presented in the previous subsection for FBMC: we solve (3.6) – (3.8) by replacing the inclusion constraint $[-ATC^-, ATC^+] \subseteq \mathcal{E}_{N-1}^{EP}$ by a set of linear inequalities that are generated iteratively. Geometrically, the algorithm constructs an outer approximation of $\mathcal{E}_{N-1}^{EP}$ via linear inequalities, so that a maximum-volume rectangle inscribed in the outer approximation is also a maximum-volume rectangle inscribed in $\mathcal{E}_{N-1}^{EP}$.

We formally define $\mathcal{E}_{N-1}^{EP} = \left\{ e \in [-E, E] \middle| \exists p \in \mathcal{P}_{N-1}^{FB-EP}, p = -Qe \right\}$, where $E_t = \sum_{l \in L(t)} F_l$ for all $t \in T$ and Q is the incidence matrix of the graph defined by the zones Z (vertices) and the interconnectors T (edges). Observe that $\mathcal{E}_{N-1}^{EP}$ is a convex polytope because $\mathcal{P}_{N-1}^{FB-EP}$ is a convex polytope (it corresponds to the intersection of finitely many polytopes). Therefore, we can describe $\mathcal{E}_{N-1}^{EP}$ using finitely many inequalities, i.e. for certain $A \in \mathbb{R}^{M \times |T|}$, $B \in \mathbb{R}^M$ we have $Ae \leq B \iff e \in \mathcal{E}_{N-1}^{EP}$. While A, B are not available, we can construct them iteratively. This line of reasoning leads to Alg. 6, which starts from a box outer approximation of $\mathcal{E}_{N-1}^{EP}$ and iteratively refines the approximation by means of two oracles:

- A maximum volume oracle $MVO(A, B)$ which, for given $A \in \mathbb{R}^{M \times |T|}$ and $B \in \mathbb{R}^M$, finds a maximum-volume rectangle inscribed in $\{e \in \mathbb{R}^{|T|} \mid Ae \leq B\}$ and returns ATC^{-*}, ATC^{+*}. In other words, the maximum

Algorithm 6 Cutting-plane algorithm for volume-maximizing ATCs.

1: Initialize $A := [I_{|T|\times|T|} \ -I_{|T|\times|T|}]^\top, B := [E^\top \ E^\top]^\top$, `inclusion` := `FALSE`

2: **while** !`inclusion` **do**
3: Call $MVO(A, B) \to ATC^-, ATC^+$
4: Call $EO(ATC^-, ATC^+) \to$ `inclusion`, (a, b)
5: $A := [A^\top \ a]^\top, \ B := [B^\top \ b]^\top$
6: **end while**
7: Terminate: ATC^-, ATC^+ are volume-maximizing ATCs.

volume oracle solves the following convex mathematical program

$$\begin{aligned} \max_{ATC} \ & \sum_{t\in T} \log\left(ATC_t^- + ATC_t^+\right) \\ \text{s.t. } & A^+ ATC^+ - A^- ATC^- \leq B, \end{aligned}$$

where $a_{ij}^+ = \max\{0, a_{ij}\}$ and $a_{ij}^- = \min\{0, a_{ij}\}$ [BV04].

- An exchange oracle $EO(ATC^-, ATC^+)$ which, for given ATCs, ATC^-, ATC^+, either certifies that $[-ATC^-, ATC^+] \subseteq \mathcal{E}_{N-1}^{EP}$ or returns a hyperplane, defined in terms of slope $a \in \mathbb{R}^{|T|}$ and intercept $b \in \mathbb{R}$. This hyperplane separates a subset of $[-ATC^-, ATC^+]$ from $\mathcal{E}_{N-1}^{EP}$.

As in the case of the algorithm for FBMC, the challenge in attaining practical performance with Alg. 6 is in designing an effective exchange oracle. An exchange oracle that checks independently all vertices of the rectangle $[-ATC^-, ATC^+]$ for all possible contingencies would consume a prohibitive amount of time on each iteration, rendering Alg. 6 ineffective. Luckily, we can check all these combinations implicitly within a single mathematical program. In order to construct such a mathematical program, we first note that

$$[-ATC^-, ATC^+] \subseteq \mathcal{E}_{N-1}^{EP} \iff d(-Qe, \mathcal{P}_{N-1}^{FB-EP}) = 0 \ \forall e \in [-ATC^-, ATC^+],$$

where $d\left(\cdot, \mathcal{P}_{N-1}^{FB-EP}\right)$ is defined according to (3.11) and that $-E \leq -ATC^- \leq ATC^+ \leq E$ by construction of Alg. 6. Further, by convexity of $\mathcal{P}_{N-1}^{FB-EP}$, $d(Qe, \mathcal{P}_{N-1}^{FB-EP}) = 0$ for e at all the vertices of $[-ATC^-, ATC^+]$ is a necessary and sufficient condition for $[-ATC^-, ATC^+] \subseteq \mathcal{E}_{N-1}^{EP}$. Then, we can generate valid separating hyperplanes using the following distance function:

$$\Delta\left(ATC^-, ATC^+\right) = \max_{s\in\{0,1\}^{|T|}} d(-Q(-ATC^- + s\odot(ATC^+ + ATC^-)), \mathcal{P}_{N-1}^{FB-EP}), \tag{3.12}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication of vectors). $\Delta\left(ATC^-, ATC^+\right)$ corresponds to the maximum distance between (i)

the zonal net position at a vertex of $[-ATC^-, ATC^+]$ and (ii) the feasible set of net positions. Based on these observations, we propose an exchange oracle based on the distance $\Delta(\cdot)$, EOD, performing the following operations at every query point $A\bar{T}C^-, A\bar{T}C^+$:

1. Compute $\tilde{b} := \Delta(A\bar{T}C^-, A\bar{T}C^+)$ and record s^*, ψ^* (see (3.11), (3.12)).
2. If $\tilde{b} = 0$, then return `TRUE`, $(0_{|T|}, 0)$.
3. Else return `FALSE`, $(-\psi^{*\top}Q, -\tilde{b} - \psi^{*\top}Q(-A\bar{T}C^- + s^* \odot (A\bar{T}C^+ + A\bar{T}C^-)))$.

We can then show the following (proof presented in section 3.A):

Proposition 5. *Algorithm 6 terminates with an optimal solution in a finite number of iterations when using EOD as exchange oracle.*

The underlying mathematical problem of $\Delta(A\bar{T}C^-, A\bar{T}C^+)$ can be cast as a bi-linear problem, where products between variables can be replaced by their McCormick envelopes, leading to a mixed integer linear program.

3.4 Two-settlement market clearing models

In this section we extend the models presented in section 3.2 to a two-stage setting where the first stage takes place in day ahead and consist of clearing energy and reserve markets simultaneously, while the second stage takes place in real time and corresponds to the actual operation of the system. Multi-settlement systems allow market participants to update their positions in the market as new information becomes available. In the context of electricity markets, new information can come from improved modeling of the grid [KO02, KO04], changes in forecasts of renewable production or residential demand, and forced outages.

While it is not the general case, certain multi-settlement systems can create systematic arbitrage opportunities, allowing market participants to earn a profit without actually producing/consuming in real time. Multi-settlement zonal electricity markets, in particular, can and have been gamed, perhaps most notoriously by the use of the *dec-game* by participants of the California Power Exchange in the early 2000's, [Fed07]. These effects have also been studied in the literature using game theoretic models and small network examples [KO02, KO04]. In this work, in order to permit the simulation of realistic systems, and as done section 3.2, we assume that participants do not game the transmission markets and that, instead, they act as price takers, by submitting their real costs to the corresponding auctions. We ignore inter-temporal constraints and we model only 1-hour snapshots. We use the two-settlement system presented in Fig. 3.4. For both day-ahead and real-time market models, we consider 3 types of market participants:

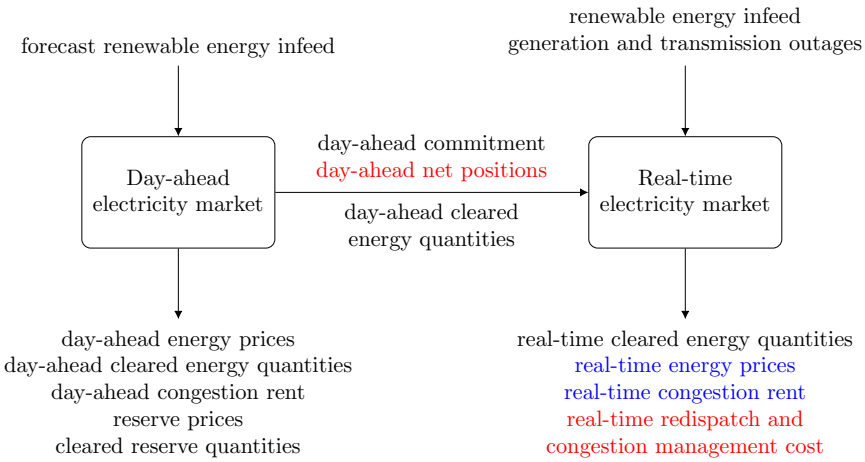


Figure 3.4: Two-settlement system. The day-ahead market is cleared based on a forecast of the renewable energy infeed, assuming that all generation and transmission elements are available. Reserve margins are imposed in order to cope with uncertainty. The real-time market is cleared using the actual renewable infeed and forced outages of generation and transmission elements. We indicate common features of nodal and zonal electricity markets using black, features only present in nodal markets using blue and features only present in zonal markets using red.

1. Consumers, who submit continuous bids for energy Q_c, P_c (demand response) to both day-ahead and real-time markets.

2. Fast generators, i.e. generators that can be started on short notice. These resources submit continuous bids for energy $Q_i, P_i, \forall i \in I(g)$ to both day-ahead and real-time markets, and submit their reserve capability Q_g^R to the day-ahead market.

3. Slow generators, i.e. thermal generators that can only be started on a 24-hour or longer notice. In the day-ahead market, slow generators submit a linked family of bids consisting of (i) a block bid associated with their minimum production level Q_g^L, K_g, (ii) continuous bids for energy $Q_i, P_i, \forall i \in I(g)$ and (iii) their reserve capability Q_g^R. These continuous bids for energy and the reserve can only be accepted if the block bid is accepted. In the real-time market, slow generators submit continuous bids for energy only if they were committed in the day-ahead market.

We compute day-ahead prices following the rules of European electricity markets for continuous bids and linked families of bids [EGN+16, MV15] with two adaptations: we model the clearing of energy and reserves simultaneously,

and we allow continuous bids to be children within linked bid families. These rules and adaptations result in the following restrictions over day-ahead prices:

1. Continuous bids must be accepted if they are in-the-money, must be rejected if they are out-of-the-money and can be partially accepted only if they are at-the-money.

2. Linked families can be accepted only if the linked family is in-the-money or at-the-money. A continuous bid i within a linked family must respect restriction 1, conditional on the acceptance of its parent $g(i)$, i.e. i must follow restriction 1 if $g(i)$ is accepted and i must be rejected if $g(i)$ is rejected (even if it is in-the-money).

Considering these features, along with the N-1 security criterion for zonal markets (described in the previous section) and the features presented in Fig. 3.4, we clear day-ahead markets by solving MILPs and real-time markets by solving LPs. The set of features modeled in the present study have not been simultaneously accounted for both by industry [AAC+15] and previous academic studies on the subject [WRH09, WRH10, MLTW13, Die17]. In what follows, we first list the nomenclature used in the clearing models, and then we introduce the nodal electricity market model and the zonal electricity market models.

Nomenclature

Sets

G	generators
I	production continuous bids
C	consumers
N	buses
L	lines
Z	bidding zones
T	interconnectors
R	reserves
$G(n), G(z)$	generators at node n, zone z
$G(r)$	generators providing reserve r
$I(n), I(z)$	production bids at bus n, zone z
$I(g)$	production bids of generator g
$C(n), C(z)$	consumers at bus n, zone z
$L(m, n)$	lines from bus m to bus n
$L(t)$	lines in interconnector t
$T(y, z)$	interconnectors from zone y to zone z
G^{SLOW}	slow generators
G^{FAST}	fast generators
L^{RT}	lines available in real time

Parameters

$P_\cdot, Q_\cdot$	price-quantity pair for a bid
$P_\cdot^{RT}, Q_\cdot^{RT}$	price-quantity pair for a bid in real time
K_g, Q_g^L	cost and quantity of block bid for technical minimum of generator g
$Q_g^{L,RT}$	post-contingency quantity of block bid of generator g
Q_g^R	reserve capability, generator g
Q_g^U	total capacity, generator g

F_l, B_l	thermal capacity and susceptance, line l	x_c	acceptance of demand response bid c
$z(g), n(g)$	zone and node, generator g	y_r	acceptance of reserve response bid r
$r(g)$	reserve of generator g	f_l	flow on line l
$z(c), n(c)$	zone and node of consumer c	θ_n	voltage angle at bus n
$m(l), n(l)$	node from and node to, line l	p_z	net position of zone z
$a(t), b(t)$	zone from and zone to, interconnector t	e_t	exchange through interconnector t
$V_{f,z}$	coefficient of zone z in facet f of $\mathcal{P}_{N-1}^{FB-EP}$	σ_g	surplus of slow generator g
W_f	constant term of facet f of $\mathcal{P}_{N-1}^{FB-EP}$	ς_i	surplus of production bid i
		τ_g	surplus of reserve bid of generator g
Variables		σ_c	surplus of demand response bid c
u_g	acceptance of block bid of generator g	ρ_n, ρ_z	energy price at node n, zone z
v_i	acceptance of production bid i	π_r	price for reserve r
w_g	acceptance of reserve procurement by generator g	γ, ϕ	dual variables of transmission constraints

3.4.1 Two-settlement nodal electricity market models

Two-settlement nodal electricity market models commonly consist of a day-ahead unit commitment (UC) model and a real-time optimal power flow (OPF) model for real time. While the real-time model that we employ is standard in the literature, our day-ahead model includes additional variables and constraints that ensure the existence of prices that respect European electricity market restrictions. These variables and constraints have been derived using duality theory, following the framework of Madani and Van Vyve [MV15]. In short, they correspond to a modified version of the dual variables and constraints of the linear relaxation of the unit commitment problem. The reader is referred to section 3.B for the full derivation of our day-ahead market clearing models.

3.4.1.1 Day-ahead market model

First, we define sets $\mathcal{D}$ describing the feasible domains for each type of agent. These sets will be common across our different clearing models for the day-ahead market.

The feasible domain for each consumer $c \in C$, presented in (3.13), allows the

energy bid to be partially accepted and ensures that its surplus is non-negative.

$$\mathcal{D}_c := \left\{(x, \sigma, \rho) \in [0,1] \times \mathbb{R}_+ \times \mathbb{R} \;\middle|\; \sigma \geq Q_c(\rho - P_c)\right\} \tag{3.13}$$

Relation (3.14) defines the feasible domain $\mathcal{D}_g$ for fast generators $g \in G^{FAST}$. It allows for continuous bids for energy and reserve to be partially accepted, it ensures that the associated surplus of each bid is non-negative and that the total capacity accepted does not surpass the maximum capacity of the generator.

$$\begin{aligned}\mathcal{D}_g := \Bigg\{ & (v, w, \varsigma, \tau, \nu, \rho, \pi) \in [0,1]^{|I(g)|} \times [0,1] \times \mathbb{R}_+^{|I(g)|} \times \\ & \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \;\Bigg|\; \varsigma_i \geq Q_i(\rho - \nu - P_i)\ \forall i \in I(g), \\ & \tau \geq Q_g^R(\pi - \nu),\ \sum_{i \in I(g)} Q_i v_i + Q_g^R w \leq Q_g^U \Bigg\}\end{aligned} \tag{3.14}$$

The feasible domain $\mathcal{D}_g$ for slow generators $g \in G^{SLOW}$, defined in relation (3.15), enforces that continuous bids for energy and reserve can be partially accepted only if the parent block bid is accepted ($u = 1$), it ensures that the associated surplus of each continuous bid is non-negative and that the total capacity accepted does not surpass the maximum capacity of the generator. The surplus of the parent block bid σ consists of the surplus of the block bid (Q_g^L, K_g) plus the surplus of the child bids in the linked family ($\varsigma_i\ \forall i \in I(g)$ and τ), i.e. σ accumulates the entire surplus of the linked family. The parameter M_g has to be large enough so that σ can be made zero by setting $u = 0$, even when the family is in-the-money (i.e. when $Q_g^L \rho + \sum_{i \in I(g)} \varsigma_i + \tau - Q_g^L \nu - K_g > 0$).

$$\begin{aligned}\mathcal{D}_g := \Bigg\{ & (u, v, w, \sigma, \varsigma, \tau, \nu, \rho, \pi) \in \{0,1\} \times \mathbb{R}_+^{|I(g)|} \times \mathbb{R}_+ \times \\ & \mathbb{R}_+ \times \mathbb{R}_+^{|I(g)|} \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \;\Bigg|\; v_i \leq u\ \forall i \in I(g), \\ & w \leq u,\ \varsigma_i \geq Q_i(\rho - \nu - P_i)\ \forall i \in I(g), \\ & \tau \geq Q_g^R(\pi - \nu),\ Q_g^L u + \sum_{i \in I(g)} Q_i v_i + Q_g^R w \leq Q_g^U, \\ & \sigma \geq Q_g^L \rho + \sum_{i \in I(g)} \varsigma_i + \tau - Q_g^L \nu - K_g - M_g(1-u) \Bigg\}\end{aligned} \tag{3.15}$$

In order to ensure feasibility of the clearing model, we cast system operators as additional agents whom submit bids for reserve response Q_r, P_r for each corresponding control area. The feasible domain for reserves is accordingly

defined as (3.16), where we enforce reserve balance and that the surplus of reserve response bids is non-negative.

$$\mathcal{D}_r := \Bigg\{ (w, y, \sigma, \pi) \in \mathbb{R}^{|G(r)|} \times [0,1] \times \mathbb{R}_+ \times \mathbb{R}_+ \;\Bigg|\; \sum_{g \in G(r)} Q_g^R w_g \geq Q_r(1-y),\ \sigma \geq Q_r(\pi - P_r) \Bigg\} \tag{3.16}$$

Using sets $\mathcal{D}_c, \mathcal{D}_g$ and $\mathcal{D}_r$, the nodal day-ahead market clearing problem can be formulated as (3.17) – (3.23).

$$\min_{\substack{u,v,w,x,y,f,\theta \\ \sigma,\varsigma,\tau,\nu,\gamma,\phi,\rho,\pi}} \quad \sum_{g \in G^{SLOW}} K_g u_g + \sum_{i \in I} P_i Q_i v_i + \sum_{c \in C} P_c Q_c x_c + \sum_{r \in R} P_r Q_r y_r \tag{3.17}$$

$$\text{s.t.} \quad \left(u_g, v_{I(g)}, w_g, \sigma_g, \varsigma_{I(g)}, \tau_g, \nu_g, \rho_{n(g)}, \pi_{r(g)}\right) \in \mathcal{D}_g \quad \forall g \in G^{SLOW},$$

$$\left(v_{I(g)}, w_g, \varsigma_{I(g)}, \tau_g, \nu_g, \rho_{n(g)}, \pi_{r(g)}\right) \in \mathcal{D}_g \quad \forall g \in G^{FAST} \tag{3.18}$$

$$\left(x_c, \sigma_c, \rho_{n(c)}\right) \in \mathcal{D}_c \quad \forall c \in C \tag{3.19}$$

$$\left(w_{G(r)}, y_r, \sigma_r, \pi_r\right) \in \mathcal{D}_r \quad \forall r \in R \tag{3.20}$$

$$\sum_{l \in L(n,\cdot)} f_l - \sum_{l \in L(\cdot,n)} f_l = \sum_{g \in G^{SLOW}(n)} Q_g^L u_g + \sum_{i \in I(n)} Q_i v_i - \sum_{c \in C(n)} Q_c(1 - x_c)\ ,\ \sum_{l \in L(n,\cdot)} B_l \phi_l = \sum_{l \in L(\cdot,n)} B_l \phi_l \quad \forall n \in N \tag{3.21}$$

$$-F_l \leq f_l \leq F_l,\ f_l = B_l\left(\theta_{m(l)} - \theta_{n(l)}\right),\ \gamma_l^- \geq 0,\ \gamma_l^+ \geq 0,$$

$$\gamma_l^+ - \gamma_l^- = \rho_{n(l)} - \rho_{m(l)} - \phi_l \quad \forall l \in L \tag{3.22}$$

$$\sum_{g \in G^{SLOW}} K_g u_g + \sum_{i \in I} P_i Q_i v_i + \sum_{c \in C} P_c Q_c x_c + \sum_{r \in R} P_r Q_r y_r \leq \sum_{n \in N} \rho_n \cdot \sum_{c \in C(n)} Q_c + \sum_{r \in R} Q_r \pi_r - \Bigg(\sum_{g \in G^{SLOW}} \sigma_g + \sum_{g \in G^{FAST}} \Bigg(\sum_{i \in I(g)} \varsigma_i + \tau_g \Bigg) + \sum_{g \in G} Q_g^U \nu_g + \sum_{c \in C} \sigma_c + \sum_{r \in R} \sigma_r + \sum_{l \in L} F_l\left(\gamma_l^+ + \gamma_l^-\right) \Bigg) \tag{3.23}$$

The objective function (3.17) consists of the total cost of production, demand response and reserve response. Constraints (3.18) – (3.20) impose the respective domains for generators, consumers and reserves. Constraints (3.21)

– (3.22) impose primal and dual constraints of the nodal power grid, including nodal power balance, DC power flow constraints, thermal limits on lines and dual restrictions on LMPs. Constraint (3.23) enforces equality of social welfare and total surplus, ensuring that European restrictions on prices are respected (see section 3.B).

3.4.1.2 Real-time market model

Following Fig. 3.4, in real time the system is bound to the commitment status decided using the day-ahead market clearing model, $\bar{u}^{DA}$, for all slow generators. Additionally, the realized renewable injections and generation outages can modify the quantities offered in the real-time market Q^{RT} with respect to what is offered in the day-ahead market Q. Transmission line outages modify the topology of the network, leaving only lines in $L^{RT} \subseteq L$ energized in real-time operation. We model the real-time market following these alterations in the power grid as problem (3.24) – (3.30), where we have removed all constant quantities cleared in the day-ahead market.

$$\min_{u,v,x,f,\theta} \sum_{i \in I} P_i Q_i^{RT} v_i + \sum_{c \in C} P_c Q_c x_c \tag{3.24}$$

$$\text{s.t.} \quad 0 \leq u_g \leq \bar{u}_g^{DA} \quad \forall g \in G^{SLOW} \tag{3.25}$$

$$0 \leq v_i \leq \bar{u}_g^{DA} \quad \forall i \in I(g), g \in G^{SLOW} \tag{3.26}$$

$$0 \leq v_i \leq 1 \quad \forall i \in I(g), g \in G^{FAST} \tag{3.27}$$

$$0 \leq x_c \leq 1 \quad \forall c \in C \tag{3.28}$$

$$\sum_{l \in L^{RT}(n,\cdot)} f_l - \sum_{l \in L^{RT}(\cdot,n)} f_l = \sum_{g \in G^{SLOW}(n)} Q_g^{L,RT} u_g + \sum_{i \in I(n)} Q_i^{RT} v_i - \sum_{c \in C(n)} Q_c (1 - x_c) \quad \forall n \in N \quad [\rho_n] \tag{3.29}$$

$$-F_l \leq f_l \leq F_l, \; f_l = B_l \left(\theta_{m(l)} - \theta_{n(l)}\right) \quad \forall l \in L^{RT} \tag{3.30}$$

The objective function (3.24) corresponds to real-time cost, comprising generation and demand response. Constraints (3.25), (3.26) restrict the production of slow generators in accordances with the results of the day-ahead market. Constraints (3.27) – (3.30) model fast generators, consumers and the available transmission grid. Real-time energy prices are obtained as the dual of the balance constraint (3.29) and are used to clear real-time quantities.

3.4.2 Zonal electricty markets with two settlements

We model two-settlement zonal markets following Fig. 3.4, which is itself based on the existing literature [ES05]: the day-ahead market is cleared using a zonal

model for the grid, and system operators perform congestion management and balancing in real time.

The day-ahead models presented in this section differ from the simplified zonal models of section 3.2 in that (i) they consider the commitment decisions of slow generators and (ii) they consider all N-1 contingencies when determining the feasible set for zonal net positions (FBMC) or inter-zonal exchanges (ATCMC). In order to solve the mathematical programs that ensure N-1 robustness, we employ the algorithms presented in section 3.3.1 for FBMC and in section 3.3.2 for ATCMC.

In what follows, we present the day-ahead models for FBMC and ATCMC, and then we introduce the real-time congestion management and balancing (CM&B) model, common for both FBMC and ATCMC.

3.4.2.1 Day-ahead flow-based market coupling with exact projection

The day-ahead market clearing problem under FBMC can be cast as the mathematical program (3.31) – (3.38), where we use the domains defined in (3.13) – (3.16) for generators, consumers and reserves.

$$\min_{\substack{u,v,w,x,y,p\\ \sigma,\varsigma,\tau,\nu,\phi,\gamma,\rho,\pi}} \quad \sum_{g\in G^{SLOW}} K_g u_g + \sum_{i\in I} P_i Q_i v_i + \sum_{c\in C} P_c Q_c x_c + \sum_{r\in R} P_r Q_r y_r \tag{3.31}$$

$$\text{s.t.} \quad \big(u_g, v_{I(g)}, w_g, \sigma_g, \varsigma_{I(g)}, \tau_g, \nu_g, \rho_{z(g)}, \pi_{r(g)}\big) \in \mathcal{D}_g^{SLOW} \quad \forall g \in G^{SLOW},$$
$$\big(v_{I(g)}, w_g, \varsigma_{I(g)}, \tau_g, \nu_g, \rho_{z(g)}, \pi_{r(g)}\big) \in \mathcal{D}_g^{FAST} \quad \forall g \in G^{FAST} \tag{3.32}$$

$$\big(x_c, \sigma_c, \rho_{n(c)}\big) \in \mathcal{D}_c \quad \forall c \in C \tag{3.33}$$

$$\big(w_{G(r)}, y_r, \sigma_r, \pi_r\big) \in \mathcal{D}_r \quad \forall r \in R \tag{3.34}$$

$$p_z = \sum_{g\in G^{SLOW}(z)} Q_g^L u_g + \sum_{i\in I(z)} Q_i v_i - \sum_{c\in C(z)} Q_c(1-x_c),$$
$$\rho_z + \phi + \sum_{f\in\mathcal{F}(\mathcal{P}_{N-1}^{FB-EP})} V_{f,z}\gamma_f = 0 \quad \forall z \in Z \tag{3.35}$$

$$\sum_{z\in Z} p_z = 0 \tag{3.36}$$

$$\sum_{z\in Z} V_{f,z} p_z \le W_f,\ \gamma_f \ge 0 \quad \forall f \in \mathcal{F}(\mathcal{P}_{N-1}^{FB-EP}) \tag{3.37}$$

$$\sum_{g\in G^{SLOW}} K_g u_g + \sum_{i\in I} P_i Q_i v_i + \sum_{c\in C} P_c Q_c x_c + \sum_{r\in R} P_r Q_r y_r \leq$$
$$\sum_{z\in Z} \rho_z \cdot \sum_{c\in C(z)} Q_c + \sum_{r\in R} Q_r \pi_r - \Bigg(\sum_{g\in G^{SLOW}} \sigma_g -$$
$$\sum_{g\in G^{FAST}} \Bigg(\sum_{i\in I(g)} \varsigma_i + \tau_g \Bigg) + \sum_{g\in G} Q_g^U \nu_g + \sum_{c\in C} \sigma_c +$$
$$\sum_{r\in R} \sigma_r + \sum_{f\in \mathcal{F}(\mathcal{P}_{N-1}^{FB-EP})} W_f \gamma_f \Bigg) \tag{3.38}$$

The objective function (3.31) and the constraints of participants (3.32), (3.33) and reserves (3.34) are analogous to those of the nodal day-ahead clearing problem (3.17) – (3.23). The only difference is that here we use zonal prices instead of nodal prices for energy. Constraints (3.35) – (3.37) model primal and dual constraints for zonal network model. Constraints (3.35) define zonal net positions p_z and zonal prices ρ_z. Constraint (3.36) enforces system-wide power balance. Constraints (3.35) enforce that zonal net positions belong in the flow based domain $\mathcal{P}_{N-1}^{FB-EP}$, by enumerating the facets of $\mathcal{P}_{N-1}^{FB-EP}$ denoted as $\mathcal{F}(\mathcal{P}_{N-1}^{FB-EP})$. Constraints (3.35) also enforce non-negativity of the variables γ_f, associated with the sensitivity of the objective to each facet $f \in \mathcal{F}(\mathcal{P}_{N-1}^{FB-EP})$. Finally, constraint (3.38) enforces equality of social welfare and total surplus, ensuring the existence of European prices.

Formulation (3.31) – (3.38) uses an exponential number of constraints (3.37) and variables γ_f to define the flow-based domain explicitly, because $\mathcal{P}_{N-1}^{FB-EP}$ results from the projection of all the nodal equations of the system under full availability and all N-1 conditions into the space of net positions. In order to solve (3.31) – (3.38) efficiently for realistic systems we use the same algorithm as in section 3.3.1 with a market clearing oracle (MCO) that solves problem (3.31)–(3.38) with matrices V, W generated by the cutting-plane procedure instead of using the facets of $\mathcal{P}_{N-1}^{FB-EP}$. Staring from a lose outer-approximation, each iteration of the cutting-plane procedure adds a new constraint of the type (3.35) and re-solves the market clearing problem. Note that, adding a new row to the system $Vp \leq W$, requires also adding a new variable γ associated to that row in the MCO. In other words, in the presence of block bids and European pricing rules, our cutting-plane procedure becomes a column-and-constraint generation algorithm. Regarding the net position oracle, we use the same definition of $\mathcal{P}_{N-1}^{FB-EP}$ as in section 3.3.1, ignoring the technical minimums of slow thermal generators in order to ensure that $\mathcal{P}_{N-1}^{FB-EP}$ is convex.

3.4.2.2 Day-ahead available-transfer-capacity market coupling

Recall that clearing the day-ahead market using ATCMC requires two steps. First, we compute ATCs by solving the mathematical program (3.6) – (3.8) with the cutting-plane algorithm described in section 3.3.2. Then, using the computed ATCs, we clear the day-ahead energy market by solving the mathematical program (3.39) – (3.45), where we use (again) the domains defined in (3.13) – (3.16) for generators, consumers and reserves.

$$\min_{\substack{u,v,w,x,y,e\\ \sigma,\varsigma,\tau,\nu,\gamma,\rho,\pi}} \sum_{g\in G^{SLOW}} K_g u_g + \sum_{i\in I} P_i Q_i v_i + \sum_{c\in C} P_c Q_c x_c + \sum_{r\in R} P_r Q_r y_r \tag{3.39}$$

$$\text{s.t.} \quad \left(u_g, v_{I(g)}, w_g, \sigma_g, \varsigma_{I(g)}, \tau_g, \nu_g, \rho_{z(g)}, \pi_{r(g)}\right) \in \mathcal{D}_g^{SLOW} \quad \forall g \in G^{SLOW},$$

$$\left(v_{I(g)}, w_g, \varsigma_{I(g)}, \tau_g, \nu_g, \rho_{z(g)}, \pi_{r(g)}\right) \in \mathcal{D}_g^{FAST} \quad \forall g \in G^{FAST} \tag{3.40}$$

$$\left(x_c, \sigma_c, \rho_{n(c)}\right) \in \mathcal{D}_c \quad \forall c \in C \tag{3.41}$$

$$\left(w_{G(r)}, y_r, \sigma_r, \pi_r\right) \in \mathcal{D}_r \quad \forall r \in R \tag{3.42}$$

$$\sum_{t\in T(z,\cdot)} e_t - \sum_{t\in T(\cdot,z)} e_t = \sum_{g\in G^{SLOW}(z)} Q_g^L u_g + \sum_{i\in I(z)} Q_i v_i - \sum_{c\in C(z)} Q_c(1-x_c) \quad \forall z \in Z \tag{3.43}$$

$$-ATC_t^- \le e_t \le ATC_t^+,\ \gamma_t^- \ge 0,\ \gamma_t^+ \ge 0,$$
$$\gamma_t^+ - \gamma_t^- = \rho_{b(t)} - \rho_{a(t)} \quad \forall t \in T \tag{3.44}$$

$$\sum_{g\in G^{SLOW}} K_g u_g + \sum_{i\in I} P_i Q_i v_i + \sum_{c\in C} P_c Q_c x_c + \sum_{r\in R} P_r Q_r y_r \le \sum_{z\in Z} \rho_z \cdot \sum_{c\in C(z)} Q_c + \sum_{r\in R} Q_r \pi_r - \Bigg(\sum_{g\in G^{SLOW}} \sigma_g - \sum_{g\in G^{FAST}} \Bigg(\sum_{i\in I(g)} \varsigma_i + \tau_g \Bigg) + \sum_{g\in G} Q_g^U \nu_g + \sum_{c\in C} \sigma_c + \sum_{r\in R} \sigma_r + \sum_{t\in T} \left(ATC_t^+ \gamma_t^+ + ATC_t^- \gamma_t^- \right) \Bigg) \tag{3.45}$$

The objective function (3.39) and constraints (3.40) – (3.42) are identical to those of the FBMC day-ahead clearing model (3.31) – (3.38). Constraints (3.43), (3.44) enforce primal and dual constraints over the transportation network between zones, with exchanges through interconnectors limited by the ATCs. Constraints (3.45) enforce strong duality between the primal and dual parts of the clearing problem, thereby, ensuring that ρ^* are supporting prices of the clearing decisions.

3.4.2.3 Real-time congestion management and balancing

We model real-time CM&B as the mathematical program (3.46) – (3.47), which is essentially the same as (3.24) – (3.30). The only difference in terms of modeling is that in (3.46) we penalize deviations from day-ahead net positions $\bar{p}^{DA}$ at $\hat{P} := \max_{i \in I} P_i$. This is done in order to account for the restriction that market participants within each zone must remain in balance in real time [Eur17a, Article 17], while maintaining feasibility of the real-time model. On the hand, there is an important difference with (3.24) – (3.30) in terms of pricing: in (3.46) – (3.47) real-time quantities are paid-as-bid [AC17] instead of using locational marginal prices.

$$\min_{u,v,x,f,\theta,p} \sum_{i \in I} P_i Q_i^{RT} v_i + \sum_{c \in C} P_c Q_c x_c + \sum_{z \in Z} \hat{P} \left| p_z - \bar{p}_z^{DA} \right| \tag{3.46}$$

$$\text{s.t.} \quad (3.25) - (3.30) \tag{3.47}$$

$$p_z = \sum_{g \in G^{SLOW}(z)} Q_g^{L,RT} u_g + \sum_{i \in I(z)} Q_i^{RT} v_i - \sum_{c \in C(z)} Q_c (1 - x_c) \quad \forall z \in Z \tag{3.48}$$

3.5 Simulation setup

We use a modified version of the CWE system of section 1.4.3. We remove Switzerland, since it is not part of the current implementation of FBMC, and we add (*i*) forced outage rates for thermal generators based on [Nor16], (*ii*) cost of CO_2 emissions, with emissions rates derived from [MBH+11] and valued at the average intra-day price of emissions in 2013-2017 [EEX18], (*iii*) forced outage rates for transmission lines from [EG15] and (*iv*) existing phase shifters in the borders of Belgium. The resulting system consists of: (*i*) 346 slow generators with a total capacity of 154 GW (including nuclear, and CHP and conventional thermal units with cold startup times of 24 hours or more); (*ii*) 301 fast thermal generators with a total capacity of 89 GW (CHP and conventional thermal units with cold startup time of less than 24 hours and aggregated generators); (iii) 1312 renewable generators with a total capacity of 149 GW; (iv) 632 buses; and (v) 945 branches. The average demand of the system amounts to 134 GW. The topology of the modified power grid and its zonal aggregation is presented in Fig. 3.5.

We consider 768 different typical snapshots for day-ahead market clearing (8 representative days, 96 snapshots per day, see section 1.4.3). Each snapshot corresponds to different demand, renewable forecasts and maintenance schedules (deratings) for thermal generators. For each snapshot, we generate 1 000 random realizations of uncertainty (renewable forecast errors, forced outages of thermal generators and forced outages of transmission lines) in order to simu-

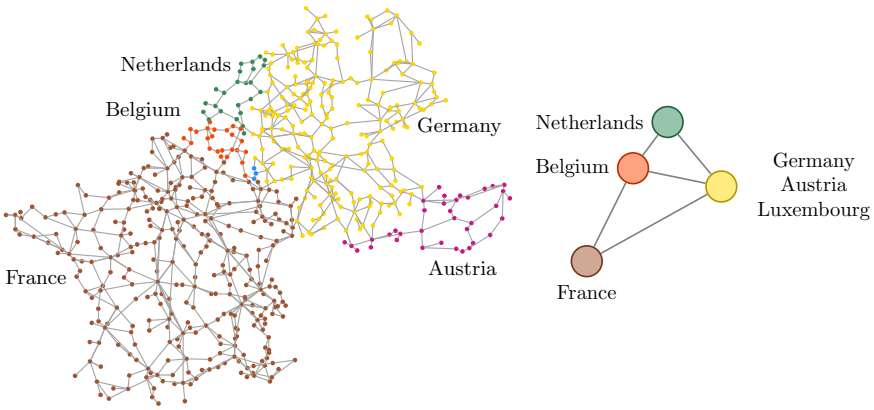


Figure 3.5: CWE network model and zonal aggregation with inter-zonal links used for ATCMC.

late real-time operation. For generators and lines, forced outages are randomly generated based on Bernoulli distributions, i.e. each element is completely unavailable with a probability equal to its forced outage rate, otherwise the element is fully available (up to derating from maintenance schedules). In total, we consider 768 000 different operating conditions for each policy. Considering that each snapshot corresponds to an hour of operation, our simulations amount to approximately 88 years of operation.

We implement the models and algorithms involved in simulating LMP, FBMC and ATCMC (presented in sections 3.3.1, 3.3.2 and 3.4) in Julia 0.6.0 [BEKS17] using JuMP 0.18.0 [DHL17] to define the required mathematical programs. We use Cplex 12.6.2 to solve the mixed-integer linear programs for day-ahead market clearing and Ipopt 3.12.8 [WB06] compiled with CoinHSL [HSL15] in the computation of volume-maximizing ATCs. We use Xpress 8.0.4 to solve the linear programs that simulate real-time operations.

We deploy our algorithms on high-performance computing clusters, using Julia's built-in message passing implementation and parallelizing over different snapshots of operation. We use the Cab cluster, hosted at the Lawrence Livermore National Laboratory, for solving day-ahead market clearing models and the Lemaitre2 cluster, hosted at the Université catholique de Louvain, for solving real-time operation models. The analysis of the results was performed using Julia and R [R C15].

3.6 Results and discussion

We present a comparison between LMP, FBMC and ATCMC. In all cases, we use arithmetic averages to aggregate results across snapshots and median

Table 3.3: Total costs and cost performance comparison between policies. Efficiency losses measured with respect to LMP total cost.

Policy	Day-ahead [M€/year]	Real-time [M€/year]	Total [M€/year]	Efficiency losses
Perfect Foresight	–	11 677	11 677	-0.93%
LMP	10 758	1 029	11 787	–
FBMC	10 693	1 787	12 480	5.88%
ATCMC	10 793	1 746	12 539	6.38%

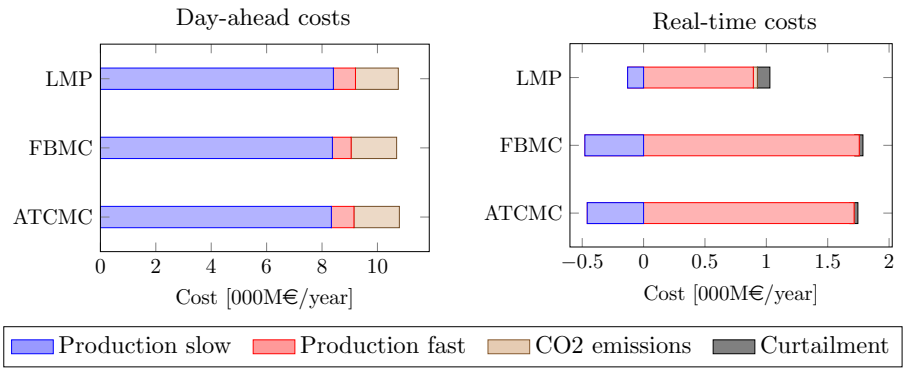


Figure 3.6: Breakdown of costs in day-ahead market and real-time operations for the three policies under study. Real-time costs only account for the differences with respect to day-ahead values.

centers [Gow74] in order to aggregate results across realizations of uncertainty. We use the median center in order to remove the effect of low-probability events from central tendency indicators.

Table 3.3 presents the overall costs of each policy at each stage. The table also contains the Perfect Foresight (PF) benchmark, which is an unreachable performance benchmark. PF and LMP achieve similar performance, which is due to the large capacity margin of the system, and due to the ability of the LMP policy to commit slow units where they are needed. LMP, FBMC and ATCMC attain similar total cost in the day-ahead market, with FBMC outperforming ATCMC by 100 M€/year (in line with the 95 M€/year estimated by the parallel run between ATCMC and FBMC [AAC+15]). On the other hand, the real-time costs of FBMC and ATCMC are notably greater than those of LMP. As indicated in Fig. 3.6, this difference stems mostly from the use of fast thermal generators (e.g. gas and oil) in real time by FBMC and ATCMC. The inefficient deployment of these units drives the difference in cost performance between zonal and nodal markets to about 6% of the total operation costs (which corresponds to approximately 720 M€/year for the CWE).

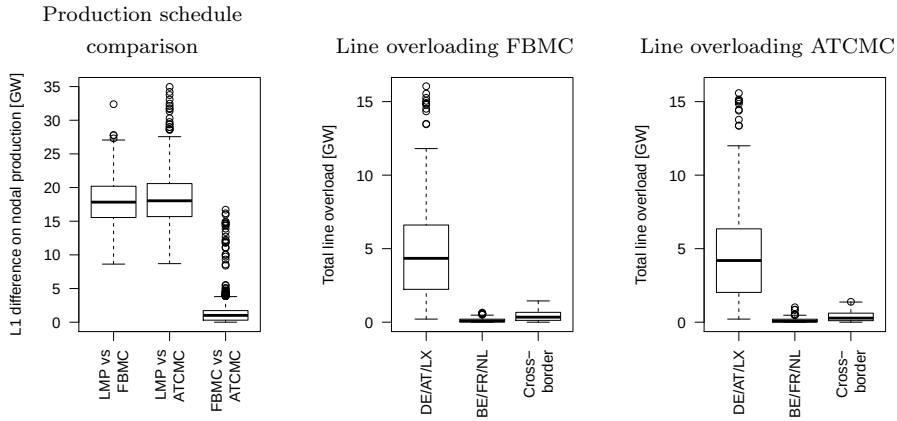


Figure 3.7: The left box plot presents the distribution of the absolute deviation between the day-ahead production schedules of the three different policies over all snapshots. The center and right box plots present the overloading pattern caused by the acceptance/rejection decisions of FBMC and ATCMC.

In addition to achieving a similar cost performance, FBMC and ATCMC result in similar acceptance/rejection decisions and implied flows. This is indicated in Fig. 3.7. The center and right box plots of Fig. 3.7 demonstrate that FBMC and ATCMC overload both internal and inter-zonal lines, with similar overloading patterns. This behavior was already observed in the 4-node network in Tables 3.1 and 3.2, and is attributed to the fact that merit order dispatch drives similar nodal injections in both models, while overlooking the implications for the grid of such injections, effectively failing at allocating transmission capacity.

The differences between FBMC and LMP are driven by the same factors that drive the differences between ATCMC and LMP observed in the literature [ES05] (see also section 4.5 in the next chapter). The major factor contributing to the inefficiency of FBMC is the suboptimal commitment decisions of this policy. In order to isolate this factor, we simulate the real-time market without imposing the requirement of maintaining zonal day-ahead net positions, and record the dual variables of the power balance constraints ρ^{RT}. We observe in Fig. 3.8 that whenever LMP commits a unit at a certain location n in the day ahead and FBMC does not, ρ_n^{RT} becomes larger than the corresponding day-ahead zonal price $\rho_{z(n)}^{DA}$, indicating the need for generation at n. Conversely, whenever FBMC commits at a certain location n in the day ahead and LMP does not, we observe that most frequently ρ_n^{RT} becomes smaller than $\rho_{z(n)}^{DA}$, indicating that generation capacity is not required the given location. FBMC fails at recognizing these locational differences as a direct consequence of the zonal aggregation in the day-ahead market. We quantify the efficiency losses

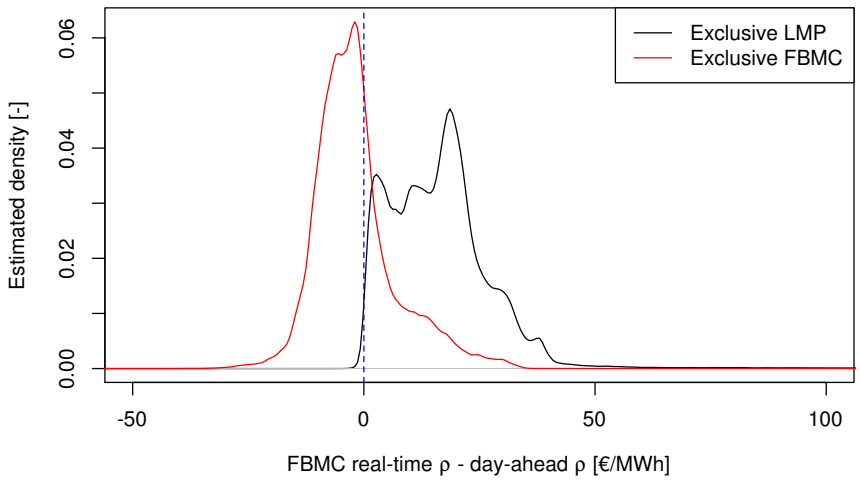


Figure 3.8: Distribution of geographical averages of marginal cost differences between real time and day ahead for FBMC, without enforcing zonal net positions in real time. For a given snapshot and a given realization of uncertainty, the "exclusive LMP" series is computed by averaging the differences $\rho_n^{RT} - \rho_{z(n)}^{DA}$ over all nodes where LMP committed slow units and FBMC did not commit units, weighted by the capacity committed exclusively by LMP. The exclusive FBMC series is computed analogously.

of FMBC caused by suboptimal commitment at 4.28% of the total operation costs. The remaining 1.60% cost difference is explained by the requirement of maintaining day-ahead net positions in the presence of renewable forecast errors in FBMC (balancing each zone independently), whereas LMP is able to exploit cross-zonal balancing.

3.7 Conclusions

We present models for flow-based market coupling and for available-transfer-capacity market coupling that do not depend on arbitrary parameters and are based purely on the technical parameters of the grid. We analyze the implications of these models on a 4-node instance and a realistic-scale instance of the CWE system. In both instances, we observe that both zonal market designs encounter challenges with allocating transmission capacity of inter-zonal lines and intra-zonal lines.

These deficiencies affect the real-time performance of the system, leading to efficiency losses of 5.88% for flow-based and 6.38% for available-transfer-capacity market coupling with respect to a nodal market design. The similarities that are unveiled between both zonal designs raise the question of whether

flow-based market coupling should be expanded to cover other zones of Europe, especially when certain zones such as Poland are already planning to switch to a nodal market design [PSE17].

Future extensions of the present work will focus on the study of systematic differences between day-ahead and real-time markets, caused by unit commitment decisions in the case of nodal markets [Hog16], and by both unit commitment decisions and the aggregated representation of the grid in the case of zonal markets.

Appendix

3.A Proofs for propositions of section 3.3

The proofs of the propositions in section 3.3 are mainly applications of known results and presented here for self-containedness. Our contribution lies on the intermediate results presented as additional propositions in this section, which guarantee the correctness of the application of cutting-plane algorithms and the conditions for convergence of the cutting-plane method.

Proposition 6. *$d(p, \mathcal{P}_{N-1}^{FB-EP})$ is a convex piece-wise linear function of p with finitely many pieces.*

Proof of Prop. 6. The convexity of $d(p, \mathcal{P}_{N-1}^{FB-EP})$ in p follows from the fact that it is the maximum over linear functions of p [BV04] which also implies that $d(p, \mathcal{P}_{N-1}^{FB-EP})$ is piece-wise linear. $d(p, \mathcal{P}_{N-1}^{FB-EP})$ can be described using finitely many pieces because for each fixed feasible u, the feasible set of the remaining variables in (3.11) is a polytope and polytopes can be described using finitely many extreme points and extreme rays. □

Proof of Prop. 4. Identical to the proof of convergence of Benders decomposition for linear programs [Ben62]. Essentially, due to Prop. 6, there exists only a finite number of separating hyperplanes that can be generated using the oracle. Once all these separating hyperplanes have been generated, the market clearing oracle will solve the original clearing problem (3.31) – (3.38), terminating either with the optimal solution or with an infeasibility certificate. □

Proposition 7. *$\Delta(ATC^-, ATC^+)$ is a convex piece-wise linear function of ATC^-, ATC^+ with finitely many pieces.*

Proof of Prop. 7. To see that $\Delta(ATC^-, ATC^+)$ is a convex piece-wise linear function of ATC^-, ATC^+, simply replace the definition of $d(p, \mathcal{P}_{N-1}^{FB-EP})$ given in (3.11) into (3.12) and fuse both max operators ($\Delta(ATC^-, ATC^+)$ is a max-type function over an objective linear in ATC^-, ATC^+ [BV04]). The proof that $\Delta(ATC^-, ATC^+)$ has finitely many pieces follows the same argument used in the proof of Prop. 6. □

Proof of Prop. 5. Identical to proof of Prop. 4, using the result of Prop. 7. □

3.B Derivation of day-ahead electricity market models with strict-linear pricing

The presence of binary commitment variables in day-ahead electricity market models renders the derivation of prices for these markets a non-trivial task.

Equilibrium prices, in the presence of integer decision variables may not exist, depending on the parameters of the market clearing problem at hand, see [Voh11, Chapter 5]. This can lead to (i) paradoxically accepted bids, i.e. bids that are accepted at a loss for the bidding participant, and (ii) paradoxically rejected bids bids, i.e. rejected bids that if accepted would generate a profit for the bidding participant. Paradoxically rejected block bids are allowed and not compensated in most electricity markets [EGN$^+$16, Fed14]. The treatment of paradoxically accepted bids, on the other hand, differs among markets: U.S. markets allow paradoxically accepted bids and cover the losses of bidding participants through uplift payments, [Fed14], while European markets do not allow paradoxically accepted bids, [EGN$^+$16].

In the following we derive a model for the nodal day-ahead market clearing problem respecting the European pricing restrictions, following the MILP framework of Madani and Van Vyve [MV15]. We start by expressing the unit commitment problem in terms of continuous bids, block bids and linked families, (3.49) – (3.63).

$$\min_{u,v,w,x,y,f,\theta} \sum_{g\in G^{SLOW}} K_g u_g + \sum_{i\in I} P_i Q_i v_i + \sum_{c\in C} P_c Q_c x_c + \sum_{r\in R} P_r Q_r y_r \tag{3.49}$$

$$\text{s.t.} \quad u_g \le 1 \qquad \forall g \in G \ [\sigma_g] \tag{3.50}$$

$$v_i \le u_{g(i)} \qquad \forall i \in I \ [\varsigma_i] \tag{3.51}$$

$$w_g \le u_g \qquad \forall g \in G^{SLOW} \ [\tau_g] \tag{3.52}$$

$$w_g \le 1 \qquad \forall g \in G^{FAST} \ [\tau_g] \tag{3.53}$$

$$Q_g^L u_g + \sum_{i\in I(g)} Q_i v_i + Q_g^R w_g \le Q_g^U \qquad \forall g \in G^{SLOW} \ [\nu_g] \tag{3.54}$$

$$\sum_{i\in I(g)} Q_i v_i + Q_g^R w_g \le Q_g^U \qquad \forall g \in G^{FAST} \ [\nu_g] \tag{3.55}$$

$$x_c \le 1 \qquad \forall c \in C \ [\sigma_c] \tag{3.56}$$

$$\sum_{c\in C(n)} Q_c(1-x_c) = \sum_{g\in G(n)} \Big(Q^L u_g + \sum_{i\in I(g)} Q_i v_i \Big) + \sum_{l\in L(\cdot,n)} f_l - \sum_{l\in L(n,\cdot)} f_l \qquad \forall n \in N \ [\rho_n] \tag{3.57}$$

$$-f_l \le F_l \qquad \forall l \in L \ [\gamma_l^-] \tag{3.58}$$

$$f_l \le F_l \qquad \forall l \in L \ [\gamma_l^+] \tag{3.59}$$

$$f_l = B_l \left(\theta_{m(l)} - \theta_{n(l)}\right) \qquad \forall l \in L \ [\phi_l] \tag{3.60}$$

$$y_r \leq 1 \qquad \forall r \in R \ [\sigma_r] \tag{3.61}$$

$$Q_r(1-y_r) - \sum_{g \in G(r)} Q_g^R w_g \leq 0 \qquad \forall r \in R \ [\pi_a] \tag{3.62}$$

$$u, v, w, x, y \geq 0, \ u_g \in \{0,1\} \ \forall g \in G \tag{3.63}$$

The objective function (3.63) corresponds to total cost of operation. This includes production, demand response and reserve response costs. Constraints (3.50) – (3.55) impose that production and reserve bids can be accepted at most once, and that the total capacity of each generator must be respected. Constraints (3.56) – (3.62) enforce limits on demand response bids, nodal power balance, transmission constraints and reserve requirements. Using the dual variables indicated in brackets aside each constraint, we can write the KKT conditions of the linear relaxation of (3.49) – (3.63) as follows:

$$0 \leq \sigma_g \perp u_g - 1 \leq 0 \quad \forall g \in G^{SLOW} \tag{3.64}$$

$$0 \leq \varsigma_i \perp v_i - u_g \leq 0 \quad \forall g \in G^{SLOW}, i \in I(g) \tag{3.65}$$

$$0 \leq \tau_g \perp w_g - u_g \leq 0 \quad \forall g \in G^{SLOW} \tag{3.66}$$

$$0 \leq \nu_g \perp Q_g^L u_g + \sum_{i \in I(g)} Q_i v_i + Q_g^R w_g - Q_g^U \leq 0 \quad \forall g \in G^{SLOW} \tag{3.67}$$

$$0 \leq \varsigma_i \perp v_i - 1 \leq 0 \quad \forall g \in G^{FAST}, i \in I(g) \tag{3.68}$$

$$0 \leq \tau_g \perp w_g - 1 \leq 0 \quad \forall g \in G^{FAST} \tag{3.69}$$

$$0 \leq \nu_g \perp \sum_{i \in I(g)} Q_i v_i + Q_g^R w_g - Q_g^U \leq 0 \quad \forall g \in G^{FAST} \tag{3.70}$$

$$0 \leq \sigma_c \perp x_c - 1 \leq 0 \quad \forall c \in C \tag{3.71}$$

$$\rho_n \perp \sum_{l \in L(n,\cdot)} f_l + \sum_{l \in L(\cdot,n)} f_l - \sum_{g \in G^{SLOW}(n)} Q_g^L u_g - \sum_{i \in I(n)} Q_i v_i + \sum_{c \in C(n)} Q_c(1-x_c) = 0 \quad \forall n \in N \tag{3.72}$$

$$0 \leq \gamma_l^- \perp -f_l - F_l \leq 0 \quad \forall l \in L \tag{3.73}$$

$$0 \leq \gamma_l^+ \perp f_l - F_l \leq 0 \quad \forall l \in L \tag{3.74}$$

$$\phi_l \perp f_l - B_l \left(\theta_{m(l)} - \theta_{n(l)}\right) = 0 \quad \forall l \in L \tag{3.75}$$

$$0 \leq \sigma_r \perp y_r - 1 \leq 0 \quad \forall r \in R \tag{3.76}$$

$$0 \leq \pi_r \perp Q_r(1-y_r) - \sum_{g \in G(r)} Q_g^R w_g \leq 0 \quad \forall r \in R \tag{3.77}$$

$$0 \leq u_g \perp \sigma_g - \sum_{i \in I(g)} \varsigma_i - \tau_g + Q_g^L \nu_g - Q_g^L \rho_{n(g)} + K_g \geq 0 \quad \forall g \in G^{SLOW} \tag{3.78}$$

$$0 \leq v_i \perp \varsigma_i + Q_i \nu_{g(i)} - Q_i \rho_{n(i)} + Q_i P_i \geq 0 \quad \forall i \in I \tag{3.79}$$

$$0 \leq w_g \perp \tau_g + Q_g^R \nu_g - Q_g^R \pi_{r(g)} \geq 0 \quad \forall g \in G \tag{3.80}$$

$$0 \leq x_c \perp \sigma_c - Q_c \rho_{n(c)} + Q_c P_c \geq 0 \quad \forall c \in C \tag{3.81}$$

$$f_l \perp -\gamma_l^- + \gamma_l^+ + \phi_l - \rho_{n(l)} + \rho_{m(l)} = 0 \quad \forall l \in L \tag{3.82}$$

$$\theta_n \perp - \sum_{l \in L(n,\cdot)} B_l \phi_l + \sum_{l \in L(\cdot,n)} B_l \phi_l = 0 \quad \forall n \in N \tag{3.83}$$

$$0 \leq y_r \perp \sigma_r - Q_r \pi_r + Q_r P_r \geq 0 \quad \forall r \in R. \tag{3.84}$$

These complementarity conditions guarantee the existence of equilibrium prices for energy and reserve, $\boldsymbol{\rho}^*$ and $\boldsymbol{\pi}^*$, for the relaxed market clearing problem. A direct consequence of this result is that if the solution set of the linear relaxation of (3.49) – (3.63) contains a binary $\boldsymbol{u}$, then $\boldsymbol{\rho}^*$ and $\boldsymbol{\pi}^*$ are also equilibrium prices for the original clearing problem. On the other hand, imposing integrality of $\boldsymbol{u}$ along with (3.64) – (3.84) can result in an infeasible system of equations.

While a complete description of the consequences of complementarity system (3.64) – (3.84) is out of the scope of this chapter, it is important to present these consequences for $\boldsymbol{u}$ in order to understand how to relax the complementarity system so that it remains feasible when integrality constraints are included in the market clearing model. For any slow generator $g \in G^{SLOW}$, if its overall surplus σ_g is positive, then by constraint (3.64) it must be the case that $u_g = 1$ (in-the-money bids must be accepted). On the other hand, if $\sigma_g = 0$ and the right side of (3.78) does not hold with equality, then $u_g = 0$ (out-of-the-money bids must be rejected). Additionally, according to (3.78), child bids in $I(g)$ and reserves can pass surplus to their parent. This allows u_g to be 1 even when $\rho_{n(g)} Q_g^L - K_g < 0$.

If the complementarity restriction of (3.64) is relaxed, one can paradoxically reject any linked bid family submitted by a slow generator. This is in line with the principles of the European market [EGN$^+$16, MV15]. Furthermore, we can guarantee the feasibility of the complementarity system with integrality constrains, independently of the parameters of the bids submitted by market participants. We formalize the latter result in the following proposition.

Proposition 8. *Assume that $F_l \geq 0$ $\forall l \in L$. Then, the system* (3.64) – (3.84)*, with complementarity relaxed for* (3.64) *and binary restrictions on $\boldsymbol{u}$, is feasible.*

Proof of Proposition 8. First, note that problem (3.49) – (3.63) is feasible since, due to our assumption, zero (all variables equal to zero) is a feasible point.

Further, let $(\boldsymbol{v}^*_{u=0}, \boldsymbol{w}^*_{u=0},\ \boldsymbol{x}^*_{u=0}, \boldsymbol{y}^*_{u=0}, \boldsymbol{f}^*_{u=0}, \boldsymbol{\theta}^*_{u=0})$ be an optimal solution to the remaining linear program after fixing $\boldsymbol{u} = \mathbf{0}$. We claim that there exist values $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}, \bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\phi}}, \bar{\boldsymbol{\rho}}, \bar{\boldsymbol{\pi}})$ such that $(\boldsymbol{v}^*_{u=0}, \boldsymbol{w}^*_{u=0}, \boldsymbol{x}^*_{u=0}, \boldsymbol{y}^*_{u=0}, \boldsymbol{f}^*_{u=0}, \boldsymbol{\theta}^*_{u=0}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\varsigma}}, \bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\phi}}, \bar{\boldsymbol{\rho}}, \bar{\boldsymbol{\pi}})$ is feasible for (3.64) – (3.84), with complementarity relaxed for (3.64).

Indeed, we can set $\bar{\nu}_g := 0$ and $\bar{\tau}_g, \bar{\varsigma}_i\ \forall i \in I(g), \bar{\sigma}_g$ to arbitrarily large values for all $g \in G^{SLOW}$, so that constraints (3.78), (3.79) and (3.80) become redundant for all linked bid families. The remaining complementarity system coincides with the KKT conditions of (3.49) – (3.63) after fixing $\boldsymbol{u} = 0$, which by feasibility of (3.49) – (3.63), we know have at least one feasible solution. □

Using the previous observations, we can formulate the nodal day-ahead market clearing model with pricing restrictions as the mathematical program with complementarity constraints (MPCC) (3.85) – (3.87), which is reformulated as an MILP in Prop. 9.

$$\min_{\substack{u,v,w,x,y,f,\theta \\ \sigma,\varsigma,\tau,\nu,\gamma,\phi,\rho,\pi}} \sum_{g \in G^{SLOW}} K_g u_g + \sum_{i \in I} P_i Q_i v_i + \sum_{c \in C} P_c Q_c x_c + \sum_{r \in R} P_r Q_r y_r \quad (3.85)$$

$$\text{s.t.} \quad 0 \leq \sigma_g,\ u_g \in \{0,1\} \quad \forall g \in G^{SLOW} \quad (3.86)$$

$$(3.65) - (3.84) \quad (3.87)$$

Proposition 9. *Problem* (3.85) – (3.87) *and problem* (3.17) – (3.23) *are equivalent in the following sense:*

1. *for each feasible point* $(u, v, w, x, y, f, \theta, \sigma, \varsigma, \tau, \nu, \gamma, \phi, \rho, \pi)$ *of* (3.85) – (3.87)*, there exists* $\tilde{\sigma}$ *such that* $(u, v, w, x, y, f, \theta, \tilde{\sigma}, \varsigma, \tau, \nu, \gamma, \phi, \rho, \pi)$ *is feasible for* (3.17) – (3.23)*.*

2. *Conversely, for each feasible point* $(u, v, w, x, y, f, \theta, \sigma, \varsigma, \tau, \nu, \gamma, \phi, \rho, \pi)$ *of* (3.17) – (3.23)*, there exists* $\tilde{\sigma}$ *such that* $(u, v, w, x, y, f, \theta, \tilde{\sigma}, \varsigma, \tau, \nu, \gamma, \phi, \rho, \pi)$ *is feasible for* (3.85) – (3.87)*.*

Proof of Proposition 9. Analogous to proof of Theorem 2 of [MV15]. □

The procedure applied in this section to obtain an MILP formulation for the nodal day-ahead market clearing problem respecting European pricing restrictions consisted of three main steps: (*i*) writing the KKT conditions, (*ii*) developing a modified version of the dual constraints that admits paradoxical rejection of linked families, and (*iii*) enforcing these dual constraints alongside primal constraints. We apply the same procedure to zonal day-ahead market clearing models, obtaining the formulation (3.31) – (3.38) for flow-based market coupling and the formulation (3.39) – (3.45) for available-transfer-capacity market coupling.

Chapter 4

Renewable energy integration in zonal electricity markets

4.1 Introduction

In this chapter we present a detailed case study of the European market coupling design as implemented at the begining of 2015 in the CWE system, revisiting the question of the relative merits of zonal and nodal markets in the context of systems with high levels of renewable energy.[1]

4.1.1 Literature review

The impact of the simplified representation of transmission constraints in day-ahead energy markets on operational efficiency was recognized in early work by Bjørndal and Jornsten [BJ01] and Ehrenmann and Smeers [ES05], whom illustrate a number of challenges in zonal markets, including efficiency losses and the difficulties of defining zones. The authors use dispatch models in order to obtain analytical insights on systems with up to 6 nodes. Van der Weijde and Hobbs [vdWH11] introduce unit commitment in the study of zonal systems. They focus on quantifying the benefits of zonal coordination in balancing, and use a two-stage model that represents the sequencing of unit commitment and dispatch decisions. Their study is focused on a 4-node system. Recent work by Oggioni *et al.* [OS12, OMS14] further refines zonal models. In [OS12] the

[1]The research presented in this chapter was conducted between 2014 and 2015. It was published in [AP17].

authors use generalized equilibrium models in order to study the effect of coordination among TSOs on operational efficiency. The authors use a standard 6-node network and a 15-node model of the CWE system. They estimate the welfare gains of LMP pricing over zonal pricing in the CWE system at 0.001%. The authors do not account for reserves or unit commitment in their analysis. In [OMS14] the authors evaluate the impacts of priority dispatch of wind in Germany using a zonal model of the CWE system. Uncertainty is accounted for using a scenario-based formulation, however the authors ignore unit commitment decisions in their analysis.

Studies that focus on renewable energy integration in Europe commonly ignore zonal network management either by directly assuming a nodal market [LWv09] or by considering a zonal transportation network without addressing congestion within zones [MOSM10, SW14, KEM14, DDG15]. Nevertheless, a number of studies have estimated the potential efficiency gains of LMP in Europe relative to a zonal design, in the context of renewable energy integration. Leuthold *et al.* [LWv08] estimate the welfare gains of LMP over uniform pricing at 0.8% using a model of Germany and its neighboring countries that consists of 309 nodes. Barth *et al.* [BAV$^+$09] study the effect of international unscheduled flows using a regional model for the entire EU. The authors use a transportation model for transmission and estimate cost savings of 0.1% of nodal relative to zonal pricing. Neuhoff *et al.* [NBB$^+$13] estimate the operating cost savings of LMP relative to zonal pricing between $1.1\% - 3.6\%$. The authors use a single-period unit commitment model of the UCTE-STUM system (4300 nodes, 6000 lines) which is simulated for two extreme operational snapshots (no wind and maximum wind). Abrell and Kunz [AK15] present a framework for day-ahead and intraday operation in a receding horizon scheme, emulating the sequential operation of day-ahead and intraday markets. Abrell and Kunz study a detailed model of the German grid for which they estimate efficiency losses of zonal markets at 0.6%. The authors do not consider uncertainty, assume that all thermal generators can update their commitment in real time and include topology control as a congestion management measure in the zonal market design.

An emerging aspect that results from the large-scale integration of renewable resources is the sub-hourly ramping capacity of a system. The state of the art in renewable energy integration often employs hourly time resolution for day-ahead and real-time operations [TMDO09, MCPR09, PO13, JWG12]. Recent work by Deane *et al.* [DDG14] and Gangammanavar *et al.* [GSZ15] underscores the importance of sub-hourly time resolution in accurately estimating the costs of integrating renewable energy. Bakirtzis *et al.* [BBLB14] propose a receding horizon model with 5-minute resolution for simulating real-time operation under large-scale renewable energy integration. In this study we develop a hybrid model that employs hourly resolution for the commitment of units, and 15-minute resolution for dispatch. This is in line with operating prac-

tice in European markets (hourly day-ahead markets [EPE15] and a quarterly real-time balancing mechanism [HAE+14]).

4.1.2 Contributions and chapter organization

This chapter contributes to the existing literature by presenting a detailed model of the ATC market coupling design and analyzing a detailed instance of the CWE region, which leads to novel insights about the performance of zonal markets in a regime of large-scale renewable energy integration. In terms of modeling, we develop a hierarchy of models for the market coupling design that includes a model for ATC computation that is guaranteed to outperform previously proposed models [NBB+13], a power exchange model that accounts for unit commitment and the treatment of non-convexities by European power exchanges, and a model that emulates the decentralized process of nominations of production and reserves after the day-ahead exchange has cleared. The latter two elements are largely absent from the current literature [BAV+09, vdWH11, OS12, NBB+13, OMS14, AK15]. We use the proposed hierarchy of models to compare the market coupling design to deterministic and stochastic unit commitment models (centralized nodal designs). Numerical results provide novel policy insights by demonstrating that the conjunction of zonal management and unit commitment decisions, in a regime of large-scale renewable energy integration, produces effects that deviate substantially from the assumptions of centrally scheduled systems.

The chapter is organized as follows. Section 4.2 outlines the proposed model for the ATC market coupling design. Section 4.3 presents the simulation setup. Section 4.4 compares the results of the ATC market coupling model to the actual performance of the CWE system over the reference year of the simulation. Section 4.5 compares the performance of the various policies that were investigated and analyzes the obtained results. Finally, section 4.6 concludes the chapter and points to directions of future research.

4.2 A model of European ATC market coupling

From an operational point of view, market coupling consists of sequential steps that are executed or supervised by power exchanges or by system operators, with some differences among countries due to local regulatory frameworks. The steps involved in the day-ahead energy market (namely, the computation of ATCs and the clearing of the energy market) are nearly standardized among countries. In contrast, there is substantial diversity in the definition of reserves. The procedures governing re-dispatch and balancing and the definitions of products are often incompatible among countries. These incompatibilities have already been resolved between Germany and Switzerland, and the TSOs

of Belgium, Germany and the Netherlands have worked towards harmonizing the definition and sharing of their reserve resources [HAE+14]. Further standardization is expected in the medium term following the ENTSO-E *network codes* [ENT14, ENT13]. In anticipation of this harmonization, in this chapter we assume that system operators adhere to a common definition of reserve products among zones.

Following the standardization of reserve products, it is expected that bidding zones will increasingly interchange secondary and tertiary reserves as is currently the case for primary reserves[2]. As opposed to primary reserves, for which shared volumes are in the order of tens of MW, the interchange of secondary and tertiary reserves might involve large volumes of power, which would require the reservation of cross-border transmission capacity between zones, as analyzed by Gebrekiros *et al.* [GDJF15]. The reservation of transmission capacity for reserve provision is currently under debate among European regulators, hence we do not include this element in our analysis in order to focus on the status quo.

We model the transmission network using a lossless DC power flow model. We assume that all dispatch decisions (production, flows) are updated every 15 minutes, whereas commitment (on/off) decisions of thermal generators are updated on an hourly basis. Thermal generators with commitment decisions are divided into two groups: slow generators, whose commitment must be determined in the day-ahead time frame, and fast generators, whose commitment can be modified in real-time operations. Following these assumptions, we model the market coupling design as depicted in Fig. 4.2.

At day ahead, the TSOs compute the ATCs between the different physically connected zones in the system, for each hour τ of the next day. Then, power exchanges collect bids from firms and clear the day-ahead energy market, modeling the exchanges between zones through a transportation network limited by the ATCs, presented in Fig. 4.1 for the CWE, while respecting the European pricing rules for block bids [EGN+16]. The day-ahead energy market is cleared with hourly resolution, and it determines a net position $\Delta Q^{MC}_{a,\tau}$ for each zone a for each hour τ of the next day, as well as a preliminary commitment for slow generators $\boldsymbol{u}^{MC}$.

After the energy market clears and before firms communicate their final schedule to the corresponding TSO, firms within each zone can trade among each other their production and reserve obligations [ELI08], [Res14]. We model this decentralized process as a cost minimizing scheduling that aims at meeting security targets for real-time operation. This procedure results in final commitment decisions for slow generators, $\boldsymbol{u}^{R}$, that comply with the energy balance and reserve requirements of each zone. Note that reserve obligations of firms are determined in monthly or weekly tenders, prior to day-ahead energy mar-

[2]Switzerland, for example, is currently sourcing primary reserve from France and Germany [Swi15].

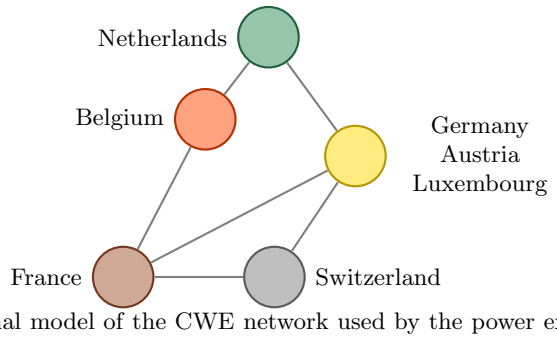


Figure 4.1: Zonal model of the CWE network used by the power exchanges to clear the day-ahead energy market. Each zone is represented as a single node and exchanges between zones are limited by the ATC values.

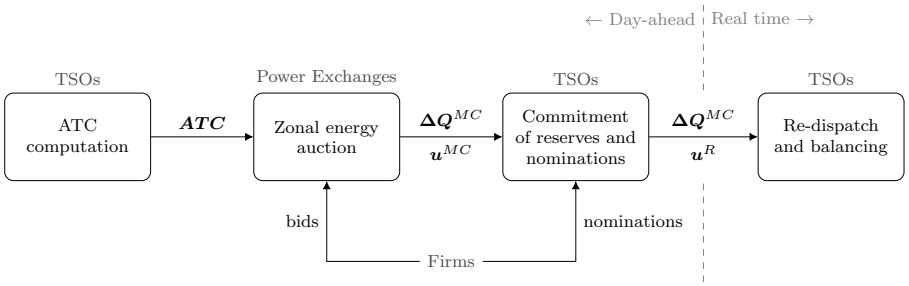


Figure 4.2: Market coupling organization model overview.

ket clearing [HAE+14]. We assume that the forward positions of the reserve tendering process are adjusted by firms after the day-ahead exchange clears, hence we do not represent these forward auctions explicitly in our analysis.

Finally, the resolution of congestion (referred to as re-dispatch) and the resolution of imbalances due to outages and forecast errors (referred to as balancing) take place in real time on 15-minute intervals, while respecting the net position of each zone $\boldsymbol{\Delta Q}^{MC}$ [ENT14] and the commitment of reserves $\boldsymbol{u}^R$. At this stage, we assume that TSOs in charge of the possibly multiple control areas[3] within each bidding zone fully integrate their operations and that they net out their scheduled net positions for balancing purposes [ESS+10]. This assumption allows us to model each individual bidding zone as if it were operated by a single TSO in real time.

The mathematical programs modeling each step of this sequential process

[3]A control area is defined as a coherent part of the interconnected system, operated by a single system operator which includes connected physical loads and/or generation units if any [Eurc]. In the CWE system almost every bidding zone corresponds to a single control area, with the exception of the German-Austrian zone which is divided into five control areas, each one operated by a different system operator.

resemble those presented in section 3.4, with two main differences: (*i*) ATCs are computed separately for each pair of areas, as done by TSOs in practice, and (*ii*) the energy market is cleared using a multi-period model where thermal generators submit mutually exclusive bids for each possible production profile throughout the clearing horizon. We present these models in detail in section 4.A in order to focus our attention on the results and analytics.

4.3 Numerical simulation settings

We simulate the operation of the CWE system, using the instance of section 1.4.3, under three major policy designs: (*i*) the market coupling policy, which is the current zonal design in the CWE system, (*ii*) deterministic unit commitment, which is the current nodal design in several US markets, and (*iii*) stochastic unit commitment, which is an ideal benchmark. Operations under these three designs are modeled in two stages. The first stage takes place in the day ahead and determines the commitment of the slow thermal generators based on a forecast for renewable energy supply. The second stage takes place in real time and corresponds to the re-dispatch and balancing performed by the system operator given the realization of multi-area renewable supply. The second stage must respect the commitment determined for slow thermal generators in the first stage in all policies. Conventional thermal generators are categorized as slow generators if their minimum up and down time is larger than 4 hours, otherwise they are categorized as fast generators.

Thermal generators, other than slow generators, are modeled as follows. The commitment of nuclear generators is decided prior to the day ahead, therefore it is considered as being fixed in the simulations. Similarly, the set point for the production of CHP units is determined based on heat demand and CHP units can adjust their production only within a limited range, therefore we fix their output and allow an adjustment of $\pm 5\%$ of their capacity in the simulations. Fast units adjust both their commitment and production in real time. Aggregated generators correspond to small producers, therefore no commitment decision is associated with them and it is assumed that they can adjust their production in real time.

The market coupling policy is simulated using the model described in section 4.2. Note that while day-ahead markets are cleared using the zonal network of Fig. 4.1, real-time operation is simulated using the nodal network of Fig. 1.1. We consider two variants of the market coupling design, representing different levels of cooperation in real-time congestion management and balancing: (*i*) *MC Net Position* penalizes deviations from day-ahead zonal net positions at the maximum marginal cost of any generator in the system [ES05], representing the operational practice whereby balancing responsible parties are required to balance their resources in real time, which implies that each zone should

maintain its trading position on the day-ahead market. (*ii*) *MC Free* allows for adjustments in zonal net positions at no penalty. This approximates the effect of intra-day markets that allow balancing responsible parties to adjust their positions as real time approaches and the conditions of the system are gradually revealed, as well as the effect of cross-border balancing.

Deterministic and stochastic unit commitment are modeled following [PO13]. Deterministic unit commitment corresponds to a centralized nodal market design with full coordination of the various products of the market (energy, reserves and transmission) and centralized scheduling and operation over all zones. The stochastic unit commitment model additionally endogenizes the uncertainty faced by the system in order to optimally adapt the commitment of reserves to multi-area renewable supply uncertainty. We selected 25 scenarios for the stochastic unit commitment model from the real-time renewable energy production samples, by resorting to the scenario selection algorithm of Heitsch and Römisch [HR07]. For these two models we used a hybrid time resolution with hourly commitment decisions and quarterly dispatch decisions. See section 4.C for a detailed presentation of these two benchmark models.

In order to estimate the performance of each policy, we perform a Monte Carlo simulation over a set of 120 samples of multi-area renewable production. We resort to high performance computing in order to parallelize the Monte Carlo simulations. The relative performance of stochastic unit commitment, deterministic unit commitment and *MC Free* is due to the difference in their day-ahead commitment schedules. The performance of *MC Net Position* is additionally affected by the requirement of adhering, in real time, to day-ahead financial positions.

Mathematical programs are implemented in Mosel/XPress [CH14] and solved on the Sierra cluster at the Lawrence Livermore National Laboratory. Given that we model production cost as a piece-wise linear convex function, all mathematical programs correspond either to LPs or MILPs. Most mathematical programs are solved directly by XPress, with two exceptions. (*i*) Stochastic unit commitment is solved within a 1% optimality gap using an early version of the distributed asynchronous algorithm presented in chapter 2. (*ii*) The zonal energy clearing model is solved using an enumeration heuristic based on column generation. The heuristic achieves a cost which is within 1% of the optimal cost, and optimal welfare which is within 10^{-4}% of optimal.

4.4 Model validation

In order to validate the accuracy of the market coupling model, we compare the results of *MC Net Position* against the historical performance of the CWE system, as reported in publicly accessible statistics. Table 4.1 presents the actual production mix of 2014 [Fed,Res,ELI,Eura] and the resulting production

mix of our model, which is obtained by averaging all samples, seasons and day types (where the contribution of each day type is weighted by the relative frequency of occurrence of each day type).

These results present a reasonable approximation to the actual production mix. The most notable differences appear in the coal production of Germany, the nuclear production of France and the conventional thermal production of the Netherlands. These differences can arise from a number of factors: (*i*) we simulate 960 days of operations over 8 representative day types in our simulation in order to exploit high performance computing and keep the study computationally tractable; (*ii*) the observed estimates of 2014 are subject to statistical error (since they correspond to 365 daily samples of operation, rather than a long-run average); (*iii*) we derate units by season instead of modeling unit-by-unit maintenance and outages, in order to capture their average effect in a season while using representative day types; (*iv*) we compute ATC values endogenously within our model, instead of using the ATCs that were used in the exchange. Notable differences between our model and the ATC values have been observed in the border between Germany and the Netherlands. In order to test the influence of ATC value differences, we have also simulated the operation of the *MC Net Position* model using the actual ATCs [Eurc]. The results better approximate the production mix for the Netherlands and France, however the approximation of the German fuel mix worsens and the average operating cost increases by 8.4%, due to a significant shift in production from France-Germany (where energy is produced at a low marginal cost) to Belgium-Netherlands (where energy is produced at a higher marginal cost). We therefore use the ATC values computed endogenously by our model, in order to maintain ATC values that are internally consistent with the CWE transmission model that we use in our simulations.

In addition to fuel mix, we compare the congestion management costs estimated by our model to those published by national TSOs. We estimate the congestion management costs as the difference between the cost of *MC Net Position* with and without thermal limits on lines. Table 4.2 presents a comparison between the estimated congestion management costs and the actual congestion management costs of the CWE system[4] [Eurc] (values correspond to January-December 2015). The congestion management costs estimated by our model correctly approximate the current situation of the Germany-Austria bidding zone, although they are greater than those observed in reality for Belgium and France. We note that our model does not account for the active control of transmission networks for relieving congestion (e.g. the use of FACTS devices at the borders and transmission switching in Belgium [HP15]). The integration of active transmission network management in our model is a potential area of future research.

[4]Congestion management costs that cover most of the CWE area were not available for 2014.

Table 4.1: Annual energy production by primary source.

Country	Primary source	MC Net Pos. results [TWh]	System statistics [TWh]
Germany	Coal	234.2	274.1
	Nuclear	87.0	97.1
	Gas	11.2	59.8
	Wind	44.4	57.3
	Solar	39.4	35.1
	Hydro	13.1	19.6
	Oil	0.2	6.1
	Other	64.3	75.9
France	Nuclear	468.0	415.5
	Hydro	55.0	67.3
	Wind	16.0	17.1
	Gas	3.0	13.1
	Coal	11.6	6.7
	Solar	5.9	5.8
	Oil	0.0	2.6
	Other	27.2	7.4
Netherlands	Thermal	38.0	91.1
	Wind	6.2	5.8
	Nuclear	4.1	4.1
	Hydro	0.1	0.1
	Solar	0.8	–
Switzerland	Hydro	32.4	37.5
	Nuclear	22.3	25.4
	Thermal	5.2	3.7
	Wind & Solar	0.9	–
Belgium	Nuclear	26.7	32.1
	Thermal	13.2	23.9
	Solar	3.4	2.8
	Wind	4.7	2.5
	Hydro	0.3	1.4
Austria	Hydro	34.5	40.2
	Thermal	13.2	13.8
	Wind	4.0	3.0
	Solar	0.5	–
Total		1291.2	1447.8

Table 4.2: Congestion management costs.

Zone	MC Net Pos. results [MM€/year]	System statistics [MM€/year]
DE/AT/LX	679.2	688.2
Belgium	118.4	0.0
France	66.4	1.1
Netherlands	19.4	–
Switzerland	8.7	–

Table 4.3: Expected policy costs and efficiency losses with respect to deterministic UC.

Policy	Expected cost [MM€/d]	Efficiency losses [%]	[MM€/year]
MC Net Position	30.42	6.2	650
MC Free	29.45	2.8	294
Deterministic UC	28.64	–	–
Stochastic UC	28.49	–0.5	–55
Perfect Foresight	28.32	–1.1	–117

4.5 Policy analysis results

We proceed with a comparison of the cost performance of the four policies described in section 4.3: stochastic unit commitment, deterministic unit commitment, *MC Free* and *MC Net Position*.

The average cost of each policy is presented in Table 4.3. In addition, the cost of perfect foresight is provided for comparison. The difference between *MC Net Position* and *MC Free* quantifies the benefits of intra-day markets and the cooperation among TSOs in balancing. These efficiency gains are estimated at 3.4% of operating costs. The difference between *MC Free* and deterministic unit commitment quantifies the efficiency gains of nodal market design relative to zonal markets, and is estimated at 2.8% of operating cost. Finally, the difference between deterministic and stochastic unit commitment corresponds to the benefits of endogenizing uncertainty relative to using fixed requirements for the commitment of reserves. This gain amounts to 0.5%. Even when considering the perfect foresight model, gains of perfectly forecasting uncertainty are no larger than 1.1%. These gains are notably lower than the aforementioned cost differences between deterministic unit commitment, *MC Free* and *MC Net Position*.

The breakdown of operating costs is presented in Table 4.4, where SLOW^{+} corresponds to the set of nuclear, CHP, slow and aggregated units. We note that the differences between policies are largely driven by the production cost

Table 4.4: Composition of the expected operating cost.

Policy	Commitment cost [MM€/d]		Production cost [MM€/d]		Load shedding
	SLOW$^+$	FAST	SLOW$^+$	FAST	[MM€/d]
MC Net Pos.	2.83	0.28	25.60	1.61	0.10
MC Free	2.83	0.20	25.39	0.97	0.05
Determ. UC	2.81	0.07	25.45	0.31	0.00
Stoch. UC	2.60	0.12	25.21	0.56	0.00

Table 4.5: Production and average marginal cost of thermal generators per policy.

Policy	Production [TWh/year]		Production curtailment	Av. marginal cost [€/MWh]	
	SLOW$^+$	FAST	[TWh/year]	SLOW$^+$	FAST
MC Net Pos.	1014.9	14.6	6.5	9.21	40.14
MC Free	1014.1	11.3	2.4	9.14	31.34
Determ. UC	1016.8	7.4	1.2	9.14	15.29
Stoch. UC	1015.4	9.0	1.3	9.06	22.99

of fast units, with the market coupling policies incurring substantially higher costs from fast units relative to deterministic unit commitment. Differences between deterministic and stochastic unit commitment are driven largely by differences in the commitment cost of slow units.

Table 4.5 presents the expected production of each generator type along with the production-weighted average marginal cost of units providing energy. We note that even though production from fast units is limited to 0.7–1.4% of the total production, the production cost of fast units corresponds to 1.2–6.3% of the total production cost. This table highlights the fact that the market coupling policies often resort to the activation of a substantial amount of fast units that are found at the far right of the fast unit supply stack. This large increase in the production costs of fast units is accompanied by an increasing amount of production curtailment that is required for alleviating congestion and balancing zones in *MC Net Position*.

We proceed with analyzing a number of factors that could explain the substantial observed cost differences among deterministic unit commitment, *MC Free* and *MC Net Position*[5]. In order to better understand the results, we compare the models in pairs, moving from the most efficient to the least efficient policy.

[5]Differences between the stochastic and deterministic UC have been analyzed in the literature [PO13].

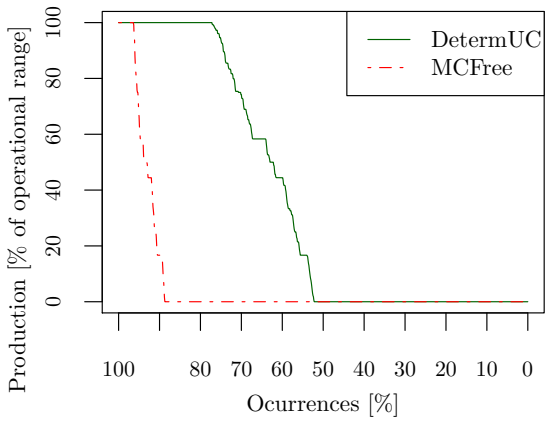


Figure 4.3: Production duration curves of slow units that are exclusively committed by deterministic unit commitment and *MC Free*. For any online unit, 0% of the operational range corresponds to its technical minimum while 100% corresponds to its maximum capacity. In green, slow units that were committed by deterministic unit commitment but not by *MC Free*. In red, slow units committed by the market coupling model but not by deterministic unit commitment.

4.5.1 Relative performance of deterministic unit commitment and MC Free

The MC design and deterministic unit commitment schedule similar amounts of slow capacity within each area, in all day types. The amount of committed capacity is mostly driven by the net demand of each area. Nevertheless, the units committed by *MC Free* are less useful in real time, as can be seen in Fig. 4.3. Units that are committed by *MC Free* and not by deterministic unit commitment remain at their technical minimum for more than 85% of the time, and are used at full capacity for less than 5% of the time. Units committed by deterministic unit commitment and not by *MC Free*, in contrast, are used significantly more.

The commitment decisions of *MC Free* are mainly driven by the merit order of different units within each area, while ignoring intra-zonal flows and misrepresenting the physical laws governing cross-border exchanges. This leads to schedules that are not necessarily feasible when considering the full network, as shown in Fig. 4.4. Slow units that are committed by *MC Free* and result in congestion need to be re-dispatched down in real-time in order to prevent overloading transmission lines, as shown in Fig. 4.5. This results in the activation of more expensive units in real-time relative to deterministic unit commitment.

Fig. 4.6 demonstrates that congestion management after day-ahead market clearing results in a more frequent use of high-cost fast units in the case of *MC*

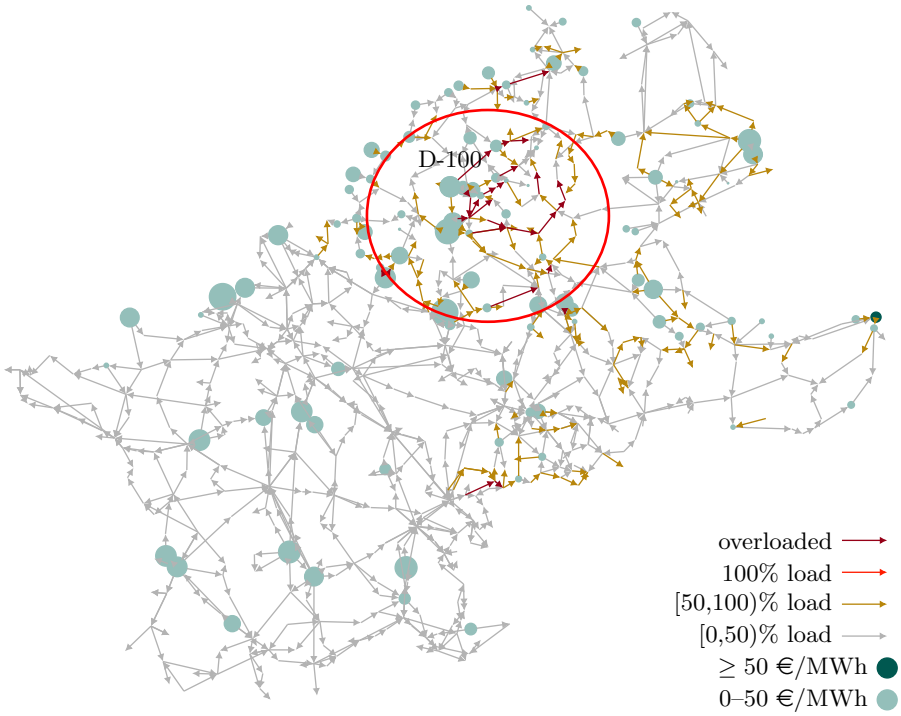


Figure 4.4: Day-ahead schedule determined by *MC Free* for a spring weekday at the 17:00–18:00 interval. Flows implied by the production schedule are infeasible for the real network since they overload lines in the west of Germany.

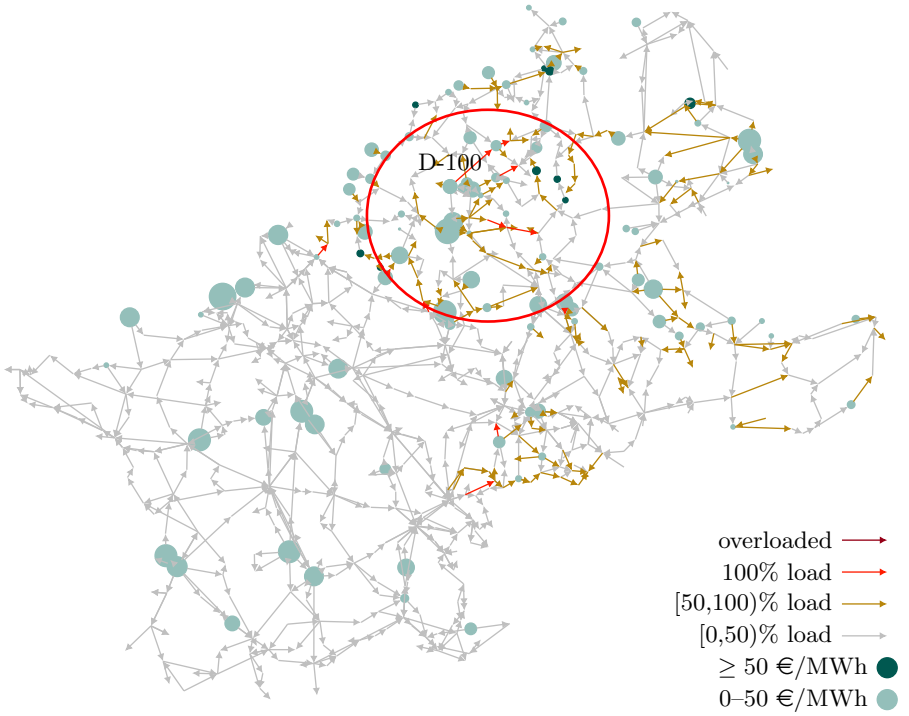


Figure 4.5: Real-time schedule under *MC Free* for a sample of spring weekdays at the 17:30–17:45 interval. The infeasible day-ahead schedule of Fig. 4.4 is altered by re-dispatching all generators at D-100 down to their technical minimum and starting up FAST units in the surrounding area. This re-dispatch pattern is present in all samples of spring weekdays within the 17:00–18:00 period.

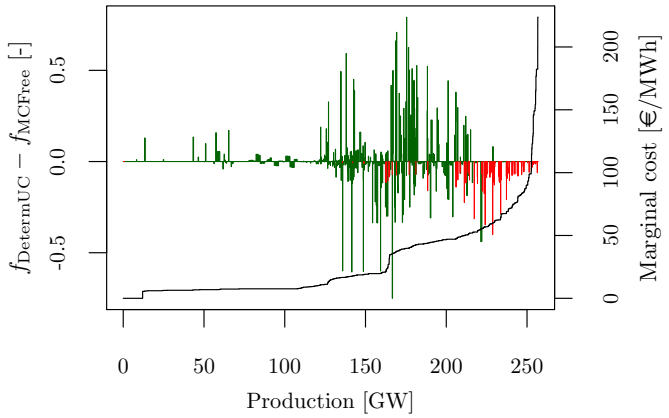


Figure 4.6: Difference in the frequency of use of supply curve increments between deterministic unit commitment and the *MC Free* policy for autumn weekdays. The frequency of use of each supply increment, for a given policy and day type, corresponds to the proportion of the increment that is used on average over all quarterly periods and all samples of real-time operation. In the plot, green stripes represent the frequency of use of supply increments corresponding to SLOW$^+$ generators, while red stripes correspond to fast generators. The marginal cost function of the system (which is measured in the right vertical axis) is presented for reference in black.

Free. Deterministic unit commitment, on the other hand, does not require a substantial modification of the day-ahead schedule in real time since the physical constraints of the transmission network are accounted for when committing slow generators (the same is true for stochastic unit commitment), as it is the case in nodal US markets [Cal15]. This highlights the need to account for unit commitment when analyzing zonal market designs, a feature that has been largely overlooked in existing literature.

4.5.2 Relative performance of MC Free and MC Net Position

The comparison of *MC Net Position* and *MC Free* provides an indication about the value of intra-day adjustments and the cooperation among TSOs for balancing operations [HAE$^+$14]. Fig. 4.7 demonstrates that *MC Free* continuously alters the day-ahead net position through cross-border balancing, whereas *MC Net Position* only deviates from the day-ahead zonal net positions in extreme situations (these correspond to outliers in the box plots).

The behavior of the *MC Net Position* model is largely driven by the renewable supply forecast error of DE/AT/LX (more specifically Germany), since adjustment in other zones is driven by adjustments in DE/AT/LX. This is demonstrated in Fig. 4.8. The figure presents the linear regression of the

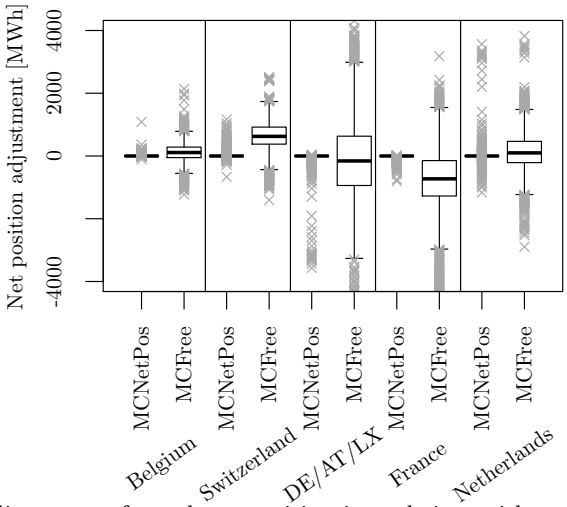


Figure 4.7: Adjustment of zonal net position in real time with respect to the day-ahead net position. A positive adjustment corresponds to a real-time net position that is larger than the day-ahead net position, and vice versa. Net position adjustments of DE/AT/LX range between -6 GW and 5 GW.

adjustments in the DE/AT/LX zone, using the forecast error in renewable production and the day-ahead net positions as independent variables. Across all day types, the forecast error (with positive correlation) and the day-ahead net position of DE/AT/LX (with negative correlation) are the factors with the greatest coefficients and highest significance levels. This indicates that when DE/AT/LX is short on its prediction of renewable supply (negative forecast error) and has available import capacity (large day-ahead net position), the *MC Free* policy will decrease the real-time net position of the zone by importing more power from other areas. Instead, according to the *MC Net Position* policy the forecast error is corrected by re-dispatching within the zone, which results in significantly higher cost, as can be observed in Fig. 4.9. This highlights a major weakness of zonal balancing in systems with substantial levels of renewable supply, which can result in major efficiency losses with respect to cross-border (system-wide) balancing. Note that this type of efficiency loss affects both the continental European system as well as the wide interconnections in US systems [MDE+10].

4.6 Conclusions

In this chapter we present a model for the sequential clearing of products in zonal electricity markets, in particular, the market coupling design currently

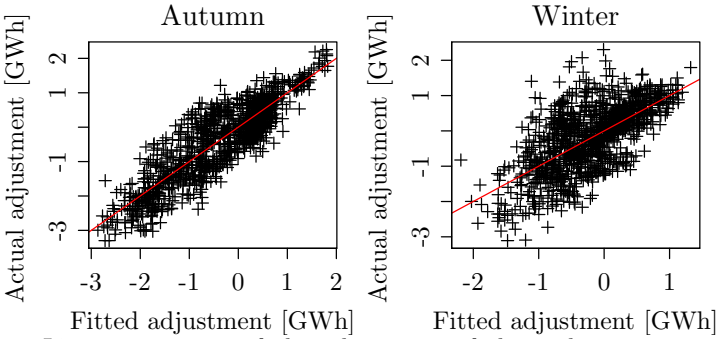


Figure 4.8: Linear regression of the adjustment of the real-time net position of DE/AT/LX under *MC Free* relative to *MC Net Position* for autumn weekdays (best fit) and winter weekdays (worst fit). The explanatory variables are the forecast error of DE/AT/LX and the day-ahead net positions of the zones. A 45 degree line is drawn in red along with the data for comparison.

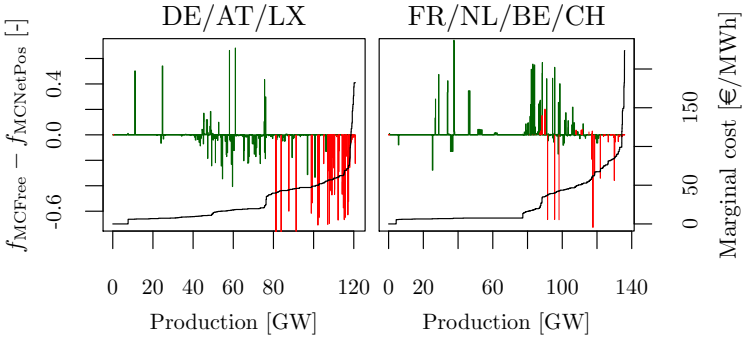


Figure 4.9: Difference in use of zonal supply function increments between *MC Free* and *MC Net Position*, for samples and periods with an adjustment of the net position of DE/AT/LX below the 87.5% quantile (12.5% larger reductions in net position of DE/AT/LX), for autumn weekdays.

as present in Europe in the beginning of 2015. Our model accounts for the simplified representation of transmission in the day-ahead time frame, the separation of energy and reserves, the separation of day-ahead unit commitment decisions from real-time dispatch and balancing, and the uncertainty stemming from renewable forecast errors. The market coupling design is compared to two centralized nodal designs: deterministic and stochastic unit commitment.

Our study finds that market design can exert an influence on physical operations, which far exceeds the benefits of stochastic unit commitment relative to deterministic unit commitment. A decentralized zonal market can undermine system performance in two ways: (*i*) by leading to suboptimal commitment of slow generators and creating significant unscheduled flows in day-ahead markets, and (*ii*) by applying suboptimal balancing strategies due to partial cooperation among multiple system operators in real time. The first type of problem affects only the European market, where institutional barriers have blocked the implementation of LMPs and the splitting of wide zones into smaller ones. In contrast, the second type of problem currently affects both the continental European system and the wide interconnections in US systems. Moreover, the lack of real-time cross-zonal balancing can harm operational security in extreme situations. This has motivated system operators of both systems to study alternatives for better balancing strategies, for instance, imbalance netting and the harmonization of balancing products in Europe.

Future extensions of the present work include (*i*) a receding horizon model of real-time operations that represents the influence of ramp rate constraints more accurately, (*ii*) the study of the impact of ramp rate constraints in a finer time scale (e.g. 5 minutes), (*iii*) the representation of resources (e.g. CCGT, hydro and nuclear) in greater detail, and (*iv*) the representation of active transmission network management.

Appendix

4.A Mathematical programming models for market coupling

4.A.1 Nomenclature

Sets

T_{60}	hourly periods, $T_{60} = \{1, \ldots, 24\}$
T_{15}	15 minute periods, $T_{15} = \{1, \ldots, 96\}$
A	zones
N	buses
L	lines
K	interconnectors
G	thermal generators
$T_{15}(\tau)$	15 minute periods within hour τ
$N(a)$	nodes in zone a
$L(n,m)$	lines between buses n and m, directed from n to m
$L_{\times}(a,b)$	cross-border lines connecting zones a and b, directed from a to b
$K(a,b)$	interconnectors between zones a and b, directed from a to b
Ξ	corners of NTC hyperrectangle
$G(n)$	thermal generators at node n (or set of nodes n)
$\mathcal{N}_{\tau}^{NTC}$	set of feasible exchanges with respect to NTCs
$\mathcal{N}_{\tau}^{OPF}$	set of feasible exchanges with respect to DC OPF
G_{SLOW}	slow generators
G_{FAST}	fast generators
I	continuous bids
J	block bids
$\bar{G}$	exclusive groups
$I(a)$	continuous bids in zone a
$J(a)$	block bids in zone a
$J_E(g)$	block bids within exclusive group g
$\mathcal{D}_g^{60}$	set of feasible generator production decisions for hourly resolution
$\mathcal{D}_g^{15,R}$	set of feasible generator production and reserve decisions for 15-minute resolution
$\mathcal{D}_g^{15}$	set of feasible production decisions for 15-minute resolution

Parameters

$\tau(t)$	corresponding hour of quarter t
$F_l^{\pm}$	flow bounds, line l
B_l	susceptance, line l
$n(l), m(l)$	departing and arrival buses, line l
$D_{n,t}$	demand at bus n on period t
$\bar{\xi}_{n,t}$	forecast renewable supply at bus n, period t
$n_{k,\tau}^{BCE}$	Base Case Exchange through interconnector k in hour τ
$Q_g^{\pm}$	minimum stable level and maximum run capacity
R_a^{FCR}	FCR requirement in zone a (similarly defined for aFRR and mFRR)
ΔT^{FCR}	delivery time of FCR (similarly defined for aFRR and mFRR)
$TTC_{k,\tau}^{\pm}$	total transfer capacity, interconnector k, hour τ
$NTC_{k,\tau}^{\pm}$	net transfer capacity, interconnector k, hour τ

$ATC^{\pm}_{k,\tau}$	available transfer capacity, interconnector k, hour τ
$a(k), b(k)$	departing and arrival zones, interconnector k
Q^i_τ	quantity offered, bid i, hour τ
P^i	unitary price, bid i
M_j	big-M parameter, block bid j
$\hat{a}(i)$	area of bid i
$C_g(q)$	hourly production cost function, generator g
Q^g_j	discrete output quantities, generator g
C^g_j	discrete production costs, generator g
m_g	number of discrete production bins, generator g
K_g	no load cost, generator g
S_g	startup cost, generator g
$h_g(\boldsymbol{\omega}, \boldsymbol{v})$	total cost of production profile $(\boldsymbol{\omega}, \boldsymbol{v})$, generator g
$\Delta Q^{MC}_{a,\tau}$	day-ahead net position of zone a, hour τ
$u^{MC}_{g,\tau}$	day-ahead preliminary commitment, generator g, hour τ
$u^R_{g,\tau}, v^R_{g,\tau}$	day-ahead definitive commitment and startup, generator g, hour τ
V	value of lost load
CL	day-ahead net position update penalty
$\zeta_{n,s,t}$	renewable supply at bus n, Monte Carlo sample s, period t
$TL^{\{60,15\}}_g$	maximum state transition level of generator g for hourly (60') and quarterly (15') formulations
$R^{\{60,15\}}_g$	maximum hourly (60') and quarterly (15') ramp of generator g

Variables

$q_g, q^i_g, q_{g,t}$	quantity produced, generator g
$f_l, f^i_l, f_{l,t}$	flow through line l
$\theta_n, \theta^i_n, \theta_{n,t}$	voltage angle, bus n
x_i	acceptance/rejection continuous bid i
y_j	acceptance/rejection block bid j
$n_{k,\tau}$	exchange through interconnector k, hour τ
$p_{a,\tau}$	energy price in zone a, hour τ
s_i, s_g	surplus of bid i (exclusive group g)
$\lambda^{\pm}_{k,\tau}$	congestion price, interconector k, hour τ
$u_{g,\tau}, v_{g,\tau}$	commitment and startup, generator g, hour τ
$\omega^g_{j,\tau}$	acceptance of production bin j, generator g, hour τ
$r^{FCR}_{g,t}$	FCR provision, generator g, period t (similarly defined for aFRR and mFRR)
$o_{n,t}$	production shedding at bus n, period t
$e_{n,t}$	load shedding at bus n, period t
$\delta_{a,\tau}$	day-ahead net position update, zone a, hour τ

4.A.2 Computation of available transfer capacity

Power exchanges use the simplified transportation network presented in Fig. 4.1 for representing transmission constraints among zones in the CWE system.

The topology of this zonal network is determined by the topology of the real network and it includes an interconnector between each pair of adjacent zones. Flows on the zonal network are limited by the ATCs, which must be computed on a daily basis by the TSOs and communicated to power exchanges.

The first step in the computation of ATCs is the determination of total transfer capacities (TTC) among zones. The ENTSO-E Operational Handbook [Uni04] defines TTC as "*the maximum exchange program between two adjacent control areas that is compatible with operational security standards applied in each system if future network conditions, generation and load patterns are perfectly known in advance*". Following this definition, we propose the following model for computing the TTC from exporting zone a to importing zone b in hour τ, $TTC^{+}_{ab,\tau}$. The notation used in the present and subsequent mathematical formulations is described in Appendix 4.A.1.

$$\max_{q,u,f,\theta} \sum_{l\in L_\times(a,b)} f_l - \sum_{l\in L_\times(b,a)} f_l \tag{4.1}$$

$$\text{s.t.} \sum_{l\in L_\times(c,d)} f_l - \sum_{l\in L_\times(d,c)} f_l = n^{BCE}_{(c,d),\tau} \quad \forall (c,d)\in K\setminus\{(a,b)\} \tag{4.2}$$

$$\sum_{g\in G(n)} q_g + \frac{1}{4}\sum_{t\in T_{15}(\tau)} \bar{\xi}_{n,t} + \sum_{l\in L(\cdot,n)} f_l = \frac{1}{4}\sum_{t\in T_{15}(\tau)} D_{n,t} + \sum_{l\in L(n,\cdot)} f_l \quad \forall n\in N \tag{4.3}$$

$$f_l = B_l\big(\theta_{n(l)} - \theta_{m(l)}\big)\ , \quad -F_l^- \le f_l \le F_l^+ \quad \forall l\in L \tag{4.4}$$

$$Q_g^- u_g \le q_g \le Q_g^+ u_g\ , \quad u_g\in\{0,1\} \quad \forall g\in G \tag{4.5}$$

$$\sum_{l\in L_\times(c,\cdot)} f_l - \sum_{l\in L_\times(\cdot,c)} f_l \le \sum_{g\in G(N(c))} Q_g^+ - \big(R_c^{FCR} + R_c^{aFRR} + R_c^{mFRR}\big) - \frac{1}{4}\sum_{\substack{n\in N(c)\\ t\in T_{15}(\tau)}} \big(D_{n,t} - \bar{\xi}_{n,t}\big) \quad \forall c\in\{a,b\} \tag{4.6}$$

The objective function (4.1) corresponds to the cross-border flow (*exchange program*) from zone a to zone b, determined as the sum of individual flows over cross-border lines. Constraint (4.2) enforces the exchanges between other pairs of areas to correspond to a baseline value, referred to as the base case exchange (BCE). Constraint (4.3) enforces hourly energy balance assuming renewable supply $\bar{\boldsymbol{\xi}}$, constraint (4.4) models the network assuming that all lines are available, constraint (4.5) models generating unit output limits, considering minimum stable and maximum production.

Constraint (4.6) models the *operational security standards* for each control area. The left-hand-side of (4.6) corresponds to the net position of zone c in terms of cross-border flows, while the right-hand-side corresponds to the

maximum net position for zone c such that primary, secondary and tertiary reserve targets, R_c^{FCR}, R_c^{aFRR} and R_c^{mFRR}, respectively, can be met. This constraint limits the exports of each area to a level that ensures that there is enough internal capacity to satisfy the internal demand for energy and reserves.

$TTC^-_{(a,b),\tau}$ (a importing, b exporting) is computed in an analogous way to $TTC^+_{(a,b),\tau}$, by minimizing the objective function (4.1). Note that in order to compute all the TTCs, problem (4.1)–(4.6) needs to be solved twice for each interconnector and hour, i.e. $2 \cdot |K| \cdot |T_{60}|$ times.

Problem (4.1)–(4.6) directly maximizes the cross-border flow between a and b in a one-shot optimization problem, which outperforms iterative methods, such as the one described in [Uni04], and methods based on net position manipulation, proposed by Neuhoff *et al.* [NBB+13].

Once the TTC value is available, the net transfer capacity $NTC^+_{(a,b),\tau}$ is computed by discounting the determined value for $TTC^+_{(a,b),\tau}$ by the Transmission Reliability Margin[6] TRM [Uni04]. Considering a proportional TRM ($0 < TRM < 1$), common to all interconnectors, $NTC^+_{(a,b),\tau}$ is computed using equation (4.7).

$$NTC^+_{k,\tau} := TTC^+_{k,\tau} - TRM \cdot \left|TTC^+_{k,\tau}\right| \cdot 1_{TTC^+_{k,\tau} - TTC^-_{k,\tau} \geq TRM \cdot \left(|TTC^+_{k,\tau}| + |TTC^-_{k,\tau}|\right)} \tag{4.7}$$

The indicator function in equation (4.7) ensures that the TRM is applied only when a sufficient margin between transfer capacities in forward and backward directions exists, i.e. it guarantees that $NTC^+_{(a,b),\tau} \geq NTC^-_{(a,b),\tau}$. For computing $NTC^-_{(a,b),\tau}$, the TRM is added in equation (4.7).

NTCs are required to be simultaneously feasible, in other words, any cross-border exchange configuration respecting the NTC values must be feasible for the real network [Ten14]. In geometric terms, the set of exchange configurations respecting the NTC values for hour τ defines a subset of $\mathbb{R}^{|K|}$, specifically an NTC hyper-rectangle $\mathcal{N}^{NTC}_\tau := \{\boldsymbol{n} \in \mathbb{R}^{|K|} \mid NTC^-_{k,\tau} \leq n_k \leq NTC^+_{k,\tau}\ \forall k \in K\}$. Simultaneous feasibility of NTCs demands that $\mathcal{N}^{NTC}_\tau$ is fully contained within the region defined by feasible exchanges for the real network, $\mathcal{N}^{OPF}_\tau$ (a $|K|$-dimensional polyhedron). In practice, however, this requirement is relaxed and replaced by vertex feasibility[7], which demands that all vertices of $\mathcal{N}^{NTC}_\tau$ are contained within $\mathcal{N}^{OPF}_\tau$ [Ten14].

[6] The transmission reliability margin is defined as a security margin that copes with uncertainties on the computed TTC values arising from: (*i*) inadvertent deviations of physical flows during operation due to the physical functioning of secondary control; (*ii*) emergency exchanges between TSOs to cope with unexpected unbalanced situations in real time; (*iii*) inaccuracies, e.g. in data collection and measurements [Eurc].

[7] Note that vertex feasibility corresponds to a relaxation of simultaneous feasibility if $\mathcal{N}^{OPF}_\tau$ is not convex.

Vertex feasibility is not necessarily satisfied by NTC values computed using equation (4.7), NTC values computed using other procedures proposed in the literature [NBB+13], or the procedure used in practice [Uni04]. The technical documentation [Ten14] states that if NTC values do not comply with vertex feasibility then "*reductions are applied to the NTC levels in a coordinated way with a view to eliminating overloads*". We model this process by computing a new set of NTC values $\boldsymbol{\chi}^+$, $\boldsymbol{\chi}^-$, as the solution of problem (4.8)–(4.14), which aims at achieving vertex feasibility with the minimum possible total reduction on the NTCs.

$$\min_{\chi,q,f,\theta} \sum_{k\in K} \Big(\big(NTC^+_{k,\tau} - \chi^+_k\big) + \big(\chi^-_k - NTC^-_{k,\tau}\big) \Big) \tag{4.8}$$

$$\text{s.t.} \quad \chi^+_k \le NTC^+_{k,\tau}\,, \quad \chi^-_k \ge NTC^-_{k,\tau}\,, \quad \chi^+_k \ge \chi^-_k \quad \forall k \in K \tag{4.9}$$

$$\sum_{l\in L_\times(a,b)} f^i_l - \sum_{l\in L_\times(a,b)} f^i_l = 1^i_{ab}\,\chi^+_{(a,b)} + (1-1^i_{ab})\,\chi^-_{(a,b)} \quad \forall (a,b)\in K, i\in \Xi \tag{4.10}$$

$$\sum_{g\in G(n)} q^i_g + \frac{1}{4}\sum_{t\in T_{15}(\tau)} \bar{\xi}_{n,t} + \sum_{l\in L(\cdot,n)} f^i_l = \frac{1}{4}\sum_{t\in T_{15}(\tau)} D_{n,t} + \sum_{l\in L(n,\cdot)} f^i_l \quad \forall n\in N, i\in\Xi \tag{4.11}$$

$$f^i_l = B_l\big(\theta^i_{n(l)} - \theta^i_{m(l)}\big)\,,\ -F^-_l \le f^i_l \le F^+_l \ \ \forall l\in L, i\in\Xi \tag{4.12}$$

$$0 \le q^i_g \le Q^+_g \quad \forall g\in G, i\in\Xi \tag{4.13}$$

$$\sum_{l\in L_\times(a,\cdot)} f^i_l - \sum_{l\in L_\times(\cdot,a)} f^i_l \le \sum_{g\in G(N(a))} Q^+_g - \big(R^{FCR}_a + R^{aFRR}_a + R^{mFRR}_a\big) - \frac{1}{4}\sum_{\substack{n\in N(a)\\ t\in T_{15}(\tau)}} \big(D_{n,t} - \bar{\xi}_{n,t}\big) \quad \forall a\in A, i\in\Xi \tag{4.14}$$

The objective function (4.8) corresponds to the total NTC reduction, i.e. the difference between the preliminary NTC values $\boldsymbol{NTC}^\pm$ and the vertex feasible NTC values $\boldsymbol{\chi}^\pm$. The constants in this objective function drop out of the optimization and can therefore be ignored, but are included here for clarity of the exposition. Constraints (4.9) establish the bounds for $\boldsymbol{\chi}^\pm$ for all interconnectors.

Each vertex $i \in \Xi$ of $\mathcal{N}^{NTC}_\tau$ can be mapped to an array of directions for cross-border flows, e.g. [*forward* on (a,b), *backward* on (c,d), ...], hence it can be represented using an indicator 1^i_{ab} for each interconnector (a,b) with $1^i_{ab} = 1$ if flow on (a,b) goes in the forward direction in vertex i and $1^i_{ab} = 0$ otherwise. Using these indicators, constraints (4.10) force the cross-border flow

on interconnector (a, b) at vertex $i \in \Xi$ to be equal to the corresponding vertex feasible NTC.

The requirement that all vertices $i \in \Xi$ of $\mathcal{N}_\tau^{NTC}$ are contained within $\mathcal{N}_\tau^{OPF}$ is enforced through constraints (4.11)–(4.14), which correspond to the linear relaxation of constraints (4.3)–(4.6) for each vertex. As the formulation employed is based on the vertices of $\mathcal{N}_\tau^{NTC}$, we have that in general the size of the problem is exponential in the number of interconnectors ($|\Xi| = 2^{|K|}$). For the CWE system the number of vertices to be considered is 2^6, which results in a large-scale linear problem that is still tractable using state-of-the-art linear solvers.

Since our market coupling model does not consider previously contracted transmission capacity, available transfer capacities are equal to the optimal simultaneously feasible NTCs, $ATC_{k,\tau}^{\pm} := (\chi_k^{\pm})^* \ \forall k \in K$.

4.A.3 Day-ahead energy market clearing

Day-ahead energy market clearing in CWE is carried out by power exchanges. Exchanges collect bids from participants and determine the acceptance/rejection decisions that maximize social welfare. Energy is cleared using a strict linear pricing scheme, which results in a series of rules regarding the acceptance/rejection of different types of bids, depending on whether they are in-the-money (bid acceptance would yield a strictly positive profit), at-the-money (bid acceptance would yield zero profit) or out-of-the-money (bid acceptance would yield strictly negative profit). Currently two main types of bids are allowed by power exchanges in the CWE system:

- Continuous bids: bids that can be accepted partially. Continuous bids that are in-the-money must be fully accepted, at-the-money continuous bids can be partially accepted and out-of-the-money continuous bids must be rejected.

- Block bids: bids that can be either fully accepted or rejected (fill-or-kill condition, which gives rise to integer variables in the clearing model). Block bids that are in-the-money or at-the-money can be accepted or paradoxically rejected, while out-of-the-money block bids must be rejected. Block bids can be arranged in linked families, in which the acceptance/rejection of certain bids is conditional on the acceptance of other bids, or in exclusive groups, in which at most one block order within the group can be accepted.

Once the bids have been collected, the power exchange clears the market using the simplified network model provided by the TSOs (Fig. 4.1) in order to represent cross-border exchanges.

Power exchange uses the EUPHEMIA algorithm [EGN+16] to clear the energy market, based on the bids submitted by firms. In the following we present an equivalent model of EUPHEMIA proposed by Madani and Van Vyve [MV15], which has been modified in order to account for exclusive groups.

Consider that buy bids (e.g. loads) correspond to bids with positive quantities $Q > 0$, while sell bids (e.g. generators) correspond to bids with negative quantities $Q < 0$. Then, the welfare maximization (market clearing) problem can be formulated as the mathematical program (4.15)–(4.21), where constraints have been grouped in primal-dual pairs.

$$\max_{\substack{x,y,n\\ s,p,\lambda}} \quad \sum_{i\in I}\Big(\sum_{\tau\in T_{60}} Q^i_\tau P^i\Big)x_i + \sum_{j\in J}\Big(\sum_{\tau\in T_{60}} Q^j_\tau P^j\Big)y_j \tag{4.15}$$

$$\text{s.t.} \quad \sum_{i\in I}\Big(\sum_{\tau\in T_{60}} Q^i_\tau P^i\Big)x_i + \sum_{j\in J}\Big(\sum_{\tau\in T_{60}} Q^j_\tau P^j\Big)y_j \geq \sum_{i\in I} s_i + \sum_{g\in\bar{G}} s_g + \sum_{\substack{k\in K\\ \tau\in T_{60}}}\Big(ATC^+_{k,\tau}\lambda^+_{k,\tau} - ATC^-_{k,\tau}\lambda^-_{k,\tau}\Big) \tag{4.16}$$

$$\sum_{i\in I(a)} Q^i_\tau x_i + \sum_{j\in J(a)} Q^j_\tau y_j = \sum_{k\in K(\cdot,a)} n_{k,\tau} - \sum_{k\in K(a,\cdot)} n_{k,\tau} \quad \forall a\in A, \tau\in T_{60} \tag{4.17}$$

$$ATC^-_{k,\tau} \leq n_{k,\tau} \leq ATC^+_{k,\tau}\,,$$

$$p_{a(k),\tau} - p_{b(k),\tau} + \lambda^+_{k,\tau} - \lambda^-_{k,\tau} = 0 \quad \forall k\in K, \tau\in T_{60} \tag{4.18}$$

$$x_i \leq 1\,, \quad s_i + \sum_{\tau\in T_{60}} Q^i_\tau p_{\hat{a}(i),\tau} \geq \sum_{\tau\in T_{60}} Q^i_\tau P^i \quad \forall i\in I \tag{4.19}$$

$$\sum_{j\in J_E(g)} y_j \leq 1 \quad \forall g\in\bar{G}$$

$$s_{g(j)} + \sum_{\tau\in T_{60}} Q^j_\tau p_{\hat{a}(j),\tau} \geq \sum_{\tau\in T_{60}} Q^j_\tau P^j - M_j(1-y_j) \quad \forall j\in J \tag{4.20}$$

$$\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\lambda} \geq 0\,;\ \boldsymbol{y}\in\{0,1\}^{|J|} \tag{4.21}$$

The objective function (4.15) corresponds to total welfare. Constraint (4.16) enforces strong duality at the solution, i.e. it enforces equality of the total welfare with the total surplus (surplus minimization is the dual of welfare maximization). As a consequence of Theorem 2 of [MV15], constraint (4.16) guarantees that a solution to (4.15)–(4.21) will satisfy (i) complementary slackness between the acceptance of bids and the surplus for continuous bids and accepted block bids, as well as (ii) complementary slackness between exchanges and congestion prices.

Energy balance at each area is expressed through constraints (4.17). Relation (4.18) corresponds to primal and dual constraints for exchanges in the simplified (transportation) network provided by the TSOs.

Constraints (4.19)–(4.20), together with constraint (4.16), establish primal restrictions and dual conditions for the acceptance of different types of bids. Constraints (4.19) and complementary slackness ($s_i \perp 1 - x_i$ and $x_i \perp s_i + \sum_{\tau \in T_{60}} Q^i_\tau p_{\hat{a}(i),\tau} - \sum_{\tau \in T_{60}} Q^i_\tau P^i, \forall i \in I$) ensure that continuous bids are accepted ($x_i = 1$) if they are in-the-money ($s_i > 0$), partially accepted ($0 \leq x_i \leq 1$) if they are at-the-money ($s_i = 0$ and $\sum_{\tau \in T_{60}} Q^i_\tau p_{\hat{a}(i),\tau} = \sum_{\tau \in T_{60}} Q^i_\tau P^i$), and rejected ($x_i = 0$) otherwise.

Constraints (4.20) deal with block bids within exclusive groups. Among the bids $J_E(g)$ of group g, at most one can be accepted and that bid must be in-the-money or at-the-money. Provided M_j are sufficiently large constants, the surplus of the group s_g is determined by the accepted bid only. The accepted bid is not necessarily the one with the maximum surplus within the group, i.e. the maximum surplus bid can be paradoxically rejected, and it is also true that the entire group can be paradoxically rejected.

4.A.4 Firm bids and energy market clearing reformulation

The solution of the market clearing model (4.15)–(4.21) complies with the rules of the CWE exchange that govern the acceptance/rejection of bids. Nevertheless, the model presumes a finite set of bids that have been bid by agents to the exchange.

In order to construct bids for all participants, we assume that they place bids in the energy market that approximate as closely as possible their feasible production/consumption possibilities[8] and their true costs/valuations. Following this assumption, loads, renewable producers and certain thermal producers can be easily modeled as submitting continuous bids.

Other thermal generators, for which we consider commitment decisions, cannot represent their constraints using continuous bids. They are modeled as submitting large exclusive groups (one group per generator) containing a discretized version of their generation possibilities for the next day. To construct this discrete set of production possibilities, the generator output is first discretized in m_g levels $\{Q^g_j\}_{j=1}^{m_g}$. The computation of production cost is accordingly discretized, as shown in Fig. 4.10. The first level of production corresponds to the technical minimum, the last level corresponds to the generator capacity and m_g must be large enough in order to allow the generator to ramp between different levels of output without violating its ramp rate.

The generator output at each hour τ can then be expressed as $\sum_{j=1}^{m_g} Q^g_j \omega^g_{j,\tau}$, where $\omega^g_{j,\tau}$ is an auxiliary binary variable such that $\sum_{j=1}^{m_g} \omega^g_{j,\tau} \leq 1$, $\forall \tau \in T_{60}$ [Glo75]. Any production profile within the discrete set is then defined by

[8] Bidding infeasible production/consumption bids would conflict with the rules of the power exchange [EPE15].

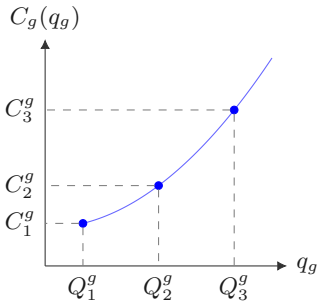


Figure 4.10: Production cost function discretization. Each production level Q_j^g is associated with a production cost C_j^g through the cost function $C_g(\cdot)$.

a certain $\boldsymbol{\omega}^g$. In order to be a feasible production profile, the production, commitment and startup associated with a certain production profile must comply with the generator constraints: technical minimum, maximum capacity, minimum up/down times and ramp rate constraints. These constraints define the hourly production domain $\mathcal{D}_g^{60}$.

Considering that each feasible production profile of generator g is included in group g, we can reformulate the energy clearing model as problem (4.22)–(4.28), where $h_g(\boldsymbol{\omega}_g, \boldsymbol{v}_g)$ corresponds to the total cost of the production profile associated with $\boldsymbol{\omega}_g$ and $\boldsymbol{v}_g$, and where $\boldsymbol{v}_g$ corresponds to the startup indicator variable. Problem (4.22)–(4.28) explicitly includes hourly approximations of the unit commitment constraints for thermal generators while respecting the rules of the power exchange for the treatment of non-convexities.

$$\max_{\substack{x,\omega,v,n\\ s,p,\lambda}} \quad \sum_{i\in I}\left(\sum_{\tau\in T_{60}} Q_\tau^i P^i\right) x_i - \sum_{g\in\bar{G}} h_g(\boldsymbol{\omega}_g, \boldsymbol{v}_g) \tag{4.22}$$

$$\text{s.t.} \quad \sum_{i\in I}\left(\sum_{\tau\in T_{60}} Q_\tau^i P^i\right) x_i - \sum_{g\in\bar{G}} h_g(\boldsymbol{\omega}_g, \boldsymbol{v}_g) \geq \sum_{i\in I} s_i + \sum_{g\in\bar{G}} s_g + \sum_{\substack{k\in K\\ \tau\in T_{60}}} \left(ATC_{k,\tau}^{+}\lambda_{k,\tau}^{+} - ATC_{k,\tau}^{-}\lambda_{k,\tau}^{-}\right) \tag{4.23}$$

$$\sum_{i\in I(a)} Q_\tau^i x_i - \sum_{g\in G(N(a))}\sum_{j=1}^{m_g} Q_j^g \omega_{j,\tau}^g = \sum_{k\in K(\cdot,a)} n_{k,\tau} - \sum_{k\in K(a,\cdot)} n_{k,\tau} \quad \forall a\in A, \tau\in T_{60} \tag{4.24}$$

$$\tag{4.25}$$

$$\sum_{j=1}^{m_g} \omega_{j,\tau}^g \leq 1 \quad \forall g \in \bar{G}, \tau \in T_{60}\,,$$

$$s_g \geq \sum_{\tau \in T_{60}} \sum_{j=1}^{m_g} Q_j^g \omega_{j,\tau}^g p_{\hat{a}(g),\tau} - h_g(\boldsymbol{\omega}_g, \boldsymbol{v}_g) \quad \forall g \in \bar{G} \tag{4.26}$$

$$\left(\sum_{j=1}^{m_g} Q_j^g \boldsymbol{\omega}_j^g\,,\, \sum_{j=1}^{m_g} \boldsymbol{\omega}_j^g\,,\, \boldsymbol{v}_g \right) \in \mathcal{D}_g^{60} \quad \forall g \in \bar{G} \tag{4.27}$$

$$(4.18)\text{–}(4.19);\ \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\lambda} \geq 0\,;\ \boldsymbol{\omega}^g \in \{0,1\}^{m_g \times |T_{60}|}\ \forall g \in \bar{G} \tag{4.28}$$

Problem (4.22)–(4.28) is analogous to (4.15)–(4.21), with the difference that block bids that were explicitly enumerated in (4.20) are now implicitly enumerated by (4.26)–(4.27). Constraint (4.26) ensures that each generator recovers at least its production cost, while constraint (4.27) enforces that the accepted production profile is feasible. We define the total cost function $h_g(\boldsymbol{\omega}_g, \boldsymbol{v}_g)$ according to equation (4.29).

$$h_g(\boldsymbol{\omega}_g, \boldsymbol{v}_g) := \sum_{\tau \in T_{60}} \left(\sum_{j=1}^{m_g} C_j^g \omega_{j,\tau}^g + K_g \sum_{j=1}^{m_g} \omega_{j,\tau}^g + S_g v_{g,\tau} \right) \tag{4.29}$$

Notice that constraint (4.26) includes a product of two variables, $\omega_{j,\tau}^g p_{\hat{a}(j),\tau}$, which can be linearized by using a big-M formulation since $\omega_{j,\tau}^g$ is binary [Glo75].

From a computational perspective, model (4.22)–(4.28) can be challenging for MILP solvers because of its large number of integer variables (proporional to $\sum_{g \in G} m_g$) and the presence of large coefficients. In order to solve (4.22)–(4.28) in cases where MILP solvers fail, we use a column-generation heuristic that operates as follows: (*i*) we solve (4.22)–(4.28) ignoring dual (pricing) constrains, that is, we solve a discretized unit commitment model over the ATC network (this provides an upper bound to (4.22)–(4.28)); (*ii*) then, we solve (4.15)–(4.21) using the production of the unit commitment model as the unique possible production profile for each generator with commitment decisions (the remaining generators submit continuous bids); (*iii*) we construct new feasible production profiles for each generator with commitment decisions based on the clearing prices of problem (4.15)–(4.21) (production profiles that would maximize the generator's profit); and (*iv*) we re-solve (4.15)–(4.21) considering all constructed profiles (this provides an upper bound to (4.22)–(4.28)). The later two steps can be repeated until convergence is attained or all production profiles have been enumerated. Note that this heuristic corresponds to the application of Dantzig-Wolfe algorithm to an integer problem [DW60].

We conclude this subsection by pointing out that the solution of (4.22)–(4.28) represents an optimistic situation in which the power exchange can decide

among a large number of possible schedules in order to maximize welfare. The model determines the preliminary commitment $\boldsymbol{u}_g^{MC}$ for slow generators and net positions $\boldsymbol{\Delta Q}_a^{MC}$ for each zone. The net position can be computed using equations (4.30) and (4.31).

$$u_{g,\tau}^{MC} := \sum_{j=1}^{m_g} (\omega_{j,\tau}^g)^* \quad \forall g \in G_{SLOW}, \tau \in T_{60} \tag{4.30}$$

$$\Delta Q_{a,\tau}^{MC} := \sum_{\substack{g \in \\ G(N(a))}} \sum_{j=1}^{m_g} Q_j^g (\omega_{j,\tau}^g)^* - \sum_{i \in I(a)} Q_\tau^i x_i^* \ \forall a \in A, \tau \in T_{60} \tag{4.31}$$

4.A.5 Treatment of boundary conditions in energy clearing model

An important consequence of replacing (4.20) by (4.26)–(4.27) is that the energy clearing model must choose one production profile for each generator. This rules out the possibility of rejecting all production profiles which, in turn, can render the market clearing problem infeasible when boundary conditions and minimum up time constraints are taken into account.

To understand why this can occur, consider an instance of problem (4.15)–(4.21) for a single zone single period market with several continuous bids and one production block bid, such that the unitary price of the block bid is higher than the unitary price of every demand bid. Existence of a feasible solution for (4.15)–(4.21) can be ensured since we can reject the block bid and find a solution using only the continuous bids. Nevertheless, if an additional constraint enforces the acceptance of the block bid, following market rules (*i*) the market price should be at least equal to the unitary price of the block bid, (*ii*) all continuous demand bids should be rejected and (*iii*) all in-the-money production bids should be accepted. Since at least one production bid has to be accepted and no demand bid can be accepted, we cannot satisfy the balance constraint and pricing restrictions at the same time. In other words, the market clearing problem becomes infeasible due to the forced acceptance of the block bid.

Boundary conditions and minimum up time constraints can effectively enforce the acceptance of non-zero production bids (e.g. generators that have been started up during the last hours of the previous day and must continue online for some hours of the next day), potentially rendering the energy clearing problem (4.22)–(4.28) infeasible. In order to avoid this type of infeasibility we proceed as follows. First, we solve (4.32) for each generator in order to identify the minimal production profile respecting boundary conditions. Then, we include the minimal profiles in the market clearing model priced at the lower

offer cap of the power exchange (-500€/MWh).

$$\begin{aligned} &\min_{\omega,v} \sum_{\tau \in T_{60}} \sum_{j=1}^{m_g} Q_j^g \omega_{j,\tau} \\ &\text{s.t.} \quad \left(\sum_{j=1}^{m_g} Q_j^g \boldsymbol{\omega}_j \, , \sum_{j=1}^{m_g} \boldsymbol{\omega}_j^g \, , \, \boldsymbol{v} \right) \in \mathcal{D}_g^{60} \end{aligned} \tag{4.32}$$

The proposed approach admits the following economic interpretation: if a generator must produce a certain amount due to its own restrictions, then it should offer that amount in the market at the minimum possible price since the cost of producing such amount is already sunk.

4.A.6 Reserves

Reserves in the CWE region can be classified into three categories: (i) primary reserves (also referred as Frequency Containment Reserve or FCR) are responsive to frequency and must be delivered within 30 seconds; (ii) secondary reserves (also referred to as automatic Frequency Restoration Reserves or aFRR) are activated following the activation of FCR, and must be delivered within 5-15 minutes; and (iii) tertiary reserves (also referred to as manual Frequency Restoration Reserves or mFRR) must be delivered within 15-30 minutes.

Following the assumption of harmonization in the definition of reserve products across zones [HAE+14], we model the reserve allocation and nomination process [ELI08, Res14] as a simultaneous cost minimizing scheduling that aims at securing the requisite FCR, aFRR and mFRR capacity. This scheduling is conducted at day ahead in each zone, after the clearing of the energy market, where firms can also trade their production obligations. This is in line with current operating practice in Belgium, France, Germany and Switzerland. Given that balancing responsible parties are required to offer reserve while maintaining a balanced position [ENT14, ENT13], we assume that the reserve scheduling must honor the net positions determined by the power exchange $\Delta Q_{a,\tau}^{MC}$ for each zone and period. Additionally, we assume that slow generators committed by the power exchange cannot be shut down when reserves are allocated.

Considering these assumptions, the commitment of reserve capacity for each 15-minute interval of the following day is formulated as the optimization problem (4.33)–(4.37), which needs to be solved separately for each zone in the system.

$$\min_{q,r,u,v} \sum_{g \in G(N(a))} \left(\frac{1}{4} \sum_{t \in T_{15}} C(q_{g,t}) + \sum_{\tau \in T_{60}} \left(K_g u_{g,\tau} + S_g v_{g,\tau} \right) \right) \tag{4.33}$$

$$\text{s.t.} \sum_{g \in G(N(a))} r^{FCR}_{g,t} \geq R^{FCR}_a ,$$

$$\sum_{g \in G(N(a))} (r^{FCR}_{g,t} + r^{aFRR}_{g,t}) \geq R^{FCR}_a + R^{aFRR}_a ,$$

$$\sum_{g \in G(N(a))} \left(r^{FCR}_{g,t} + r^{aFRR}_{g,t} + r^{mFRR}_{g,t} \right) \geq R^{FCR}_a + R^{aFRR}_a + R^{mFRR}_a \quad \forall t \in T_{15} \tag{4.34}$$

$$\sum_{\substack{g \in G(N(a)) \\ t \in T_{15}(\tau)}} q_{g,t} + \sum_{\substack{n \in N(a) \\ t \in T_{15}(\tau)}} \left(\bar{\xi}_{n,t} - D_{n,t} \right) = 4 \Delta Q^{MC}_{a,\tau} \quad \forall \tau \in T_{60} \tag{4.35}$$

$$u_{g,\tau} \geq u^{MC}_{g,\tau} \quad \forall g \in G_{\text{SLOW}}(N(a)), \tau \in T_{60} \tag{4.36}$$

$$\left(\boldsymbol{q}_g, \boldsymbol{r}^{FCR}_g, \boldsymbol{r}^{aFRR}_g, \boldsymbol{r}^{mFRR}_g, \boldsymbol{u}_g, \boldsymbol{v}_g \right) \in \mathcal{D}^{15,R}_g \quad \forall g \in G(N(a)) \tag{4.37}$$

The objective function (4.33) corresponds to the total cost of the planned operation of zone a. Constraint (4.34) enforces the reserve requirements for each period. Constraint (4.35) requires that each zone maintains its day-ahead position in the energy market, and constraint (4.36) requires that slow units committed by the energy exchange remain ON during the commitment of reserves. Constraint (4.37) corresponds to generator constraints for providing energy and reserves that must be respected with a 15-minute time resolution.

The reserve allocation (4.33)–(4.37) can modify the output of the energy market clearing model by turning on additional slow generators to supply reserves. Consequently, the output of the reserve allocation is a new vector of commitment for slow generators $\boldsymbol{u}^R, \boldsymbol{v}^R$, which guarantees the availability of the required reserves for each zone and for each period of the next day.

4.A.7 Redispatch and balancing

During real-time operations, Kirchhoff's laws determine the flows on the network. Moreover, renewable energy supply can differ from its forecast $\bar{\boldsymbol{\xi}}$. To mitigate the effects of network congestion and forecast errors, the system operator can modify dispatch decisions and the commitment of fast generators with the objective of minimizing the real-time operating cost.

In order to evaluate the real-time cost performance of the day-ahead commitment decisions, the actual renewable injection is assumed to be modeled by the random vector $\boldsymbol{\zeta}$. The real-time operating cost of the system is then estimated by solving the redispatch and balancing model (4.38)–(4.44). The

average performance of the system is estimated through Monte Carlo simulation, i.e. problem (4.38)–(4.44) is solved for each $\boldsymbol{\zeta}_s,\ s \in S_{\text{sim}}$, where S_{sim} is the set of random samples.

$$\min_{\substack{q,u,v,\delta \\ f,\theta,e,o}} \sum_{g\in G}\left(\frac{1}{4}\sum_{t\in T_{15}} C_g(q_{g,t}) + \sum_{\tau\in T_{60}}\left(K_g u_{g,\tau} + S_g v_{g,\tau}\right)\right) + V\sum_{n\in N}\sum_{t\in T_{15}} e_{n,t} + CL\sum_{a\in A}\sum_{\tau\in T_{60}}\delta_{a,\tau} \tag{4.38}$$

$$\text{s.t.}\quad \delta_{a,\tau} \geq \left|4\Delta Q^{MC}_{a,\tau} - \left(\sum_{\substack{l\in L_\times(a,\cdot)\\ t\in T_{15}(\tau)}} f_{l,t} - \sum_{\substack{l\in L_\times(\cdot,a)\\ t\in T_{15}(\tau)}} f_{l,t}\right)\right| \quad \forall a\in A, \tau\in T_{60} \tag{4.39}$$

$$\sum_{g\in G(n)} q_{g,t} + \zeta_{n,s,t} + \sum_{l\in L(\cdot,n)} f_{l,t} + e_{n,t} = D_{n,t} + \sum_{l\in L(n,\cdot)} f_{l,t} + o_{n,t} \quad \forall n\in N, t\in T_{15} \tag{4.40}$$

$$f_{l,t} = B_l\left(\theta_{n(l),t} - \theta_{m(l),t}\right),\ -F_l^- \leq f_{l,t} \leq F_l^+ \quad \forall l\in L, t\in T_{15} \tag{4.41}$$

$$0 \leq o_{n,t} \leq \sum_{g\in G(n)} q_{g,t} + \zeta_{n,s,t},\ 0 \leq e_{n,t} \leq D_{n,t} \quad \forall n\in N, t\in T_{15} \tag{4.42}$$

$$(\boldsymbol{q}_g, \boldsymbol{u}_g, \boldsymbol{v}_g) \in \mathcal{D}_g^{15} \quad \forall g\in G \tag{4.43}$$

$$\boldsymbol{u}_g = \boldsymbol{u}_g^R,\ \boldsymbol{v}_g = \boldsymbol{v}_g^R \quad \forall g\in G_{\text{SLOW}} \tag{4.44}$$

Problem (4.38)–(4.44) commits fast units, with the additional requirement of respecting day-ahead zonal net positions and the commitment schedule of slow generators. The requirement of respecting the zonal day-ahead net positions is imposed as a soft constraint, the violation of which is penalized through an L1 penalty term defined by equation (4.39). Assuming that, in practice, TSOs prefer to redispatch any generator instead of violating their net position, the penalty CL can be established as the maximum marginal cost of any generator within a zone. In contrast, the requirement of respecting the day-ahead commitment of slow generators is imposed as a hard constraint in equation (4.44).

By maintaining day-ahead zonal net positions, the redispatch and balancing model (4.38)–(4.44) captures the partial cooperation of system operators in real time.

4.B Feasible production sets of generators

Production possibilities of thermal generators are limited by technical constraints whose description varies depending on the temporal resolution used

to model the system. For hourly market models the production domain $\mathcal{D}_g^{60}$ of thermal generator g is defined by relation (4.45). $\mathcal{D}_g^{60}$ considers technical minimum, maximum capacity, ramp rate limits (as formulated by Frangioni *et al.* [FGL09]) and minimum up/down times constraints (as formulated by Rajan *et al.* [RT05]).

$$\begin{aligned}\mathcal{D}_g^{60} := \Bigg\{ & \left(\boldsymbol{q}_g, \boldsymbol{u}_g, \boldsymbol{v}_g\right) \in \mathbb{R}_+^{|T_{60}|} \times \{0,1\}^{|T_{60}|} \times \{0,1\}^{|T_{60}|} \;\Bigg|\\ & Q_g^- u_{g,\tau} \le q_{g,\tau} \le Q_g^+ u_{g,\tau}\,,\\ & q_{g,\tau} - q_{g,\tau-1} \le TL_g^{60} - (TL_g^{60} - R_g^{60})\, u_{g,\tau-1}\,,\\ & q_{g,\tau-1} - q_{g,\tau} \le TL_g^{60} - (TL_g^{60} - R_g^{60})\, u_{g,\tau}\,,\\ & \sum_{\substack{\sigma=(\tau-\\ UT_g+1)}}^{\tau} v_{g,\sigma} \le u_{g,\tau}\,, \quad \sum_{\substack{\sigma=(\tau-\\ DT_g+1)}}^{\tau} v_{g,\sigma} \le 1 - u_{g,\tau-DT_g}\,,\\ & v_{g,\tau} \ge u_{g,\tau} - u_{g,\tau-1} \quad \forall \tau \in T_{60} \Bigg\}\end{aligned} \tag{4.45}$$

In quarterly market models we employ an hybrid resolution model for thermal generators where production and reserve provision can vary from quarter t to quarter $t+1$, and commitment status must remain constant within each hour $\tau(t)$. Technical minimum and capacity constraints are formulated as (4.46). Slow generators are allowed to provide reserves only when they are online, while fast generators can provide mFRR even if they are offline.

$$\begin{aligned} g \in G_{SLOW} : \; & Q_g^- u_{g,\tau(t)} \le q_{g,t}\,,\\ & q_{g,t} + r_{g,t}^{FCR} + r_{g,t}^{aFRR} + r_{g,t}^{mFRR} \le Q_g^+ u_{g,\tau(t)} \quad \forall t \in T_{15}\\ g \in G_{FAST} : \; & Q_g^- u_{g,\tau(t)} \le q_{g,t}\,,\\ & q_{g,t} + r_{g,t}^{FCR} + r_{g,t}^{aFRR} \le Q_g^+ u_{g,\tau(t)}\,,\\ & q_{g,t} + r_{g,t}^{FCR} + r_{g,t}^{aFRR} + r_{g,t}^{mFRR} \le Q_g^+ \quad \forall t \in T_{15}\end{aligned} \tag{4.46}$$

We formulate ramp rate constraints with reserves at a quarterly resolution as relations (4.47)–(4.50), $\forall t \in T_{15}$. Constraints (4.47) enforce that $r_{g,t}^{FCR}$ can be provided within ΔT^{FCR} minutes (e.g. 0.5' in the CWE system [HAE+14]). Similarly, constraints (4.48) and (4.49) enforce delivery times for aFRR and mFRR. Note that constraints (4.47)–(4.49) along with the requirement that $r_{g,t}^{FCR}, r_{g,t}^{aFRR}, r_{g,t}^{mFRR}$ are non-negative, enforce also the ramp up rate limit on the production variables. Ramp down rate limits on production variables are enforced through constraints (4.50), $\forall t \in T_{15}$.

$$\left(q_{g,t} + \tfrac{15}{\Delta T^{FCR}}\, r^{FCR}_{g,t}\right) - q_{g,t-1} \leq TL^{15}_g - \left(TL^{15}_g - R^{15}_g\right) u_{g,\tau(t-1)} \tag{4.47}$$

$$\left(q_{g,t} + \tfrac{15}{\Delta T^{aFRR}}\left(r^{FCR}_{g,t} + r^{aFRR}_{g,t}\right)\right) - q_{g,t-1} \leq TL^{15}_g - \left(TL^{15}_g - R^{15}_g\right) u_{g,\tau(t-1)} \tag{4.48}$$

$$\left(q_{g,t} + r^{FCR}_{g,t} + r^{aFRR}_{g,t} + r^{mFRR}_{g,t}\right) - q_{g,t-1} \leq TL^{15}_g - \left(TL^{15}_g - R^{15}_g\right) u_{g,\tau(t-1)} \tag{4.49}$$

$$q_{g,t-1} - q_{g,t} \leq TL^{15}_g - \left(TL^{15}_g - R^{15}_g\right) u_{g,\tau(t)} \tag{4.50}$$

Using the previous hybrid constraints, the feasible production set for generator g at a quarterly resolution with reserves $\mathcal{D}^{15,R}_g$ can then be defined as (4.51). Similarly, the quarterly feasible production set without reserves $\mathcal{D}^{15}_g$ can be defined as (4.52).

$$\begin{aligned}\mathcal{D}^{15,R}_g := \Bigg\{ & \left(\boldsymbol{q}_g, \boldsymbol{r}^{FCR}_g, \boldsymbol{r}^{aFRR}_g, \boldsymbol{r}^{mFRR}_g, \boldsymbol{u}_g, \boldsymbol{v}_g\right) \in \\ & \mathbb{R}_+^{|T_{15}|} \times \mathbb{R}_+^{|T_{15}|} \times \mathbb{R}_+^{|T_{15}|} \times \mathbb{R}_+^{|T_{15}|} \times \{0,1\}^{|T_{60}|} \times \{0,1\}^{|T_{60}|} \Big| \\ & (4.46)\text{–}(4.50), \\ & \sum_{\substack{\sigma=(\tau-\\ UT_g+1)}}^{\tau} v_{g,\sigma} \leq u_{g,\tau}\,, \quad \sum_{\substack{\sigma=(\tau-\\ DT_g+1)}}^{\tau} v_{g,\sigma} \leq 1 - u_{g,\tau-DT_g}\,, \\ & v_{g,\tau} \geq u_{g,\tau} - u_{g,\tau-1} \quad \forall \tau \in T_{60} \Bigg\}\end{aligned} \tag{4.51}$$

$$\mathcal{D}^{15}_g := \left\{ \left(\boldsymbol{q}_g, \boldsymbol{u}_g, \boldsymbol{v}_g\right) \in \mathbb{R}_+^{|T_{15}|} \times \{0,1\}^{|T_{60}|} \times \{0,1\}^{|T_{60}|} \,\middle|\, \left(\boldsymbol{q}_g, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{u}_g, \boldsymbol{v}_g\right) \in \mathcal{D}^{15,R}_g \right\} \tag{4.52}$$

Boundary conditions are parameters of the previous production set definitions (4.45), (4.51), (4.52), expressed using the same notation of production and commitment variables but with $\tau \leq 0$ or $t \leq 0$.

4.C Benchmark models: deterministic and stochastic unit commitment

In contrast to market coupling, deterministic and stochastic unit commitment models clear energy, reserves and transmission simultaneously over the entire

system in the day-ahead market. These models are classic of the literature and are presented in this section for self-containment, without further description. The deterministic unit commitment model is formulated as problem (4.53)–(4.58).

$$\min_{\substack{q,r,u,v\\ f,\theta,e,o}} \sum_{g\in G}\left(\frac{1}{4}\sum_{t\in T_{15}} C_g(q_{g,t}) + \sum_{\tau\in T_{60}}\left(K_g u_{g,\tau} + S_g v_{g,\tau}\right)\right) + V\sum_{n\in N}\sum_{t\in T_{15}} e_{n,t} \quad (4.53)$$

$$\text{s.t.} \quad \sum_{g\in G(n)} q_{g,t} + \bar{\xi}_{n,t} + \sum_{l\in L(\cdot,n)} f_{l,t} + e_{n,t} = D_{n,t} + \sum_{l\in L(n,\cdot)} f_{l,t} + o_{n,t} \quad \forall n\in N, t\in T_{15} \quad (4.54)$$

$$f_{l,t} = B_l\left(\theta_{n(l),t} - \theta_{m(l),t}\right),\ -F_l^- \le f_{l,t} \le F_l^+ \quad \forall l\in L, t\in T_{15} \quad (4.55)$$

$$0 \le o_{n,t} \le \sum_{g\in G(n)} q_{g,t} + \bar{\xi}_{n,t},\ 0 \le e_{n,t} \le D_{n,t} \quad \forall n\in N, t\in T_{15} \quad (4.56)$$

$$\left(\boldsymbol{q}_g, \boldsymbol{r}_g^{FCR}, \boldsymbol{r}_g^{aFRR}, \boldsymbol{r}_g^{mFRR}, \boldsymbol{u}_g, \boldsymbol{v}_g\right) \in \mathcal{D}_g^{15,R} \quad \forall g\in G \quad (4.57)$$

$$\sum_{g\in G(a)} r_{g,t}^{FCR} \ge R_a^{FCR},$$

$$\sum_{g\in G(a)} \left(r_{g,t}^{FCR} + r_{g,t}^{aFRR}\right) \ge R_a^{FCR} + R_a^{aFRR},$$

$$\sum_{g\in G(a)} \left(r_{g,t}^{FCR} + r_{g,t}^{aFRR} + r_{g,t}^{mFRR}\right) \ge R_a^{FCR} + R_a^{aFRR} + R_a^{mFRR} \quad \forall a\in A, t\in T_{15} \quad (4.58)$$

Following Papavasiliou *et. al* [PO13], the stochastic unit commitment model is formulated as the two-stage stochastic program (4.59) – (4.64). The notation is identical to that of the deterministic unit commitment model, with additional indexation by scenario $s \in S$ and state variables $\boldsymbol{w}, \boldsymbol{z}$.

$$\min_{\substack{q,u,w,v,z \\ f,\theta,e,o}} \sum_{s\in S}\pi_s\Bigg(\sum_{g\in G}\Bigg(\frac{1}{4}\sum_{t\in T_{15}} C_g(q_{g,t}) + \sum_{\tau\in T_{60}}\big(K_g u_{g,\tau} + S_g v_{g,\tau}\big)\Bigg) + V\sum_{n\in N}\sum_{t\in T_{15}} e_{n,t}\Bigg) \tag{4.59}$$

$$\text{s.t.} \quad \sum_{g\in G(n)} q_{g,s,t} + \xi_{n,s,t} + \sum_{l\in L(\cdot,n)} f_{l,s,t} + e_{n,s,t} = D_{n,t} + \sum_{l\in L(n,\cdot)} f_{l,s,t} + o_{n,s,t} \quad \forall n\in N, s\in S, t\in T_{15} \tag{4.60}$$

$$f_{l,s,t} = B_l\big(\theta_{n(l),s,t} - \theta_{m(l),s,t}\big)\,,\ -F_l^- \le f_{l,s,t} \le F_l^+ \quad \forall l\in L, s\in S, t\in T_{15} \tag{4.61}$$

$$0 \le o_{n,s,t} \le \sum_{g\in G(n)} q_{g,s,t} + \xi_{n,s,t}\,,\ 0 \le e_{n,s,t} \le D_{n,t} \quad \forall n\in N, t\in T_{15} \tag{4.62}$$

$$\big(\boldsymbol{q}_{g,s}, \boldsymbol{u}_{g,s}, \boldsymbol{v}_{g,s}\big) \in \mathcal{D}_g^{15} \quad \forall g\in G, s\in S \tag{4.63}$$

$$u_{g,s,\tau} = w_{g,\tau}\ ,\ \ v_{g,s,\tau} = z_{g,\tau} \quad \forall g\in G_{\text{SLOW}}, s\in S, \tau\in T_{60} \tag{4.64}$$

Bibliography

[AAC+15] Amprion, APG, Creos, Elia, RTE, TenneT, and Transnet BW. CWE flow based market-coupling project: Parallel run performance report, May 2015.

[ABC+10a] APX Group, Belpex, Cegedel Net, EEX, ELIA Group, EnBw, E-On Netz, Powernext, RTE, RWE, and TenneT. A report for the regulators of the Central West European (CWE) region on the final design of the market coupling solution in the region, by the CWE MC Project, January 2010. URL: http://www.epexspot.com/en/market-coupling/documentation_cwe.

[ABC+10b] APX Group, Belpex, Cegedel Net, EEX, ELIA Group, EnBw, E-On Netz, Powernext, RTE, RWE, and TenneT. A report for the regulators of the Central West European (CWE) region on the final design of the market coupling solution in the region, by the CWE MC Project, January 2010. URL: http://www.epexspot.com/en/market-coupling/documentation_cwe.

[ABP06] APX Group, Belpex, and Powernext. Trilateral Market Ccoupling algorithm, March 2006.

[AC17] ACER and CEER. Annual report on the results of monitoring the internal electricity and gas markets in 2016, October 2017.

[AGK+15] Shabbir Ahmed, Renan Garcia, Nan Kong, Lewis Ntaimo, Feng Qiu Gyana Parija, and Suvrajeet Sen. Siplib: A stochastic integer programming test problem library, 2015. URL: http://www.isye.gatech.edu/~sahmed/siplib.

[Ahm13] Shabbir Ahmed. A scenario decomposition algorithm for 0–1 stochastic programs. *Operations Research Letters*, 41(6):565 – 569, 2013. doi:10.1016/j.orl.2013.07.009.

[Ahm15] Shabbir Ahmed. Corrigendum to "A scenario decomposition algorithm for 0–1 stochastic programs" [Oper. Res. Lett. 41(6) (2013)

565 – 569]. *Operations Research Letters*, 43(2):215 – 217, 2015. doi:10.1016/j.orl.2015.01.006.

[AK15] Jan Abrell and Friedrich Kunz. Integrating intermittent renewable wind generation - a stochastic multi-market electricity model for the european electricity market. *Networks and Spatial Economics*, 15(1):117–147, 2015. doi:10.1007/s11067-014-9272-4.

[AP15] I. Aravena and A. Papavasiliou. A distributed asynchronous algorithm for the two-stage stochastic unit commitment problem. In *2015 IEEE PES General Meeting — Conference Exposition*, pages 1–5, July 2015. (accepted).

[AP17] I. Aravena and A. Papavasiliou. Renewable energy integration in zonal markets. *IEEE Transactions on Power Systems*, 32(2):1334–1349, March 2017. doi:10.1109/TPWRS.2016.2585222.

[APP17] Ignacio Aravena, Anthony Papavasiliou, and Alex Papalexopoulos. A distributed computing architecture for the large-scale integration of renewable energy and distributed resources in smart grids. In Wen-Jyi Hwang, editor, *Recent Progress in Parallel and Distributed Computing*, chapter 3. InTech, Rijeka, 2017. doi:10.5772/67791.

[Aus] Austrian Power Grid AG (APG). Market information. URL: http://www.apg.at/en/market/.

[AW09] Kurt M. Anstreicher and Laurence A. Wolsey. Two "well-known" properties of subgradient optimization. *Mathematical Programming*, 120(1):213–220, 2009.

[BAV+09] Rüdiger Barth, Jürgen Apfelbeck, Philip Vogel, Peter Meibom, and Christoph Weber. Load-flow based market coupling with large-scale wind power in Europe. In *Proceedings of the 8th International Workshop on Large Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Farms*, pages 296–303, Bremen, October 2009.

[BBLB14] E.A. Bakirtzis, P.N. Biskas, D.P. Labridis, and A.G. Bakirtzis. Multiple time resolution unit commitment for short-term operations scheduling under high renewable penetration. *Power Systems, IEEE Transactions on*, 29(1):149–159, Jan 2014. doi:10.1109/TPWRS.2013.2278215.

[BEKS17] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. doi:10.1137/141000671.

[Ben62] J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4(1):238–252, Dec 1962. doi:10.1007/BF01386316.

[BJ01] Mette Bjørndal and Kurt Jørnsten. Zonal pricing in a deregulated electricity market. *The Energy Journal*, 22(1):51–73, 2001.

[BT96] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.

[BTEN09] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princenton Series in Applied Mathematics. Princeton University Press, 2009.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[Cal15] California ISO. Business Practice Manual for Market Operations, November 2015.

[Cal18] California ISO. Fifth Replacement FERC Electric Tariff, July 2018.

[CBFL+09] Santiago Cerisola, Álvaro Baíllo, José M. Fernández-López, Andrés Ramos, and Ralf Gollmer. Stochastic power generation unit commitment in electricity markets: A novel formulation and a comparison of solution methods. *Operations Research*, 57(1):32–46, 2009. doi:10.1287/opre.1080.0593.

[CDR15] Sorathan Chaturapruek, John C. Duchi, and Christopher Ré. Asynchronous stochastic convex optimization: the noise is in the noise and SGD don't care. In *Proceedings of the 2015 Neural Information Processing Systems (NIPS 2015)*, Montréal, Canada, 2015.

[CGCR96] Pierre Carpentier, Guy Gohen, Jean-Christophe Culioli, and Arnaud Renaud. Stochastic optimization of unit commitment: a new decomposition framework. *IEEE Transactions on Power Systems*, 11(2):1067–1073, May 1996. doi:10.1109/59.496196.

[CGSM+15] Kwok Cheung, Dinakar Gade, César Silva-Monroy, Sarah M. Ryan, Jean-Paul Watson, Roger J.-B. Wets, and David L. Woodruff. Toward scalable stochastic unit commitment. Part 2: solver configuration and performance assessment. *Energy Systems*, 6(3):417–438, 2015. doi:10.1007/s12667-015-0148-6.

[CH14] Y. Colombani and S. Heipcke. Multiple models and parallel solving with Mosel, February 2014. Accessed: 2014-11-20. URL: http://community.fico.com/docs/DOC-1141.

[CNH+16] Michael Caramanis, Elli Ntakou, William W. Hogan, Aranta Chakrabortty, and Jens Schoene. Co-optimization of power and reserves in dynamic t amp;d power markets with nondispatchable renewable generation and distributed energy resources. *Proceedings of the IEEE*, 104(4):807–836, April 2016. doi:10.1109/JPROC.2016.2520758.

[Com18] Federal Energy Regulatory Comission. Market Oversight, Electric Power Markets, 2018. Accessed: 2018-04-26. URL: https://www.ferc.gov/market-oversight/mkt-electric/overview.asp.

[CPMS13] CEPS, PSE, MAVIR, and SPS. Unplanned flows in the CEE region: In relation to the common market area Germany - Austria, January 2013.

[CRE17] CREG. Functioning and design of the Central West European day-ahead flow based market coupling for electricity: Impact of TSOs discretionary actions, December 2017.

[CS99] Claus C. Carøe and Rüdiger Schultz. Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24(1–2):37 – 45, 1999. doi:10.1016/S0167-6377(98)00050-9.

[DA15] Ilias Dimoulkas and Mikael Amelin. Probabilistic day-ahead CHP operation scheduling. In *2015 IEEE Power Energy Society General Meeting*, pages 1–5, July 2015. doi:10.1109/PESGM.2015.7285962.

[DCDN14] C.B. Davis, A. Chmieliauskas, G.P.J. Dijkema, and I. Nikolic. Enipedia. Energy & Industry group, Faculty of Technology, Policy and Management, TU Delft, 2014. URL: http://enipedia.tudelft.nl.

[DDG14] J.P. Deane, G. Drayton, and B.P. Ó Gallachóir. The impact of sub-hourly modelling in power systems with significant levels of renewable generation. *Applied Energy*, 113:152 – 158, 2014. doi:10.1016/j.apenergy.2013.07.027.

[DDG15] J.P. Deane, Á. Driscoll, and B.P Ó Gallachóir. Quantifying the impacts of national renewable electricity ambitions using a north–west european electricity market model. *Renewable Energy*, 80:604 – 609, 2015.

[DHL17] Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi:10.1137/15M1020575.

[Die17] C. Dierstein. Impact of generation shift key determination on flow based market coupling. In *2017 14th International Conference on the European Energy Market (EEM)*, pages 1–7, June 2017. doi:10.1109/EEM.2017.7981901.

[DW60] George B. Dantzig and Philip Wolfe. Decomposition principle for linear programs. *Operations Research*, 8(1):101–111, 1960. doi:10.1287/opre.8.1.101.

[EE14] ENTSO-E. Yearly statistics & adequacy retrospect 2013, 2014. URL: http://www.entsoe.eu/publications/statistics/yearly-statistics-and-adequacy-retrospect/.

[EEX18] EEX AG. European Emission Allowances, 2018. Accessed: 2018-02-14. URL: https://www.eex.com/en/market-data/environmental-markets/spot-market/european-emission-allowances.

[EG15] S. Ekisheva and H. Gugel. North American AC circuit outage rates and durations in assessment of transmission system reliability and availability. In *2015 IEEE Power Energy Society General Meeting*, pages 1–5, July 2015. doi:10.1109/PESGM.2015.7285782.

[EGI+14] Jonas Egerer, Clemens Gerbaulet, Richard Ihlenburg, Friedrich Kunz, Benjamin Reinhard, Christian von Hirschhausen, Alexander Weber, and Jens Weibezahn. *Electricity sector data for policy-relevant modeling: data documentation and applications to the German and European electricity markets.* Number 72 in Data documentation / DIW. Berlin : DIW, 2014.

[EGN+16] EPEX Spot, GME, Noord Pool, OMIE, OPCOM, OTE, and TGE. EUPHEMIA Public Description, July 2016.

[ELI] ELIA Group. Elia Grid Data. Accessed: 2014-11-09. URL: http://www.elia.be/fr/grid-data/.

[ELI08] ELIA Group. The CIPU contract: a set framework for taking part in the high-voltage grid management, 2008.

[EN17] Energy Markets Inspectorate (EI, Sweden) and Norwegian Water Resources and Energy Directorate (NVE). Reduced capacity on German-Nordic interconnectors, regulatory framework and socioeconomic effects on the European electricity market, June 2017.

[ENT13] ENTSO-E. Network Code on Load-Frequency Control and Reserves, June 2013. URL: http://networkcodes.entsoe.eu/operational-codes/load-frequency-control-reserves/.

[ENT14] ENTSO-E. Network Code on Electricity Balancing, August 2014. URL: http://networkcodes.entsoe.eu/market-codes/electricity-balancing/.

[EPE15] EPEX Spot. EPEX Spot Market Rules and Regulations, September 2015. Accessed: 2015-11-24. URL: http://www.epexspot.com/en/extras/download-center/documentation.

[Erm83] Yuri Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics*, 9(1-2):1–36, 1983. doi:10.1080/17442508308833246.

[ES05] Andreas Ehrenmann and Yves Smeers. Inefficiencies in European congestion management proposals. *Utilities Policy*, 13(2):135 – 152, 2005. doi:10.1016/j.jup.2004.12.007.

[ESS+10] B. Ernst, C. Scholz, U. Schreier, H.P. Erbring, F. Berster, S. Schlunke, J.H. Pease, and Y.V. Makarov. Large-scale wind and solar integration in germany. Technical Report PNNL-19225, Pacific Northwest National Laboratory, February 2010.

[Eura] European Commission. Eurostat, supply of electicity – monthly data. product code: nrg_105m, accessed: 2015-11-12. URL: http://ec.europa.eu/eurostat/.

[Eurb] European Energy Exchange (EEX). EEX transparency platform. Accessed: 2014-11-09. URL: http://www.transparency.eex.com/.

[Eurc] European Network of Transmission System Operators for Electricity (ENTSO-E). Accessed: 2014-11-09. URL: http://www.entsoe.eu/.

[Eur03] European Parliament, Council of the European Union. Regulation 1228/2003 of the European Parliament and of the Council of 26 June 2003 on conditions for access to the network for cross-border exchanges in electricity, June 2003.

[Eur09] European Parliament, Council of the European Union. Regulation (EC) No 714/2009 of the European Parliament and of the Council

of 13 July 2009 on conditions for access to the network for cross-border exchanges in electricity and repealing Regulation (EC) No 1228/2003, July 2009.

[Eur15a] European Commission. Commission Regulation (EU) 2015/1222 of 24 July 2015 establishing a guideline on capacity allocation and congestion management, July 2015.

[Eur15b] European Network of Transmission System Operators for Electricity (ENTSO-E). ENTSO-E transparency platform, 2015. Accessed: 2018-01-22. URL: https://transparency.entsoe.eu/.

[Eur17a] European Commission. Commission Regulation (EU) 2017/2195 of 23 November 2012 establishing a guideline on electricity balancing, November 2017.

[Eur17b] European Network of Transmission System Operators for Electricity (ENTSO-E). Statistical Factsheet 2016, 2017.

[Eur18] Eurostat. Population on 1 January, 2018. Accessed: 2018-04-26. URL: http://ec.europa.eu/eurostat/tgm/table.do?language=en&pcode=tps00001.

[Fai16] Fair Isaac Corporation (FICO). Xpress-Optimizer Reference Manual, February 2016.

[Fed] Federal Statistical Office, Germany. DESTATIS, gross electricity production. Accessed: 2015-11-12. URL: https://www.destatis.de/EN/FactsFigures/EconomicSectors/Energy/Production/Tables/GrossElectricityProduction.html.

[Fed07] Federal Energy Regulatory Commission (FERC). Initial Decision, Docket No. EL03-180-000, 2007.

[Fed14] Federal Energy Regulatory Commission (FERC). Uplift in RTO and ISO markets, August 2014.

[FGL09] A. Frangioni, C. Gentile, and F. Lacalandra. Tighter approximated milp formulations for unit commitment problems. *IEEE Transactions on Power Systems*, 24(1):105–113, Feb 2009. doi:10.1109/TPWRS.2008.2004744.

[FH14] Frank Fischer and Christoph Helmberg. A parallel bundle framework for asynchronous subspace optimization of nonsmooth convex functions. *SIAM Journal on Optimization*, 24(2):795–822, 2014. doi:10.1137/120865987.

[For15] Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard, Version 3.1.* High Performance Computing Center Stuttgart, 2015.

[FR13] Olivier Fercoq and Peter Richtárik. Smooth minimization of nonsmooth functions with parallel coordinate descent methods. Optimization Online, 2013.

[GDJF15] Yonas Gebrekiros, Gerard Doorman, Stefan Jaehnert, and Hossein Farahmand. Reserve procurement and transmission capacity reservation in the northern european power market. *International Journal of Electrical Power & Energy Systems*, 67:546 – 559, 2015. doi:10.1016/j.ijepes.2014.12.042.

[Geo10] Arthur M. Geoffrion. *Lagrangian Relaxation for Integer Programming*, pages 243–281. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-540-68279-0_9.

[Glo75] Fred Glover. Improved linear integer programming formulations of nonlinear integer problems. *Management Science*, 22(4):455–460, 1975. doi:10.1287/mnsc.22.4.455.

[Gow74] J. C. Gower. Algorithm AS 78: The mediancentre. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(3):466–470, 1974. URL: http://www.jstor.org/stable/2347150.

[GS01] H.I. Gassmann and E. Schweitzer. A comprehensive input format for stochastic linear programs. *Annals of Operations Research*, 104(1):89–125, 2001. doi:10.1023/A:1013138919445.

[GSZ15] H. Gangammanavar, S. Sen, and V.M. Zavala. Stochastic optimization of sub-hourly economic dispatch with wind energy. *Power Systems, IEEE Transactions on*, PP(99):1–11, 2015. doi:10.1109/TPWRS.2015.2410301.

[HAA+17] 50Hertz, Amprion, APG, Creos, Elia, EPEXSpot, RTE, TenneT, and Transnet BW. Documentation of the CWE FB MC solution, September 2017.

[HAE+14] 50Hertz Transmission GmbH, Amprion GmbH, Elia System Operator NV, TenneT TSO B.V., TenneT TSO GmbH, and TransnetBW GmbH. Potential cross-border balancing cooperation between the Belgian, Dutch and German electricity Transmission System Operators, October 2014. Report prepared by the Institute of Power Systems and Power Economics and E-Bridge Consulting GmbH.

[HB13] N. Hutcheon and J.W. Bialek. Updated and validated power flow model of the main continental European transmission network. In *2013 IEEE PowerTech Grenoble*, pages 1–5, June 2013. doi:10.1109/PTC.2013.6652178.

[Hog16] William W. Hogan. Virtual bidding and electricity market design. *The Electricity Journal*, 29(5):33 – 47, 2016. doi:https://doi.org/10.1016/j.tej.2016.05.009.

[HP15] J. Han and A. Papavasiliou. The impacts of transmission topology control on the european electricity network. *IEEE Transactions on Power Systems*, PP(99):1–12, 2015. doi:10.1109/TPWRS.2015.2408439.

[HR07] Holger Heitsch and Werner Römisch. A note on scenario reduction for two-stage stochastic programs. *Operations Research Letters*, 35(6):731–738, November 2007.

[HSL15] HSL. A collection of fortran codes for large scale scientific computation, 2015. URL: http://www.hsl.rl.ac.uk/.

[HZPA15] David Hurlbut, Ella Zhou, Kevin Porter, and Dougla J. Arent. Renewables-friendly grid development strategies: Experience in the United States, potential lessons for China. Technical Report NREL/TP-6A20-64940, National Renewable Energy Laboratory (NREL), October 2015.

[Int18] International Renewable Energy Agency (IRENA). Renewable capacity statistics 2018, 2018.

[JKP17] T. V. Jensen, J. Kazempour, and P. Pinson. Cost-optimal atcs in zonal electricity markets. *IEEE Transactions on Power Systems*, PP(99):1–1, 2017. doi:10.1109/TPWRS.2017.2786940.

[Joi15] Joint Allocation Office (JAO). Yearly long term auctions, 2015. URL: http://www.jao.eu/marketdata/yearlylongtermauctions.

[JWG12] Ruiwei Jiang, Jianhui Wang, and Yongpei Guan. Robust unit commitment with wind power and pumped storage hydro. *Power Systems, IEEE Transactions on*, 27(2):800–810, May 2012. doi:10.1109/TPWRS.2011.2169817.

[KEM14] KEMA Consulting GmbH. Impact of a significant share of renewable energies on the european power generation system. Technical Report 9011-700, European Commission, June 2014.

[KO02] R. Kamat and S. S. Oren. Multi-settlement systems for electricity markets: zonal aggregation under network uncertainty and market power. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 739–748, Jan 2002. doi:10.1109/HICSS.2002.993956.

[KO04] Rajnish Kamat and Shmuel S. Oren. Two-settlement systems for electricity markets under network uncertainty and market power. *Journal of Regulatory Economics*, 25(1):5–37, Jan 2004. doi:10.1023/B:REGE.0000008653.08554.81.

[KZ15] Kibaek Kim and Victor M. Zavala. Algorithmic innovations and software for the dual decomposition method applied to stochastic mixed-integer programs. Optimization Online, 2015.

[LHPA13] Miles Lubin, J. A. Julian Hall, Cosmin G. Petra, and Mihai Anitescu. Parallel distributed-memory simplex for large-scale stochastic lp problems. *Computational Optimization and Applications*, 55(3):571–596, 2013. doi:10.1007/s10589-013-9542-y.

[LMPS13] Miles Lubin, Kipp Martin, Cosmin G. Petra, and Burhaneddin Sandıkçı. On parallelizing dual decomposition in stochastic integer programming. *Operations Research Letters*, 41(3):252 – 258, 2013. doi:10.1016/j.orl.2013.02.003.

[LS04] Guglielmo Lulli and Suvrajeet Sen. A branch-and-price algorithm for multistage stochastic integer programming with application to stochastic batch-sizing problems. *Management Science*, 50(6):786–796, 2004. doi:10.1287/mnsc.1030.0164.

[LWR+15] Ji Liu, Stephen J. Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16(1):285–322, 2015.

[LWv08] Florian Leuthold, Hannes Weigt, and Christian von Hirschhausen. Efficient pricing for european electricity networks – the theory of nodal pricing applied to feeding-in wind in germany. *Utilities Policy*, 16(4):284 – 291, 2008. European Regulatory Perspectives. doi:10.1016/j.jup.2007.12.003.

[LWv09] F. Leuthold, H. Weigt, and C. von Hirschhausen. When the wind blows over europe – a simulation analysis and the impact of grid extensions. *Dresden University of Technology, Chair of Energy Economics, Electricity Market Working Paper*, (WP-EM-31), 2009.

[LWv12] FlorianU. Leuthold, Hannes Weigt, and Christian von Hirschhausen. A large-scale spatial optimization model of the european electricity market. *Networks and Spatial Economics*, 12(1):75–107, 2012. doi:10.1007/s11067-010-9148-1.

[MBH+11] W. Moomaw, P. Burgherr, G. Heath, M. Lenzen, J. Nyboer, and A. Verbruggen. Annex II: Methodology. In O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, and C. von Stechow, editors, *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*. Cambridge University Press, Cambridge, United Kingdom, 2011.

[McC76] Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I — convex underestimating problems. *Mathematical Programming*, 10(1):147–175, 1976. URL: https://doi.org/10.1007/BF01580665, doi:10.1007/BF01580665.

[MCPR09] J.M. Morales, A.J. Conejo, and J. Perez-Ruiz. Economic valuation of reserves in power systems with high penetration of wind power. *Power Systems, IEEE Transactions on*, 24(2):900–910, May 2009. doi:10.1109/TPWRS.2009.2016598.

[MDE+10] Y.V. Makarov, R. Diao, P.V. Etingov, S. Malhara, N. Zhou, R.T. Guttromson, J. Ma, P. Du, N.A. Samaan, and C. Sastry. Analysis methodology for balancing authority cooperation in high penetration of variable generation. Technical Report PNNL-19229, Pacific Northwest National Laboratory, February 2010.

[MLTW13] A. Marien, P. Luickx, A. Tirez, and D. Woitrin. Importance of design parameters on flowbased market coupling implementation. In *2013 10th International Conference on the European Energy Market (EEM)*, pages 1–8, May 2013. doi:10.1109/EEM.2013.6607298.

[Mon10] David Monniaux. Quantifier elimination by lazy model enumeration. In Tayssir Touili, Byron Cook, and Paul Jackson, editors, *Computer Aided Verification*, pages 585–599, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[MOR15] Lluís-Miquel Munguía, Geoffrey Oxberry, and Deepak Rajan. PIPS-SBB: A parallel distributed-memory branch-and-bound algorithm for stochastic mixed-integer programs. Optimization Online, 2015.

[Mor18] Craig Morris. When is the grid full?, January 2018. Accessed: 2018-05-02. URL: https://energytransition.org/2018/01/when-is-the-grid-full/.

[MOSM10] T. Mirbach, S. Ohrem, S. Schild, and A. Moser. Impact of a significant share of renewable energies on the european power generation system. In *Energy Market (EEM), 2010 7th International Conference on the European*, pages 1–6, June 2010. doi:10.1109/EEM.2010.5558688.

[MPS01] Hans W. Moritsch, Georg Ch. Pflug, and Mariusz Siomak. Asynchronous nested optimization algorithms and their parallel implementation. *Wuhan University Journal of Natural Sciences*, 6(1):560–567, 2001. doi:10.1007/BF03160302.

[MV15] Mehdi Madani and Mathieu Van Vyve. Computationally efficient MIP formulation and algorithms for european day-ahead electricity market auctions. *European Journal of Operational Research*, 242(2):580 – 593, 2015.

[NB01] Angelia Nedić and Dimitri P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001. doi:10.1137/S1052623499362111.

[NBB01] Angelia Nedić, Dimitri P. Bertsekas, and Vivek S. Borkar. Distributed asynchronous incremental subgradient methods. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Studies in Computational Mathematics, pages 381–407. Elsevier, 2001.

[NBB+13] Karsten Neuhoff, Julian Barquin, Janusz W. Bialek, Rodney Boyd, Chris J. Dent, Francisco Echavarren, Thilo Grau, Christian von Hirschhausen, Benjamin F. Hobbs, Friedrich Kunz, Christian Nabe, Georgios Papaefthymiou, Christoph Weber, and Hannes Weigt. Renewable electric energy integration: Quantifying the value of design of markets for international transmission capacity. *Energy Economics*, 40(0):760 – 772, 2013. doi:10.1016/j.eneco.2013.09.004.

[Ned02] Angelia Nedić. *Subgradient Methods for Convex Optimization*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, June 2002.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Springer US, 1st edition, 2004.

[Nes05] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. doi:10.1007/s10107-004-0552-5.

[Nes12] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi:10.1137/100802001.

[New16a] New York ISO. Market Administration and Control Area Services Tariff, February 2016.

[New16b] New York ISO. Transmission and Dispatching Operations Manual, February 2016.

[Nor] North American Electric Reliability Corporation (NERC). Standard BAL-001-1 – Real Power Balancing Control Performance. Enforcement date: April 1, 2014.

[Nor12] North American Electric Reliability Corporation (NERC). Interchange Reference Guidelines, March 2012.

[Nor16] North American Electric Reliability Corporation (NERC). Generating availability data system (GADS), 2016. URL: http://www.nerc.com/pa/{RAPA}/gads/Pages/Reports.aspx.

[OMS14] G. Oggioni, F.H. Murphy, and Y. Smeers. Evaluating the impacts of priority dispatch in the European electricity market. *Energy Economics*, 42:183 – 200, 2014. doi:10.1016/j.eneco.2013.12.009.

[OS12] Giorgia Oggioni and Yves Smeers. Degrees of coordination in market coupling and counter-trading. *The Energy Journal*, 33(3):39 – 90, 2012. doi:10.5547/01956574.33.3.3.

[OSS11] Welington Oliveira, Claudia Sagastizábal, and Susana Scheimberg. Inexact bundle methods for two-stage stochastic programming. *SIAM Journal on Optimization*, 21(2):517–544, 2011. doi:10.1137/100808289.

[PO13] Anthony Papavasiliou and Shmuel S. Oren. Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network. *Operations Research*, 61(3):578–592, 2013. doi:10.1287/opre.2013.1174.

[Pol69] Boris T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9:14–29, 1969.

[POR15] Anthony Papavasiliou, Shmuel S. Oren, and Barry Rountree. Applying high performance computing to transmission-constrained stochastic unit commitment for renewable energy integration. *IEEE Transactions on Power Systems*, 30(3):1109–1120, May 2015.

[PSE17] PSE S.A. PSE S.A. informs on the conclusion of the tendering procedure for the supply and implementation of the Electricity Balancing Market Management System in Poland, November 2017. URL: `https://www.pse.pl/web/pse-eng/-/pse-s-a-informs-on-the-conclusion-of-the-tendering-procedure-for-the-supply-and-implementation-of-the-electricity-balancing-market-management-system-i`.

[R C15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL: `https://www.R-project.org/`.

[Res] Reseau de transport d'électricité (RTE). Customer's portal & éCO2mix. URLs: `http://clients.rte-france.com/index_en.jsp`, `http://www.rte-france.com/fr/eco2mix/eco2mix/`. Accessed: 2014-11-28.

[Res14] Reseau de transport d'électricité (RTE). Règles Services Système, July 2014.

[RRA16] K. Ryan, D. Rajan, and S. Ahmed. Scenario decomposition for 0-1 stochastic programs: Improvements and asynchronous implementation. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 722–729, May 2016. `doi:10.1109/IPDPSW.2016.119`.

[RT05] Deepak Rajan and Samer Takriti. Minimum up/down polytopes of the unit commitment problem with start-up costs. IBM Research Report RC23628, Thomas J. Watson Research Center, June 2005.

[RTE06] RTE. Réferentiel Technique, July 2006.

[SB04] Takayuki Shiina and John R. Birge. Stochastic unit commitment problem. *International Transactions in Operational Research*, 11(1):19–32, 2004. `doi:10.1111/j.1475-3995.2004.00437.x`.

[SCTB88] F.C. Schweppe, M.C. Caramanis, R.D. Tabors, and R.E. Bohn. *Spot Pricing of Electricity*. Power Electronics and Power Systems. Springer US, 1988.

[SGM15] Tim Schulze, Andreas Grothey, and Ken McKinnon. A stabilised scenario decomposition algorithm applied to stochastic unit commitment problems. Optimization Online, 2015.

[SJA09] B. Stott, J. Jardim, and O. Alsac. DC power flow revisited. *Power Systems, IEEE Transactions on*, 24(3):1290–1300, Aug 2009. doi:10.1109/TPWRS.2009.2021235.

[SMA14] A. Street, A. Moreira, and J. M. Arroyo. Energy and reserve scheduling under a joint generation and transmission security criterion: An adjustable robust optimization approach. *IEEE Transactions on Power Systems*, 29(1):3–14, Jan 2014. doi:10.1109/TPWRS.2013.2278700.

[Sou15] Southwest Power Pool. Interchange Scheduling Reference Manual, October 2015.

[SW14] Stephan Spiecker and Christoph Weber. The future of the european electricity system and the impact of fluctuating renewable energy – a scenario analysis. *Energy Policy*, 65:185 – 197, 2014.

[Swi15] Swissgrid Ltd. Basic principles of acillary service products, February 2015.

[TBL96] Samer Takriti, John R. Birge, and Erik Long. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems*, 11(3):1497–1508, Aug 1996. doi:10.1109/59.535691.

[Ten] TenneT. Energieinfo. Accessed: 2014-12-10. URL: http://energieinfo.tennet.org/.

[Ten09] TenneT. Quality & capacity plan 2010-2016, 2009. URL: http://www.tennet.eu/nl/about-tennet/news-press-publications/publications/technical-publications.html.

[Ten14] TenneT. Determining securely available cross-border transmission capacity, April 2014.

[TMDO09] A. Tuohy, P. Meibom, E. Denny, and M. O'Malley. Unit commitment for systems with significant wind penetration. *Power Systems, IEEE Transactions on*, 24(2):592–601, May 2009. doi:10.1109/TPWRS.2009.2016470.

[Tse01] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. doi:10.1023/A:1017501703105.

[TvAFL15] Milad Tahanan, Wim van Ackooij, Antonio Frangioni, and Fabrizio Lacalandra. Large-scale unit commitment under uncertainty. *4OR*, 13(2):115–171, 2015. doi:10.1007/s10288-014-0279-y.

[Umb16] UMBRELLA Project Final Report, January 2016. URL: http://www.e-umbrella.eu/documents.

[Uni04] Union for the Coordination of Transmission of Electricity (UCTE). UCTE Operation Handbook, 2004. URL: https://www.entsoe.eu/publications/system-operations-reports/operation-handbook/Pages/default.aspx.

[Uni16] United Nations. Report of the Conference of the Parties on its twenty-first session, held in Paris from 30 November to 13 December 2015. Part two: Action taken by the Conference of the Parties at its twenty-first session, January 2016. URL: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

[vAM16] Wim van Ackooij and Jérôme Malick. Decomposition algorithm for large-scale two-stage unit-commitment. *Annals of Operations Research*, 238(1):587–613, 2016. doi:10.1007/s10479-015-2029-8.

[VBD16] Kenneth Van den Bergh, Jonas Boury, and Erik Delarue. The flow-based market coupling in central western europe: Concepts and definitions. *The Electricity Journal*, 29(1):24 – 29, 2016. URL: http://www.sciencedirect.com/science/article/pii_S104061901530004X, doi:10.1016/j.tej.2015.12.004.

[vdWH11] Adriaan Hendrik van der Weijde and Benjamin F. Hobbs. Locational-based coupling of electricity markets: benefits from coordinating unit commitment and balancing markets. *Journal of Regulatory Economics*, 39(3):223–251, 2011. doi:10.1007/s11149-011-9145-4.

[Voh11] Rakesh V. Vohra. *Mechanism Design: A Linear Programming Approach*. Cambridge University Press, 2011. doi:10.1017/CBO9780511835216.

[WB06] Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006. doi:10.1007/s10107-004-0559-y.

[WRH09] D. Waniek, C. Rehtanz, and E. Handschin. Analysis of market coupling based on a combined network and market model. In *2009 IEEE Bucharest PowerTech*, pages 1–6, June 2009. doi:10.1109/PTC.2009.5282231.

[WRH10] D. Waniek, C. Rehtanz, and E. Handschin. Flow-based evaluation of congestions in the electric power transmission system. In *2010 7th International Conference on the European Energy Market*, pages 1–6, June 2010. doi:10.1109/EEM.2010.5558762.

[Wri15] Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. doi:10.1007/s10107-015-0892-3.

[ZMW+17] Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, and Tie-Yan Liu. Asynchronous stochastic gradient descent with delay compensation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4120–4129, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.