This article was downloaded by: [2a00:1398:4:6c02:dcd6:f884:3020:56e3] On: 09 September 2024, At: 04:16
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Tailored Benders Decomposition for a Long-Term Power Expansion Model with Short-Term Demand Response

Timo Lohmann, Steffen Rebennack

To cite this article:
Timo Lohmann, Steffen Rebennack (2017) Tailored Benders Decomposition for a Long-Term Power Expansion Model with Short-Term Demand Response. Management Science 63(6):2027-2048. https://doi.org/10.1287/mnsc.2015.2420

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2016, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 6, June 2017, pp. 2027–2048
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Tailored Benders Decomposition for a Long-Term Power Expansion Model with Short-Term Demand Response

Timo Lohmann,[a] **Steffen Rebennack**[a]

[a] Division of Economics and Business, Colorado School of Mines, Golden, Colorado 80401

Contact: timo.lohmann@uniper.energy (TL); srebenna@mines.edu (SR)

Received: April 21, 2015
Revised: July 9, 2015; September 15, 2015
Accepted: November 20, 2015
Published Online in Articles in Advance: June 8, 2016

https://doi.org/10.1287/mnsc.2015.2420

Copyright: © 2016 INFORMS

Abstract. We present a long-term power generation expansion planning model that features a long planning horizon, an hourly time resolution, multiperiod investment and retirement decisions, transmission constraints, start-up restrictions, and short-term demand response. Demand response is the capability of power load to react to short-term changes in electricity prices. It plays an increasingly important role in today's electricity markets but has not been taken into consideration in long-term power generation expansion planning problems, which mostly treat demand as perfectly inelastic. Given mild assumptions for the underlying demand function, the resulting model is a large-scale, concave, linearly constrained maximization problem. We exploit the model structure by developing a new approach to generalized Benders decomposition (GBD). In particular, we present two algorithmic ideas: (1) solving the nonlinear Benders subproblem as a linear programming (LP) problem with the aid of dynamic linear overestimation, referred to as the LP-based method, and (2) directly calculating all necessary optimal primal and dual variable values, referred to as the calculation-based method. We consider three special cases of our expansion planning model and show that solving mathematical programming problems can become entirely obsolete in the calculation-based method. We demonstrate the efficiency of all proposed algorithms for the Texas power system, comparing our tailored decomposition methods to a monolithic approach and a state-of-the-art GBD implementation. Our LP-based method is up to 3,822 times faster than the monolithic approach and up to 55 times faster than the GBD. The calculation-based method dramatically improves the solution time, being an average factor of 20 faster than solving LPs and 107,074 times faster than the monolithic approach (for the largest solvable instance by a commercial solver). The overall largest instance we solve, containing more than 79 million variables and constraints, converges in less than one minute using the calculation-based method. The modeling language GAMS and its latest features were used to efficiently implement all algorithms.

History: Accepted by Yinyu Ye, optimization.
Supplemental Material: Data are available available at https://doi.org/10.1287/mnsc.2015.2420.

Keywords: power generation planning • convex nonlinear programming • Benders decomposition • short-term demand response • Benders cut calculation • Texas power system

1. Introduction

Optimization problems in the power industry can be characterized by their time horizon. Power control problems typically span time horizons of seconds to minutes, short-term optimization problems tend to span several days to weeks, midterm optimization problems can span several years, and long-term models consider several decades. Each of these models has a different purpose. While short-term optimization problems typically solve daily operation problems at the power plant level, midterm models might provide necessary information for the short-term models, e.g., water values for hydroelectric plants.

Long-term power generation expansion planning models inform about strategic decisions and were one of the first applications of linear programming (LP) in the 1950s (Massé and Gibrat 1957). They determine a least-cost capacity expansion plan for an existing power system over a planning horizon of 20 years and longer. They are characterized by a forecasted (inelastic) demand or load that has to be met by the power system. Thus, these models are typically used in the context of generation planning in a monopoly. A review of these models is presented by Anderson (1972).

With the changes in the power industry and the electricity market, the focus has shifted from a cost-minimizing to a profit-maximizing perspective (Hobbs 1995). Deregulation and competition play important roles in today's power industry. Often, electricity

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

demand can not be assumed to be a fixed quantity, because demand response programs allow consumers to respond to electricity prices and adapt their electricity consumption accordingly (Albadi and El-Saadany 2008, Cappers et al. 2010). Consequently, price-sensitive demand, also referred to as demand response, needs to be incorporated into these models. This allows the representation of a market in which power producers are in competition with each other to sell power to customers and, more importantly, a system operator who has the ability to communicate demand peaks (i.e., high electricity prices) to consumers in advance. Several authors have investigated the impact of price-sensitive demand in a competitive electricity market (Murphy and Smeers 2005, Borenstein 2005, Bushnell 2010, DeJonghe et al. 2012). Although their models consider some of the aspects of hourly time resolution, none of them combines a long time horizon with an hourly time resolution for a realistically sized power system in one integrated model such as we propose in this paper.

Besides academic models, commercial long-term generation expansion planning models are available and applied in both industry and the political sector. We mention two prominent representatives. The National Energy Modeling System (NEMS) (U.S. Energy Information Administration 2015) by the U.S. Department of Energy is an energy-economics modeling system, built on several modules, that allows the projection of production, consumption, and prices of energy for the United States over a 25-year horizon and can further be used to analyze the impact of energy policies. AURORAxmp (EPIS 2015) by EPIS Inc. is a commercial electric sector model. It decomposes the horizon into blocks of one week. Each block is first solved as a unit commitment model and then as a dispatch model. AURORAxmp allows us to model a wide range of features, including hourly market clearing, a zonal or nodal market design, ramping restrictions, and fixed demand. It can make investment decisions as well as retirement decisions of generators.

We present a long-term power generation expansion planning model, denoted as (PGEP), that considers short-term demand response; i.e., as the electricity price increases, consumers reduce their demand. In economic theory, this is equivalent to a downward sloping demand function to represent the bidding behavior of electricity consumers. We model the electricity market with an hourly time resolution, which results in a large-scale problem that contains a market clearance condition for each hour. The demand function leads to a (convex) nonlinear programming (NLP) problem and, in the special case of a linear demand function, to a quadratic programming (QP) problem (DeJonghe et al. 2011b). Our expansion model is motivated by a model of Fell and Linn (2013). We enhance their model by using start-up restrictions for coal and nuclear plants, transmission constraints, multiple investment periods, investments in transmission lines, and the option to decommission existing power plants. Further, they use a genetic algorithm to solve their expansion planning problem, which they terminate after 24 hours of CPU time. In contrast, we propose a novel and tailored Benders decomposition-type approach to solve for the optimal investment decision that maximizes the welfare of the entire power system with respect to general demand functions beyond a linear function. In this regard, our work can also be seen as an extension of the work of DeJonghe et al. (2011b), who discuss several ways to solve the long-term generation expansion planning problem that includes short-term demand response.

The result is a stylized electricity market model that features short-term demand response in a deregulated market environment and can be applied to long time horizons with detailed time resolutions. Given perfect competition, the bid-based dispatch equals the central dispatch. As such, central dispatching is a good proxy when studying deregulated market environments, if no player can execute market power (Gross and Finlay 2000). Effects of market power on the electricity market are typically quantified by equilibrium models, e.g., Cournot and Bertrand models, and Stackelberg models (Bushnell 2003). Although it is not the focus of this work, note that the purpose of our model is not to perfectly forecast electricity prices but to describe the impact of electricity policies and regulations on the behavior of the system.

The application of Benders decomposition to the power generation expansion planning problem in various forms has been widely studied in the literature. Depending on the structure of the subproblem, either classical Benders decomposition (Kim et al. 2011, Baringo and Conejo 2011) or generalized Benders decomposition can be applied (Bloom 1982, 1983; Bloom et al. 1984). In our case, we decompose (PGEP) into a master problem that contains the investment and decommission decisions and a nonlinear subproblem that contains the market clearing condition in each hour. However, instead of applying general Benders decomposition, we propose to dynamically linearize the subproblem to obtain a more efficient solution algorithm. Given certain assumptions for the demand function, our subproblem is a convex NLP with linear constraints. Linearization of demand surplus in the electricity market context has been studied before (García et al. 1999), but not in combination with Benders decomposition.

In particular, with these mild assumptions for the demand function, we can show that our algorithm solves the expansion planning problem to proven optimality for any given tolerance. We show how the

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

special structure of our subproblem can be further exploited and even allows us to compute the dual variables needed for the Benders optimality cuts. Our core model without start-up and transmission constraints can be decomposed and solved up to five orders of magnitude faster than a monolithic model with a state-of-the-art commercial solver. Given that only small instances of the model can be solved as a monolith because of time and memory restrictions, we expect even greater factors for the large instances this algorithm solves. If start-up and transmission constraints are present, the dual variables can be calculated for certain hours, improving the convergence time of the Benders decomposition algorithm as well.

The unique contributions of this paper are the following:

- We present a long-term generation expansion planning model that features transmission constraints, start-up restrictions, and demand response with an hourly time resolution. Investment decisions, as well as binary decommissioning decisions, are made on a yearly basis.
- We propose several Benders decomposition approaches, including nested Benders decomposition (NBD), which dynamically linearize the nonlinear subproblem. Finite convergence and correctness of the proposed decomposition algorithms are preserved via a dynamic overestimation approach given certain assumptions for the demand response function hold. The resulting algorithms are up to 55 times faster than a classical generalized Benders decomposition (GBD).
- We show how the structure of (PGEP)'s special cases can be exploited by presenting methods to explicitly calculate the necessary dual solution in the subproblem to construct the Benders optimality cuts. Solving linearized subproblems is no longer required, which leads to very efficient algorithms. The resulting algorithms are 20 times faster on average than the overestimation approach above and up to 876 times faster than a GBD. Compared to the largest solvable instance with a monolithic approach, this method is 107,074 times faster.

2. Long-Term Expansion Planning Model

We develop a long-term power generation expansion planning model that can be used to evaluate energy market regulations and policies. We want to obtain insights on their impact both at an operational level and at a strategic level, i.e., long-term investment and retirement decisions. Therefore, a compromise in modeling detail and computational tractability must be achieved. The following list describes the key desired features:

- Planning horizon of more than 20 years to capture long-term trends and structural changes.
- Multiperiod investment and decommissioning decisions of private independent power producers to account for change in regulations and policies over time.
- Representation of a competitive electricity wholesale market in which buyers of electricity face sellers. Sellers are represented by supply bids, and a downward sloping demand curve allows the incorporation of short-term demand response.
- Detailed time resolution to capture daily demand and supply patterns as well as interaction of renewable generation sources and base-load generators such as coal.
- Start-up restriction for coal generators to model their inflexibility with regard to hourly shifts in load.
- Representation of transmission lines to model zonal market prices and congestion.

The model presented in the following subsections describes our approach to include the above features. We discuss possible extensions at the end of this section.

2.1. The Full Model

We begin by stating the notation:

Indices and Sets:

$i \in \mathbb{I}$:	all generators [-]
$i \in \mathbb{I}^{\mathrm{EX}}$:	existing generators [-]
$i \in \mathbb{I}^{\mathrm{N}}$:	new generators [-]
$i \in \mathbb{I}^{\mathrm{R}}$:	generators with start-up restrictions [-]
$i \in \mathbb{I}_u$:	set of generators i belonging to bus u [-]
$h \in \mathbb{H}$:	hours [-]
$(h', t') \in \mathscr{A}_{ht}$:	previous hour–year combination (h', t') of hour h in year t [-,-]
$t \in \mathbb{T}$:	years [-]
$t \in \mathbb{T}^{\mathrm{I}}$:	years with investment [-]
$u \in \mathbb{U}$:	buses (zones; nodes) [-]
$uv \in \Omega$:	transmission line from bus u to bus v [-]

Parameters:

β:	discount factor [%]
C_{ih}:	capacity factor of generator i in hour h [\$/MWh]
c_{it}:	(marginal) generator cost of generator i in year t [\$/MWh]
$\hat{c}_{it}$:	start-up cost of generator i in year t [\$/MW]
F^{I}_{it}:	investment cost of generator i in year t [\$/MW]
F^{OM}_{i}:	operation & maintenance cost of generator i [\$/MW]
$F^{\max}_{uv}$:	maximum transmission capacity of transmission line uv [MWh]
F^{T}_{uvt}:	investment cost of transmission line uv in year t [\$/MW]
K_i:	capacity of generator i [MW]
K^{T}_{uv}:	maximum capacity expansion of transmission line uv [MW]
$P^{\min}_{i}$:	minimum generation of generator i, if running [MWh]
$P^{\max}_{i}$:	maximum generation of generator i [MWh]

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Decision Variables:

δ_{iht}: (continuous) start-up of generator i in hour h of year t [%]

γ^{D}_{it}: (binary)

$$\begin{cases} 0 & \text{if existing generator } i \text{ is decommissioned in year } t, \\ 1 & \text{if existing generator } i \text{ is online in year } t \end{cases}$$

γ^{N}_{it}: (binary)

$$\begin{cases} 1 & \text{if new generator } i \text{ is built in year } t, \\ 0 & \text{otherwise} \end{cases}$$

$\gamma^{\mathrm{T}}_{uvt}$: (continuous) investment decision transmission line uv in year t [%]

Q_{uht}: (continuous) electricity load at bus u in hour h of year t [MWh]

q_{iht}: (continuous) dispatch of generator i in hour h of year t [MWh]

q^{L}_{iht}: (continuous) per unit generation up to $P^{\min}_i$ of generator i in hour h of year t [%]

q^{U}_{iht}: (continuous) per unit generation between $P^{\min}_i$ and $P^{\max}_i$ of generator i in hour h of year t [%]

P_{uht}: (continuous) inverse demand function of bus u in hour h and year t [$/MWh]

x_{uvht}: (continuous) transmission of line uv in hour h of year t [MWh]

The long-term power generation expansion planning model (PGEP) reads as follows.

- Objective function (see Section 2.2)

$$\begin{aligned} W^*_{\mathrm{NLP}} := \Bigg\{ \max & -\sum_{t\in\mathbb{T}^{\mathrm{I}}} \beta^{(t-1)} \left[\sum_{i\in\mathbb{I}^{\mathrm{N}}} \gamma^{\mathrm{N}}_{it} F^{\mathrm{I}}_{it} K_i + \sum_{uv\in\Omega} \gamma^{\mathrm{T}}_{uvt} F^{\mathrm{T}}_{uvt} K^{\mathrm{T}}_{uv} \right] \\ & + \sum_{t\in\mathbb{T}} (\beta)^t \Bigg[-\sum_{i\in\mathbb{I}^{\mathrm{N}}} \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \gamma^{\mathrm{N}}_{it'} F^{\mathrm{OM}}_i K_i - \sum_{i\in\mathbb{I}^{\mathrm{EX}}} \gamma^{\mathrm{D}}_{it} F^{\mathrm{OM}}_i K_i \\ & \quad + \sum_{h\in\mathbb{H}} \Bigg(\sum_{u\in\mathbb{U}} \int_0^{Q_{uht}} P_{uht}(x)\,dx \\ & \quad - \sum_{i\in\mathbb{I}} c_{it} q_{iht} - \sum_{i\in\mathbb{I}^{\mathrm{R}}} \hat{c}_{it} K_i \delta_{iht} \Bigg) \Bigg] \Bigg\}. \end{aligned} \tag{1}$$

- Generator capacity bounds (see Section 2.3)

$$\text{s.t. } 0 \le q_{iht} \le \gamma^{\mathrm{D}}_{it} C_{ih} K_i \quad \forall\, i\in\mathbb{I}^{\mathrm{EX}},\ h\in\mathbb{H},\ t\in\mathbb{T}, \tag{2}$$

$$0 \le q_{iht} \le \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \gamma^{\mathrm{N}}_{it'} C_{ih} K_i \quad \forall\, i\in\mathbb{I}^{\mathrm{N}},\ h\in\mathbb{H},\ t\in\mathbb{T}. \tag{3}$$

- Start-up restrictions (see Section 2.4)

$$\text{s.t. } P^{\min}_i q^{\mathrm{L}}_{iht} + (P^{\max}_i - P^{\min}_i) q^{\mathrm{U}}_{iht} = q_{iht} \quad \forall\, i\in\mathbb{I}^{R},\ h\in\mathbb{H},\ t\in\mathbb{T}, \tag{4}$$

$$q^{\mathrm{L}}_{iht} - q^{\mathrm{U}}_{iht} \ge 0 \quad \forall\, i\in\mathbb{I}^{R},\ h\in\mathbb{H},\ t\in\mathbb{T}, \tag{5}$$

$$q^{\mathrm{L}}_{iht} - q^{\mathrm{L}}_{ih't'} \le \delta_{iht} \quad \forall\, i\in\mathbb{I}^{R},\ h\in\mathbb{H},\ t\in\mathbb{T},\ (h',t')\in\mathcal{A}_{ht}, \tag{6}$$

$$\delta_{iht} \ge 0 \quad \forall\, i\in\mathbb{I}^{R},\ h\in\mathbb{H},\ t\in\mathbb{T}, \tag{7}$$

$$0 \le q^{\mathrm{L}}_{iht},\ q^{\mathrm{U}}_{iht} \le 1 \quad \forall\, i\in\mathbb{I}^{R},\ h\in\mathbb{H},\ t\in\mathbb{T}. \tag{8}$$

- Transmission constraints (see Section 2.5)

$$\text{s.t. } \sum_{vu\in\Omega} x_{vuht} - \sum_{uv\in\Omega} x_{uvht} + \sum_{i\in\mathbb{I}_u} q_{iht} = Q_{uht} \quad \forall\, u\in\mathbb{U},\ h\in\mathbb{H},\ t\in\mathbb{T}, \tag{9}$$

$$0 \le x_{uvht} \le F^{\max}_{uv} + \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \gamma^{\mathrm{T}}_{uvt'} K^{\mathrm{T}}_{uv} \quad \forall\, uv\in\Omega,\ h\in\mathbb{H},\ t\in\mathbb{T}, \tag{10}$$

$$Q_{uht} \ge 0 \quad \forall\, u\in\mathbb{U},\ h\in\mathbb{H},\ t\in\mathbb{T}. \tag{11}$$

- Investment and decommissioning (see Section 2.6)

$$\text{s.t. } \gamma^{\mathrm{D}}_{it} \le \gamma^{\mathrm{D}}_{i,t-1} \quad \forall\, i\in\mathbb{I}^{\mathrm{EX}},\ t\in\mathbb{T},\ t\ge 2, \tag{12}$$

$$\sum_{t\in\mathbb{T}^{\mathrm{I}}} \gamma^{\mathrm{N}}_{it} \le 1 \quad \forall\, i\in\mathbb{I}^{\mathrm{N}}, \tag{13}$$

$$\sum_{t\in\mathbb{T}^{\mathrm{I}}} \gamma^{\mathrm{T}}_{uvt} \le 1 \quad \forall\, u\in\mathbb{U},\ v\in\Omega_u, \tag{14}$$

$$\gamma^{\mathrm{N}}_{it} \in \{0,1\} \quad \forall\, i\in\mathbb{I}^{\mathrm{N}},\ t\in\mathbb{T}^{\mathrm{I}}, \tag{15}$$

$$\gamma^{\mathrm{D}}_{it} \in \{0,1\} \quad \forall\, i\in\mathbb{I}^{\mathrm{EX}},\ t\in\mathbb{T}, \tag{16}$$

$$0 \le \gamma^{\mathrm{T}}_{uvt} \le 1 \quad \forall\, uv\in\Omega,\ t\in\mathbb{T}^{\mathrm{I}}. \tag{17}$$

2.2. Objective Function

The objective function (1) maximizes the welfare of the system and can be divided into two parts: the discounted hourly market clearing term

$$\sum_{t\in\mathbb{T}} (\beta)^t \left[\sum_{h\in\mathbb{H}} \left(\sum_{u\in\mathbb{U}} \int_0^{Q_{uht}} P_{uht}(x)\,\mathrm{d}x - \sum_{i\in\mathbb{I}} c_{it} q_{iht} - \sum_{i\in\mathbb{I}^{\mathrm{R}}} \hat{c}_{it} K_i \delta_{iht} \right) \right],$$

which is described in Section 2.3, and the annual cost terms. These are operation and maintenance costs for new and existing generators

$$-\sum_{t\in\mathbb{T}} (\beta)^t \left[\sum_{i\in\mathbb{I}^{\mathrm{N}}} \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \gamma^{\mathrm{N}}_{it'} F^{\mathrm{OM}}_i K_i + \sum_{i\in\mathbb{I}^{\mathrm{EX}}} \gamma^{\mathrm{D}}_{it} F^{\mathrm{OM}}_i K_i \right],$$

as well as capital costs for investment in new generators and transmission lines

$$-\sum_{t\in\mathbb{T}^{\mathrm{I}}} (\beta)^{(t-1)} \left[\sum_{i\in\mathbb{I}^{\mathrm{N}}} \gamma^{\mathrm{N}}_{it} F^{\mathrm{I}}_i K_i + \sum_{uv\in\Omega} \gamma^{\mathrm{T}}_{uvt} F^{\mathrm{T}}_{uv} K^{\mathrm{T}}_{uv} \right].$$

Section 2.6 describes this component in detail. All terms are discounted by a factor of β. This factor represents the model's combined assumptions on, among others, inflation, equity structure, and taxes. For instance, for the initial investment decision in year zero, it is used to calculate that investment's net present value of monetary flows in future years.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

2.3. Market Clearing Condition

The market clearing mechanism of the electricity market is the core component of (PGEP). For notational convenience and without loss of generality, we drop the time indices h and t and bus index u in this section. For each hour and zone, demand function f is defined as

$$Q = f(P). \tag{18}$$

We assume f to be a continuously differentiable, strictly monotonically decreasing function on $[0, +\infty)$; i.e., for two prices $P_1 > P_2$, it follows that $f(P_1) = Q_1 < Q_2 = f(P_2)$. This implies that the first derivative $f'(x)$ is nonzero for all $x \in [0, +\infty)$. Further, suppose $f(x) = b$. Then, its inverse function f^{-1} exists at b and is continuous and strictly monotonically decreasing. The existence of the inverse function f^{-1} enables us to rewrite (18) equivalently in terms of quantity Q as the inverse demand function

$$P = f^{-1}(Q). \tag{19}$$

Function $f^{-1}(Q)$ is more commonly referred to as $P(Q)$. The market clearance condition is depicted in Figures 1–3.

The dashed lines show price P^* and quantity Q^* at equilibrium for a given hour. The gray blocks symbolize generators or, equivalently, supply bids. The width of a block describes the respective generator's (effective) capacity; the height of a block describes the respective generator's marginal cost. The generators are sorted in ascending order of their marginal cost (from left to right); examples of generators with zero marginal cost are wind or solar units.

The equilibrium price is determined by the running generator with the highest marginal cost (Figure 1) or by the total generation in the system in case demand cannot be met, either because there is not enough capacity at the offered price (Figure 2) or not enough capacity in general (Figure 3). We avoid the case depicted in Figure 3 by introducing an artificial generator with a sufficiently high marginal cost, resulting in the case shown in Figure 2.

We are interested in computing the welfare, i.e., consumer surplus plus producer surplus, at equilibrium. The welfare is given by the area between the inverse demand function and the gray blocks from 0 to Q^*, as illustrated in Figures 1–3. We calculate the welfare using the inverse demand function $P(x)$, as defined in (19); the electricity quantity at equilibrium Q^*; and

Figure 1. Most Expensive Generator Defines Market Clearing Price

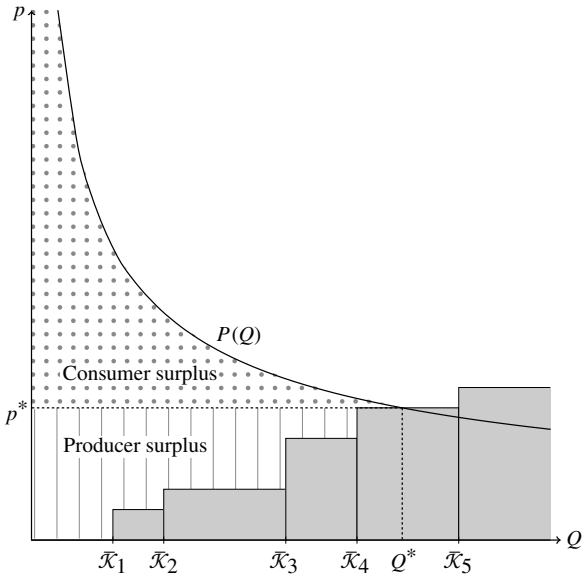


Figure 2. Available Capacity at Offered Price Defines Market Clearing Price

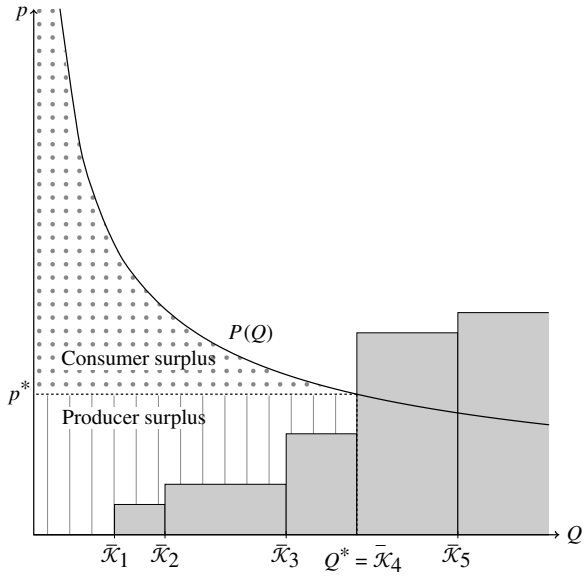
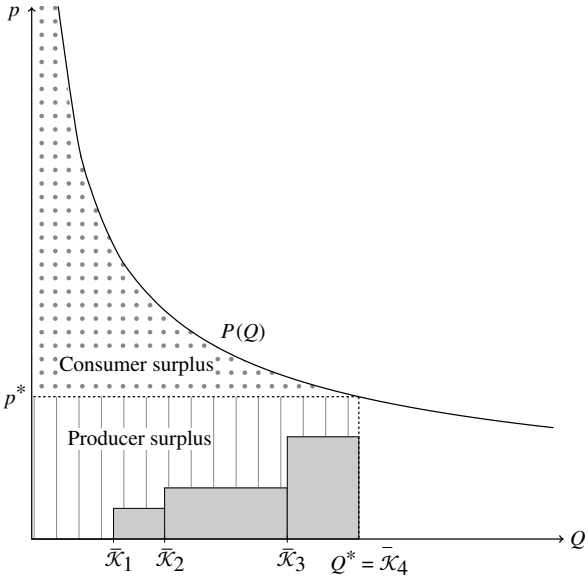


Figure 3. Available Capacity Defines Market Clearing Price



Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

the generators' marginal cost c_i through

$$W^*(Q^*) := \int_0^{Q^*} P(x)\,dx - \sum_{i\in\mathbb{I}} c_i q_i^*, \tag{20}$$

with the condition

$$\sum_{i\in\mathbb{I}} q_i^* = Q^*.$$

Next, for function $W^*(Q)$, we establish the conditions for which the function (i) exists, i.e., assumes a finite value, and (ii) is concave. We require the concave shape for our tailored Benders decomposition-type approaches (see Section 3). Below, we argue that the required assumptions are naturally met by the application at hand.

For condition (i), we need the (lower) limit of the integral

$$\lim_{Q_1\to 0} \int_{Q_1}^{Q_2} P(x)\,dx$$

to exist for any finite Q_2. We are not concerned about the existence of the limit for $Q_2 \to +\infty$ because we can always ensure a finite value for $Q_2 = Q^*$ by adding an artificial generator with a significantly high enough marginal cost and effective capacity, as done for the case depicted in Figure 3.

For condition (ii), consider the following.

Proposition 1 (Boyd and Vandenberghe 2004, Section 3.1.4). *Let f be a real function that is differentiable on the open interval (a, b). Then, f is strictly concave on (a, b) if and only if its derivative f' is strictly decreasing on (a, b).*

Loosely speaking, Proposition 1 states that concavity is equivalent to a (decreasing) negative slope. Per our assumptions, we have that the inverse demand function $P(Q)$ is strictly decreasing ($P(Q)$ assumes the role of f' in Proposition 1). Thus, by Proposition 1, $W^*(Q)$ in (20) is a concave function in Q on $(0, b)$ for any $b > 0$.

In summary, we assume that

A1: $P(Q)$ is a strictly decreasing, continuous function, and that

A2: $W^*(Q)$ exists.

Table 1 contains a (nonexhaustive) list of commonly used demand functions in economics. Demand with a price elasticity ϵ between zero and one is considered inelastic and demand with $\epsilon > 1$ is considered elastic. Demand with $\epsilon = 1$ is also called "unit elastic demand." Electricity wholesale markets are considered to be inelastic or perfectly inelastic, depending on the level of demand response in the market. The most commonly used functions to model short-term demand response are linear and partial-log demand functions. An application of the partial-log demand function can be found in Bushnell (2010). Although the isoelastic demand function is not suited for the electricity market context for $\epsilon > 1$, it satisfies assumptions A1 and A2 and, thus, can be used in the solution framework presented in Section 3. Perfectly inelastic demand functions satisfy neither A1 nor A2. A perfectly inelastic demand is either used in a cost-minimizing setting or when applying value-of-lost-load pricing (Stoft 2002). The case of perfectly elastic demand functions assumes a special role in our framework. Though assumption A1 is not satisfied, the corresponding integral function is a linear function. Therefore, perfectly elastic demand functions can be handled by our framework as well. In summary, we use the linear and partial-log demand function to model electricity demand as inelastic.

Next, we describe a mathematical programming formulation that determines the welfare for a given hour and given power system when the investment decisions have been made. Consider the following nonlinear programming (NLP) problem:

$$W^*_{\text{NLP}} := \left\{ \max \int_0^{Q} P(x)\,dx - \sum_{i\in\mathbb{I}} c_i q_i \right\} \tag{21}$$

$$\text{s.t.} \quad Q = \sum_{i\in\mathbb{I}} q_i, \tag{22}$$

$$0 \le q_i \le C_i K_i \quad \forall\, i \in \mathbb{I}. \tag{23}$$

Problem (21)–(23) is a box-constrained nonlinear maximization problem (after substituting the relation (22) for Q into the objective function (21)) with a concave objective function. It belongs to the class of

Table 1. Commonly Used Demand Functions and Their Properties

Name	$Q = f(P)$	$P(Q)$	$\int P(Q)\,dQ$	Parameters	A1	A2
Linear	$Q = a - b \cdot P$	$P = \frac{a}{b} - \frac{Q}{b}$	$\frac{a}{b}Q - \frac{Q^2}{2b}$	$a > 0$: intercept $b > 0$: slope	3	3
Isoelastic	$Q = K \cdot P^{-\epsilon}$	$P = \left(\frac{Q}{K}\right)^{-1/\epsilon}$	$\frac{\epsilon Q (Q/K)^{-1/\epsilon}}{\epsilon - 1}$	K: scale $\epsilon > 0$: elasticity	3	(3) (for $\epsilon > 1$)
Partial-log	$Q = a - b \cdot \log P$	$P = e^{(a-Q)/b}$	$-be^{(a-Q)/b}$	$a > 0$: intercept $b > 0$: slope	3	3
Perfectly inelastic	$Q = c$	—	Unbounded	c: constant	—	—
Perfectly elastic	—	$P = c$	cQ	c: constant	—	3

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

polynomially solvable problems, though it is an open problem as to whether there exists a strongly polynomial solution algorithm for general convex NLPs (Vavasis 1991). Problem (21)–(23) (for each hour h and year t) is the core component of (PGEP).

We utilize an hourly resolution for the market clearing condition for the following reasons:

- (Day-ahead) electricity markets typically have an hourly time resolution and, thus, hourly electricity prices (Maurer and Barroso 2011).
- An hourly time resolution allows a detailed representation of demand and supply patterns that occur during a day. In particular, wind and solar power supply may significantly change throughout the course of a day. It is important to represent the correlation of supply of these generation technologies and demand.
- An hourly time resolution implicitly incorporates multiple supply and demand scenarios, e.g., wind scenarios, compared to models with a load block representation.
- Ultimately, we are interested in capturing the volatility of renewable generators in combination with base load plants such as coal plants. Thus, we require a high time resolution. The start-up restrictions described in Section 2.4 are meaningless in the context of lower time resolutions.

We therefore refrain from only modeling a small number of representative hours but model the day-ahead market in its full resolution. Examples for well-known expansion models that feature a number of representative hours are the U.S. Energy Information Administration's NEMS, the National Renewable Energy Laboratory's Regional Energy Deployment System (ReEDS) (Martinez et al. 2013), the Electric Power Research Institute's U.S. Regional Economy, Greenhouse Gas, and Energy (US-REGEN) model (Blanford et al. 2014), and Resources For the Future's Haiku (Paul et al. 2009). We decided against using a higher time resolution (e.g., 15-minute time intervals) because of the additional complexity it would add, for instance, in the presence of ramping constraints (see Section 2.4). Also, since we assume that we have perfect foresight and do not consider stochastic demand or stochastic outages in our model, we can also assume that we are able to perfectly schedule all generators in an hourly day-ahead market.

2.4. Start-up Restrictions

The operation of a generator is restricted by its operation in previous hours, and we are therefore interested in representing this dependency in (PGEP). We briefly review ramp constraints as they appear in the literature and motivate our choice for (PGEP).

There are three processes to consider:

(i) starting an offline generator until it has reached its minimum generation level,

(ii) load cycling an online generator between its minimum and maximum generation level, and

(iii) shutting down an online generator after generation has been reduced to its minimum generation level.

Mixed-integer linear programming (MILP) techniques allow us to model all three processes above and are widely used. Unit commitment models (Guan et al. 1992, Tseng et al. 2000, Hobbs et al. 2001) belong to this category, and great detail in modeling start-up (i) and shut-down (iii) power trajectories can be achieved (Arroyo and Conejo 2004). However, load cycling an online generator, as described in (ii), is typically not seen as critical with an hourly time resolution (Gollmer et al. 2000, Lindsay and Dragoon 2010). Currently, coal plants are used in a more flexible operation and can load cycle from their minimum to their maximum generation level within the course of an hour, although their flexibility strongly depends on the plant type, e.g., subcritical versus supercritical. Processes (i) and (ii) are of importance because different start-up costs have to be applied (Lindsay and Dragoon 2010), dependent on the duration for which a generator has been offline. Describing the physical behavior of the generators in even greater detail leads to mixed-integer nonlinear programming (MINLP) or NLP models with variable ramping costs that reflect the physical changes of the generators (Wang and Shahidehpour 1995) and model valve-point effects of generators (Han et al. 2001, Xia and Elaiw 2010).

However, all of these modeling approaches have in common that they affect the structure of the problem by either introducing binary variables or nonlinearities. Although (PGEP) is an MINLP, we have shown in Section 2.3 that the nonlinear term in (1) is concave. Binary variables γ^{N}_{it} and γ^{D}_{it} can be handled by a decomposition approach that we present in Section 3. We thereby avoid the introduction of additional nonlinear and integral terms into (PGEP) that would destroy its structure.

Linearized ramp constraints describe an alternative to these binary and nonlinear constructs (DeJonghe et al. 2011a, Warland et al. 2008). In particular, Warland et al. (2008) discuss start-up costs for a system with thermal generators and hydropower plants and show that linear constraints are a good proxy. Hydrothermal scheduling models are typically solved with Benders decomposition-type algorithms, which require LP subproblems (Pereira and Pinto 1991). For our long-term generation expansion problem, linearized start-up constraints (4)–(8) are a good trade-off between modeling detail and complexity. In this regard, our model represents a compromise between an MILP unit commitment model and a load block representation without

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

any form of the start-up restrictions applied by most expansion models.

Constraints (4) divide the load of the generators q_{iht} into load below the assumed minimum generation, q^{L}_{iht}, and between minimum and maximum generation, q^{U}_{iht}. For instance, assuming a minimum generation of 40% for a given generator i, one would assign $P^{\min}_i = 0.4C_{ih}K_i$ and $P^{\max}_i = C_{ih}K_i$. Then, if $q_{iht} = C_{ih}K_i$, it follows that $q^{\mathrm{L}}_{iht} = q^{\mathrm{U}}_{iht} = 1$. Constraints (5) enforce the load below minimum generation to always exceed the load above minimum generation. Constraints (6) define the variables δ_{iht} as start-up of a generator and are penalized in the objective function. Note that generators can run at a level below their respective minimum generation using these constraints. Thus, they only describe an approximation of the real world. In particular, constraints (4)–(6) approximate process (i). The start-up costs are proportional to the respective generator's capacity, and the respective generator's capacity is used in constraint (4).

2.5. Transmission Constraints

Model (PGEP) assumes that the electricity wholesale market is divided into multiple zones (or nodes), each with different demand functions, loads, and, ultimately, electricity prices. Each zone has a certain demand and associated generators to supply this demand. Transmission between zones is possible and is governed by transmission lines. High demand in a zone with low supply can lead to congestion in transmission lines between zones. This results in zonal prices of electricity instead of a single price for the entire power system. In past years, efforts (Gomez-Exposito et al. 2008, Kazerooni and Mutale 2010) have been made to increase the modeling detail with the aim of better accounting for demand and supply imbalances and the resulting congestion in transmission lines, leading to nodal models. As an example, the ERCOT region was divided into four load zones until 2010, but the independent system operator's current nodal model divides the market into 4,000 price nodes to better capture and price the congestion.

In this work, we restrict ourselves to a zonal representation of the power market in which power flow does not adhere to the laws of physics but is represented in the form of a simple network model. Constraints (9) and (10) model the flow and its capacity restrictions and describe what we refer to as a "pipeline representation," although oil flowing through a pipeline is governed by laws of physics as well (De Wolf and Smeers 1996). Thus, our representation stands in contrast to a direct current (DC) or alternating current (AC) model (Frank et al. 2012). However, (PGEP) can be adopted for a nodal market by replacing constraints (9) and (10) with the linear representation of DC power flow (Castillo et al. 2002). Since our model is only comprised of four major load zones, a DC model introduces unnecessary detail and complexity. In fact, constraints (9) and (10) are common practice in this setting to guarantee computational tractability and can be found in most expansion planning models (U.S. Energy Information Administration 2015, EPIS 2015, Nolden et al. 2013).

However, the algorithm presented in Section 3.1 is able to handle a DC representation since the constraints remain linear. Parts of the algorithm in Section 3.2.5 rely on the representation through constraints (9) and (10) and would have to be adapted to work with DC flows. An AC representation of power flow introduces nonlinearities into the constraint set that cannot be handled by either algorithm.

2.6. Investment and Decommissioning Decisions

(PGEP) features multiperiod investment and annual decommission decisions of power plants. In both cases, we either let the model make the optimal decision or implement official announcements of new projects or closing plants. Both investment and decommissioning are effective immediately; i.e., new generators and transmission lines can be used in the same year and existing generators are turned off completely.

We represent investment decisions using binary variables in (PGEP), and constraints (13) ensure that a generator is built at most once. All algorithms presented in Section 3 are able to handle continuous investment decisions as well. In fact, continuous investments would drastically reduce the number of variables in (PGEP), because generators of the same technology could be aggregated into one variable. This represents a viable simplification for technologies such as solar and wind generators, because of their small capacities. Furthermore, a portfolio of discrete investment projects may be associated with certain strategic decisions, and one may argue that, by adding investment options, these strategic decisions change. Such a project portfolio might be fitting from an individual investor's perspective but could be problematic from a policy analysis standpoint.

Either modeling assumption for the type of investment variable works with start-up restrictions (4)–(8). Although the capacity of new generators depends on the investment decisions γ^{N}_{it}, the correct start-up cost is incurred in the case of γ^{N}_{it} being continuous. Consider the following example: assume that there is only one investment period, i.e., $|\mathbb{T}^{\mathrm{I}}| = 1$, and further assume that new generator i with $\gamma^{\mathrm{N}}_i < 1$ is offline in hour $h-1$ and is ramped up to full capacity, i.e., $C_{ih}K_i\gamma^{\mathrm{N}}_i$, in hour h. It follows that $q^{\mathrm{L}}_{ih} = \gamma^{\mathrm{N}}_i = \delta_{ih}$. Thus, the ramping cost is $\hat{c}_{it}\gamma^{\mathrm{N}}_i K_i$, which is correct.

We set an artificial investment limit for new transmission lines to allow the use of continuous decision variables. Constraints (14) ensure that the limit across all investment periods is not exceeded. In case this limit is binding, we increase it and resolve the model.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Existing plants can be decommissioned to avoid their annual operation and maintenance costs. We might be interested in knowing which plants or technologies become obsolete under certain policies and scenarios or implement announcements of closing plants. Constraints (12) enforce that retired plants stay offline. Decommission decisions were chosen to be binary to avoid partial retirement decisions.

2.7. Model Extensions

We briefly address possible model extensions and state the reasons why we chose not to incorporate them into our model.

Uncertainty in fuel prices and capacity factors allows us to hedge our investment decisions against shifts in market prices and wind volatility. For instance, Jin et al. (2011) consider a two-stage model with uncertain demand and fuel prices. We already addressed wind volatility in Section 2.3, so let us consider the fuel prices. Given the size of model (1)–(17), modeling uncertainty is a difficult task and might lead to an intractable model, even for small scenario trees using decomposition methods. It is certainly possible to introduce uncertainty for special cases of the model such as the core model discussed in Sections 3.2.1 and 3.2.4. However, from a case study standpoint, we are much more interested in running specific fuel price scenarios than hedging against multiple ones at the same time.

Imperfect competition in the market is an assumption that is not possible in our welfare maximization framework. We assume that all market participants are price takers in a competitive market. With decreasing natural gas prices in the past years and a resulting flat supply curve, exercising market power through decreasing supply bids of base load plants such as coal is very limited. In back-testing, (PGEP) yields electricity prices that are close to the observed prices.

Risk aversion with respect to the investment and decommissioning decisions is not represented, but it is possible to incorporate. In our model, an investment decision is favorable as soon as its discounted returns over the planning horizon exceed its initial cost. Risk aversion is typically represented by a concave utility function and, dependent on its shape, keeps us from making marginally profitable investments or retiring marginally unprofitable generators. The utility function would enter the objective function and, thus, preserve its concavity. However, we decide against modeling risk aversion since we do not explicitly incorporate uncertainty.

3. Solution Algorithms

Model (PGEP) is a very large-scale MINLP because of the hourly resolution in combination with a planning horizon of more than two decades. Consequently, (PGEP) cannot be solved efficiently as a monolith; see Section 4. Thus, we must exploit the problem structure: for given investment decisions γ^{N}_{it} and $\gamma^{\mathrm{T}}_{uvt}$, and decommissioning decisions γ^{D}_{it}, the model decomposes into $|\mathbb{H}|\cdot|\mathbb{T}|$ zonal market clearing problems coupled by constraints (6). This naturally suggests a type of Benders decomposition approach (Benders 1962).

More specifically, because the subproblem is a convex NLP, a GBD approach must be applied (Geoffrion 1972). GBD is an extension of classical Benders decomposition (Benders 1962, 2005) in that it can handle NLPs. For computational speed and robustness, we propose a novel Benders decomposition-type approach in which we linearly and dynamically overestimate the concave objective function of the subproblems, transforming the NLPs into LPs. Although we solve LPs, our approach yields the optimal solution for all special cases of (PGEP) considered as shown later in this section. We show computational results of solving NLPs versus LPs as subproblems in Section 4.

For an LP with two types of variables, the basic idea of Benders decomposition is to treat the so-called complicating variables—these are the variables that connect the problem—in a master problem and the other type of variables in the subproblem. The information of the subproblem is passed to the master problem via an outer linearization with two types of hyperplanes: the Benders feasibility and optimality cuts. The key concept is that the feasible region of the dual problems of the subproblems is independent of any selection of the γ^{N}_{it}, $\gamma^{\mathrm{T}}_{uvt}$, and γ^{D}_{it} variables. This allows for an exact representation of the subproblem with a finite number of hyperplanes. The same method can be applied toward an MILP if all integer variables are moved to the master problem while the subproblem remains an LP.

3.1. Generalized Benders Decomposition Using Dynamic Overestimation

We follow the idea of (generalized) Benders decomposition, in that we decompose (PGEP) into two problems: a relaxed master problem (Master) and a subproblem (Sub). Problem (Master) reads

$$
\begin{aligned}
w^* := \Bigg\{ \max & -\sum_{t\in\mathbb{T}^{\mathrm{I}}} \beta^{(t-1)} \left[\sum_{i\in\mathbb{I}^{\mathrm{N}}} \gamma^{\mathrm{N}}_{it} F^{\mathrm{I}}_{it} K_i + \sum_{uv\in\Omega} \gamma^{\mathrm{T}}_{uvt} F^{\mathrm{T}}_{uvt} K^{\mathrm{T}}_{uv} \right] \\
& - \sum_{t\in\mathbb{T}} (\beta)^t \left[\sum_{i\in\mathbb{I}^{\mathrm{N}}} \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \gamma^{\mathrm{N}}_{it'} F^{\mathrm{OM}}_i K_i + \sum_{i\in\mathbb{I}^{\mathrm{EX}}} \gamma^{\mathrm{D}}_{it} F^{\mathrm{OM}}_i K_i \right] + \eta \Bigg\} \\
\text{s.t.} \quad & \eta \le \sum_{t\in\mathbb{T}} \sum_{i\in\mathbb{I}^{\mathrm{EX}}} \omega^{\mathrm{D}}_{ijt} \gamma^{\mathrm{D}}_{it} \\
& \quad + \sum_{t\in\mathbb{T}^{\mathrm{I}}} \left(\sum_{i\in\mathbb{I}^{\mathrm{N}}} \omega^{\mathrm{N}}_{ijt} \gamma^{\mathrm{N}}_{it} + \sum_{uv\in\Omega} \omega^{\mathrm{T}}_{uvt} \gamma^{\mathrm{T}}_{uvt} \right) \\
& \quad + \omega^{\mathrm{c}}_j \quad \forall j \in \mathbb{J}, \quad (12)\text{–}(17). \qquad (24)
\end{aligned}
$$

Unrestricted variable η is commonly referred to as "future benefit function," although "future welfare function" is a more appropriate term in our setting.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

The GBD subproblem (Sub), for trial values $\bar{\gamma} := (\bar{\gamma}^{\mathrm{D}}, \bar{\gamma}^{\mathrm{N}}, \bar{\gamma}^{\mathrm{T}})$, reads

$$W^*(\bar{\gamma}) := \max \sum_{t\in\mathbb{T}} \sum_{h\in\mathbb{H}} (\beta)^t \left(\sum_{u\in\mathbb{U}} \int_0^{Q_{uht}} P_{uht}(x)\,\mathrm{d}x - \sum_{i\in\mathbb{I}} c_{it} q_{iht} - \sum_{i\in\mathbb{I}^{\mathrm{R}}} \hat{c}_{it} K_i \delta_{iht} \right) \tag{25}$$

$$\text{s.t.} \quad 0 \le q_{iht} \le \bar{\gamma}^{\mathrm{D}}_{it} C_{ih} K_i \quad \forall\, i \in \mathbb{I}^{\mathrm{EX}},\, h \in \mathbb{H},\, t \in \mathbb{T}, \quad (\boldsymbol{\lambda}^*), \tag{26}$$

$$0 \le q_{iht} \le \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \bar{\gamma}^{\mathrm{N}}_{it'} C_{ih} K_i \quad \forall\, i \in \mathbb{I}^{\mathrm{N}},\, h \in \mathbb{H},\, t \in \mathbb{T}, \quad (\boldsymbol{\lambda}^*), \tag{27}$$

$$0 \le x_{uvht} \le F^{\max}_{uv} + \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\le t}} \bar{\gamma}^{\mathrm{T}}_{uvt'} K^{\mathrm{T}}_{uv} \quad \forall\, uv \in \Omega,\, h \in \mathbb{H},\, t \in \mathbb{T}, \quad (\boldsymbol{\pi}^*), \tag{28}$$

$$(4)\text{–}(9),\ (11), \tag{29}$$

with some optimal dual solution vectors $\boldsymbol{\lambda}^*$ and $\boldsymbol{\pi}^*$. Note that (Sub) is a convex NLP and, thus, possesses a zero duality gap. With that, its dual solutions are well defined, though multiple dual optimal solutions might exist.

Cut coefficients $\omega^{\mathrm{D}}_{ijt}$, $\omega^{\mathrm{N}}_{ijt}$, and $\omega^{\mathrm{T}}_{juvt}$, as well as right-hand side constant ω^{c}_{j} for the Benders optimality cuts (24) are obtained as follows, for some cut index $j \in \mathbb{J}$, and all $t \in \mathbb{T}$:

$$\omega^{\mathrm{D}}_{ijt} = \sum_{h\in\mathbb{H}} \lambda^{\mathrm{c}}_{iht} C_{ih} K_i \quad \forall\, i \in \mathbb{I}^{\mathrm{EX}}, \tag{30}$$

$$\omega^{\mathrm{N}}_{ijt} = \sum_{h\in\mathbb{H}} \sum_{\substack{t'\in\mathbb{T} \\ t'\ge t}} \lambda^{\mathrm{c}}_{iht'} C_{ih} K_i \quad \forall\, i \in \mathbb{I}^{\mathrm{N}}, \tag{31}$$

$$\omega^{\mathrm{T}}_{juvt} = \sum_{h\in\mathbb{H}} \sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}} \\ t'\ge t}} \pi^{\mathrm{c}}_{uvht} K^{\mathrm{T}}_{uv} \quad \forall\, uv \in \Omega, \tag{32}$$

$$\omega^{\mathrm{c}}_{jt} = W^*(\bar{\gamma}) - \sum_{t\in\mathbb{T}} \sum_{i\in\mathbb{I}^{\mathrm{EX}}} \omega^{\mathrm{D}}_{ijt} \bar{\gamma}^{\mathrm{D}}_{it} - \sum_{t\in\mathbb{T}^{\mathrm{I}}} \left(\sum_{i\in\mathbb{I}^{\mathrm{N}}} \omega^{\mathrm{N}}_{ijt} \bar{\gamma}^{\mathrm{N}}_{it} + \sum_{uv\in\Omega} \omega^{\mathrm{T}}_{juvt} \bar{\gamma}^{\mathrm{T}}_{uvt} \right). \tag{33}$$

We omit feasibility cuts for (Master) because (Sub) is feasible for any choice $\bar{\gamma}$ satisfying (12)–(17). That means it is feasible to have no generation capacity, i.e., no (or low) investments and complete decommissioning of the existing fleet. However, the resulting electricity price would be extremely high (in the case of a linear demand function; see Table 1) or infinite (in the case of a partial-log demand function; see Table 1), which is not desirable and affects algorithm performance. In Section 2.3, we stated that we introduce an artificial generator with a sufficiently high marginal cost. This artificial generator is not allowed to be decommissioned, which (1) avoids dealing with infinite electricity prices and (2) leads to numerically superior dual variables.

For (Sub), we propose to linearly overestimate the concave functions

$$\zeta_{uht}(Q_{uht}) := \int_0^{Q_{uht}} P_{uht}(x)\,\mathrm{d}x$$

in variables Q_{uht} present in the objective function (25) (the overestimation is crucial for the correctness of the obtained upper bounds; see below) at a finite number of break points. This yields the following linearized and overestimated subproblem (Sub:O) for trial $\bar{\gamma}$:

$$\bar{W}(\bar{\gamma}) := \max \sum_{t\in\mathbb{T}} \sum_{h\in\mathbb{H}} (\beta)^t \left(\sum_{u\in\mathbb{U}} \phi_{uht} - \sum_{i\in\mathbb{I}} c_{it} q_{iht} - \sum_{i\in\mathbb{I}^{\mathrm{R}}} \hat{c}_{it} K_i \delta_{iht} \right) \tag{34}$$

$$\text{s.t.} \quad \phi_{uht} \le \omega^{\mathrm{s}}_{\kappa uht} Q_{uht} + \omega^{\mathrm{c}}_{\kappa uht} \quad \forall\, \kappa \in \mathbb{K}_{uht},\, u \in \mathbb{U},\, h \in \mathbb{H},\, t \in \mathbb{T},$$
$$(26)\text{–}(29),$$

where $\omega^{\mathrm{s}}_{\kappa uht}$ and $\omega^{\mathrm{c}}_{\kappa uht}$ are slope and constant of the affine function overestimating $\zeta_{uht}(\cdot)$, for cut index $\kappa \in \mathbb{K}_{uht}$. To distinguish the cuts for $\zeta_{uht}(\cdot)$ from the Benders optimality cuts, we call $\kappa \in \mathbb{K}_{uht}$ the "break point set" and κ the "break point index" for the remainder of this paper.

At break point y, the slope of the linear overestimator for break point index κ is given by

$$\omega^{\mathrm{s}}_{\kappa uht}(y) := \frac{\mathrm{d}}{\mathrm{d}y} \zeta_{uht}(y) = P_{uht}(y) \quad \forall\, u \in \mathbb{U},\, h \in \mathbb{H},\, t \in \mathbb{T},$$

and its constant is calculated via

$$\omega^{\mathrm{c}}_{\kappa uht}(y) := \zeta_{uht}(y) - \omega^{\mathrm{s}}_{\kappa uht}(y) y \quad \forall\, u \in \mathbb{U},\, h \in \mathbb{H},\, t \in \mathbb{T}.$$

We propose to start with a coarse grid of break points, to initialize the break point set $\mathbb{K}_{uht}$ for all $u \in \mathbb{U}$, $h \in \mathbb{H}$, $t \in \mathbb{T}$ in the preprocessing step of the algorithm. This is a static discretization. Throughout the Benders iterations, we add break points dynamically at computed quantities Q_{uht} if the objective function value of the overestimated problem is too far away from the exact objective function value evaluated for that Q_{uht}. A detailed description of the generalized Benders decomposition algorithm using dynamic linear overestimators (BD-DO) is given below (for some user-defined tolerance $\epsilon > 0$).

Benders Decomposition Using Dynamic Overestimator (BD-DO) to Solve (PGEP)

1. *Initialize*. Empty Benders cut set $\mathbb{J} = \varnothing$, initialize lower bound $LB = -\infty$, initialize set of overestimator cuts $\mathbb{K}_{uht}$, and define some trial $\bar{\gamma}$ satisfying (12)–(17).

2. *Solve* (*Sub*:*O*). For trial $\bar{\gamma}$, solve (Sub:O) to obtain dual solutions $\boldsymbol{\lambda}^*$ and $\boldsymbol{\pi}^*$ as well as equilibrium quantities Q^*_{uht} and overestimator value ϕ^*_{uht}.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

3. *Update overestimator*. If

$$\sum_{t\in\mathbb{T}}\sum_{h\in\mathbb{H}}\sum_{u\in\mathbb{U}}(\beta)^t\left(\phi^*_{uht}-\zeta_{uht}(Q^*_{uht})\right)>\frac{\epsilon}{2},$$

construct overestimating cut (34) at $y=Q^*_{uht}$ and add additional break point index to $\mathbb{K}_{uht}$ for all zones $u\in\mathbb{U}$, all hours $h\in\mathbb{H}$, and all years $t\in\mathbb{T}$.

4. *Update LB*. Calculate

$$\begin{aligned}\underline{w}=\bar{W}^*(\bar{\gamma})&+\sum_{t\in\mathbb{T}}\sum_{h\in\mathbb{H}}\sum_{u\in\mathbb{U}}(\beta)^t(\zeta_{uht}(Q^*_{uht})-\phi^*_{uht})\\&-\sum_{t\in\mathbb{T}^{\mathrm{I}}}\beta^{(t-1)}\left[\sum_{i\in\mathbb{I}^{\mathrm{N}}}\bar{\gamma}^{\mathrm{N}}_{it}F^{\mathrm{I}}_{it}K_i+\sum_{uv\in\Omega}\bar{\gamma}^{\mathrm{T}}_{uvt}F^{\mathrm{T}}_{uvt}K^{\mathrm{T}}_{uv}\right]\\&-\sum_{t\in\mathbb{T}}(\beta)^t\left[\sum_{i\in\mathbb{I}^{\mathrm{N}}}\sum_{\substack{t'\in\mathbb{T}^{\mathrm{I}}\\t'\le t}}\bar{\gamma}^{\mathrm{N}}_{it'}F^{\mathrm{OM}}_{i}K_i+\sum_{i\in\mathbb{I}^{\mathrm{EX}}}\bar{\gamma}^{\mathrm{D}}_{it}F^{\mathrm{OM}}_{i}K_i\right],\end{aligned}$$

and update lower bound $LB=\max\{LB,\underline{w}\}$.

5. *Construct cut*. Construct the Benders optimality cut for (Master) via (30)–(32) and

$$\begin{aligned}\omega^{\mathrm{c}}_{it}=\bar{W}^*(\bar{\gamma})&-\sum_{t\in\mathbb{T}}\sum_{i\in\mathbb{I}^{\mathrm{EX}}}\omega^{\mathrm{D}}_{iit}\bar{\gamma}^{\mathrm{D}}_{it}\\&-\sum_{t\in\mathbb{T}^{\mathrm{I}}}\left(\sum_{i\in\mathbb{I}^{\mathrm{N}}}\omega^{\mathrm{N}}_{iit}\bar{\gamma}^{\mathrm{N}}_{it}+\sum_{uv\in\Omega}\omega^{\mathrm{T}}_{\iota uvt}\bar{\gamma}^{\mathrm{T}}_{uvt}\right);\end{aligned}$$

add additional cut index ι to $\mathbb{J}$.

6. *Solve* (*Master*). Solve (Master) to obtain new trial $\bar{\gamma}=(\gamma^{\mathrm{D},*},\gamma^{\mathrm{N},*},\gamma^{\mathrm{T},*})$ and to update upper bound $UB=w^*$.

7. *Check convergence*. If $UB-LB\le\epsilon$, STOP (optimal solution found); otherwise, go to step 2.

Proposition 2. *The Benders decomposition-type algorithm BD-DO is correct and converges after finitely many iterations to an ϵ optimal solution for* (*PGEP*) *for any $\epsilon>0$, i.e., a feasible solution to* (*PGEP*) *with objective function value $\ge W^*_{\mathrm{NLP}}-\epsilon$.*

Proof. We need to show two things: (1) the Benders optimality cuts derived from solving (Sub:O) are valid (step 5 in BD-DO), and (2) finitely many optimality cuts suffice to reach convergence.

1. Subproblem (Sub:O) overestimates the true optimal objective function value; i.e., we have $\bar{W}^*(\bar{\gamma})\ge W^*(\bar{\gamma})$ for any choice of $\bar{\gamma}=(\bar{\gamma}^{\mathrm{D}},\bar{\gamma}^{\mathrm{N}},\bar{\gamma}^{\mathrm{T}})$ satisfying (12)–(17). Because the Benders optimality cuts derived from (Sub:O) overestimate the (piecewise linear, concave) function $\bar{W}(\cdot)$, the cut is also valid for (Sub).

2. We can reach an ϵ optimality tolerance in finitely many iterations as follows. We observe that objective function (25) can be approximated using cuts (34) such that $|W^*(\cdot)-\bar{W}(\cdot)|\le\epsilon/2$. Because function (25) is continuous, finitely many cuts suffice. Because the Benders decomposition method for LPs converges to a tolerance of $\epsilon/2$ in finitely many steps, finite convergence is guaranteed. □

Remark 1. Note that the optimality cuts derived from (Sub:O) might not be tight in some iterations, especially in the early ones.

Remark 2. Instead of using an absolute tolerance criterium, ϵ, one might use a relative one. For $\bar{\varepsilon}>0$, we then ask $UB-LB\le\bar{\varepsilon}\cdot LB$ in step 7 of BD-DO. Step 3 can be changed as follows.

3. *Update overestimator*. For all zones $u\in\mathbb{U}$, all hours $h\in\mathbb{H}$, and all years $t\in\mathbb{T}$, if

$$\phi^*_{uht}-\zeta_{uht}(Q^*_{uht})>\frac{\bar{\varepsilon}}{2}\left(\zeta_{uht}(Q^*_{uht})-\sum_{i\in\mathbb{I}}c_{it}q^*_{iht}-\sum_{i\in\mathbb{I}^{\mathrm{R}}}\hat{c}_{it}K_i\delta^*_{iht}\right),$$

then construct overestimating cut (34) at $y=Q^*_{uht}$; add additional break point index to $\mathbb{K}_{uht}$.
This avoids adding overestimator cuts to hours that are already sufficiently approximated.

3.2. Special Cases of the Subproblem

Although the above decomposition BD-DO allows us to solve large problem instances of (PGEP), this approach is limited by the computational tractability of (Sub:O). In other words, solving (Sub:O) as a monolith becomes very time and memory intensive. In our case, (Sub:O) quickly exceeds the memory of off-the-shelf personal computers for instances that span several years (see Section 4). We therefore consider three special cases of model (PGEP) in order to further exploit (Sub:O)'s structure. We begin by describing three "traditional" mathematical programming approaches in which primal and dual solution values of the (Sub:O) are obtained by solving corresponding mathematical programming problems via a standard LP solver (Sections 3.2.1–3.2.3). Then, we demonstrate that necessary primal and dual solution values for the Benders subproblem can be explicitly calculated in some cases (Sections 3.2.4–3.2.6). This leads to highly efficient algorithms that significantly outperform their "traditional" counterparts. Thus, the analysis of these special cases in this section serves two purposes: (1) we demonstrate how to make the large (Sub:O) tractable, and (2) we are able to identify structures that can be exploited in cut calculation approaches. The three special cases are as follows.

(*Core*). We define the core model as the combination of the hourly market clearing component (see Section 2.3) for one load zone and annual investment and decommissioning decisions (and Section 2.6). This means that neither start-up restrictions (4)–(8) nor transmission constraints (9) and (10) are part of this model. In Sections 3.2.1 and 3.2.4, two efficient algorithms are presented to solve (Core).

(*Core*:*T*). We extend the core model by modeling multiple load zones using transmission constraints (9) and (10) and annual transmission capacity expansions. We present the modified algorithms in Sections 3.2.2 and 3.2.5.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

(*Core*:*R*). We extend the core model using start-up restrictions (4)–(8), which introduces a coupling of hours. We present the modified algorithms in Sections 3.2.3 and 3.2.6.

3.2.1. GBD with Dynamic Linear Overestimation for (Core). Model (Core) simplifies both the relaxed Master problems (Master) as well as the linearized subproblem (Sub:O). We call the resulting subproblem (Sub:Core) and (Sub:Core:O). Most notably, the subproblem (Sub:Core:O) decomposes into $|\mathbb{H}| \cdot |\mathbb{T}|$ independent linear programming problems. Consequently, if (Core) is a reasonable simplification of (PGEP), memory restrictions are no longer an issue, because (Sub:O) can be further decomposed into problems as small as one hour. Furthermore, the solution time drastically decreases, because parallel programming schemes become applicable.

3.2.2. GBD with Dynamic Linear Overestimation for (Core:T). Similar to (Core), in model (Core:T), the subproblem (Sub:O) decomposes into $|\mathbb{H}| \cdot |\mathbb{T}|$ independent LPs denoted by $(\text{Sub:T:O})_{ht}$ for $h \in \mathbb{H}$ and $t \in \mathbb{T}$. Each such LP is then an hourly market-clearing problem considering multiple load zones. We denote the subproblem (Sub:O) for model (Core:T) by (Sub:T:O). The benefits of simplifying (PGEP) into (Core:T) are the same as described in Section 3.2.1.

3.2.3. GBD with Dynamic Linear Overestimation for (Core:R). In contrast to (Core) and (Core:T), the presence of linear start-up restrictions (4)–(8) in (Core:R) introduces a coupling of consecutive hours (Sub:R:O), which denotes the resulting overestimated subproblem (Sub:O) for (Core:R). The Benders subproblem (Sub:R:O) no longer decomposes with h and t. This coupling makes it significantly harder to solve subproblem (Sub:R:O), compared to the (Core) and (Core:T) models. Although (Core:R) is a special case of (PGEP), the following description directly leads to a solution method for (Sub:O) and, therefore, (PGEP).

To solve LP (Sub:R:O), we suggest an NBD algorithm (Ho and Manne 1974, Birge and Louveaux 2011). This algorithm allows the decomposition into $|\mathbb{H}| \cdot |\mathbb{T}|$ one-hour problems (having one zone each), coupled from one hour to the next via state variables $\bar{q}^{\mathrm{L}}_{ih't'}$. In the following discussion, we omit the zonal index u, where applicable. For some trial $\bar{\gamma} = (\bar{\gamma}^{\mathrm{N}}, \bar{\gamma}^{\mathrm{D}})$ and all $h \in \mathbb{H}$ and $t \in \mathbb{T}$, these one-hour problems are given by $(\text{Sub:R:O})_{ht}$

$$
\begin{aligned}
&\tilde{W}^*_{ht}(\bar{\gamma}, \bar{q}^{\mathrm{L}}_{ih't'}) \\
&\quad := \max\left\{(\beta)^t\left(\phi_{ht} - \sum_{i\in\mathbb{I}} c_{it} q_{iht} - \sum_{i\in\mathbb{I}^{\mathrm{R}}} \hat{c}_{it} K_i \delta_{iht}\right) + \tilde{\eta}_{ht}\right\} \\
&\qquad \text{s.t.} \quad \phi_{ht} \le \omega^{\mathrm{s}}_{\kappa ht} Q_{ht} + \omega^{\mathrm{c}}_{\kappa ht} \quad \forall\, \kappa \in \mathbb{K}_{ht}, \\
&\qquad\qquad 0 \le q_{iht} \le \bar{\gamma}^{\mathrm{D}}_{it} C_{ih} K_i \quad \forall\, i \in \mathbb{I}^{\mathrm{EX}}, \qquad (35) \\
&\qquad\qquad 0 \le q_{iht} \le \sum_{\substack{t' \in \mathbb{T}^{\mathrm{I}} \\ t' \le t}} \bar{\gamma}^{\mathrm{N}}_{it'} C_{ih} K_i \quad \forall\, i \in \mathbb{I}^{\mathrm{N}}, \qquad (36)
\end{aligned}
$$

$$
\begin{aligned}
&P^{\min}_i q^{\mathrm{L}}_{iht} + (P^{\max}_i - P^{\min}_i) q^{\mathrm{U}}_{iht} = q_{iht} \quad \forall\, i \in \mathbb{I}^{R}, \\
&q^{\mathrm{L}}_{iht} - q^{\mathrm{U}}_{iht} \ge 0 \quad \forall\, i \in \mathbb{I}^{R}, \\
&q^{\mathrm{L}}_{iht} \le \delta_{iht} + \bar{q}^{\mathrm{L}}_{ih't'}, \\
&\qquad\qquad \forall\, i \in \mathbb{I}^{R}, (h', t') \in \mathcal{A}_{ht}, \quad (\tilde{\lambda}^*), \qquad (37) \\
&\tilde{\eta}_{ht} \le \sum_{i\in\mathbb{I}^{\mathrm{R}}} \tilde{\omega}^{\mathrm{s}}_{ijht} q^{\mathrm{L}}_{iht} + \tilde{\omega}^{\mathrm{c}}_{jht} \quad \forall\, j \in \tilde{\mathbb{J}}_{ht}, \qquad (38) \\
&0 \le q^{\mathrm{L}}_{iht},\ q^{\mathrm{U}}_{iht} \le 1 \quad \forall\, i \in \mathbb{I}^{R}.
\end{aligned}
$$

LP $(\text{Sub:R:O})_{ht}$ also contains Benders optimality cuts (38) to "price" the future benefit associated with q^{L}_{iht}. For $h \in \mathbb{H}$ and $t \in \mathbb{T}$, these cuts are calculated through

$$
\tilde{\omega}^{\mathrm{s}}_{ijh't'} = \tilde{\lambda}^{\mathrm{c}}_{iht} \quad \forall\, i \in \mathbb{I}^{R}, \qquad (39)
$$

$$
\tilde{\omega}^{\mathrm{c}}_{jh't'} = \tilde{W}^*_{ht}(\bar{\gamma}, \bar{q}^{\mathrm{L}}_{ih't'}) - \sum_{i\in\mathbb{I}^{\mathrm{R}}} \tilde{\omega}^{\mathrm{s}}_{ijh't'} \bar{q}^{\mathrm{L}}_{ih't'}, \qquad (40)
$$

for some index $j \in \tilde{\mathbb{J}}_{ht}$ and $(h', t') \in \mathcal{A}_{ht}$. Vector $\tilde{\lambda}^{\mathrm{c}}$ is an optimal dual solution associated with constraint (37), and $\tilde{\mathbb{J}}_{ht}$ is the cut set for hour $h \in \mathbb{H}$ in year $t \in \mathbb{T}$. Again, we can omit feasibility cuts because problem (Sub:R:O) is feasible for any choice and combination of $\bar{\gamma}$ and $\bar{q}^{\mathrm{L}}_{ih't'}$.

Note that $(\text{Sub:R:O})_{ht}$ does not necessarily have to be a one-hour problem but can be a block of multiple consecutive hours. In that case, state variables enter the problem in the first hour of the block, and the optimality cuts are associated with the last hour of the block. This results in the following trade-off: the more hours each problem contains, the fewer Benders iterations for convergence are expected; however, the effort to solve each such problem increases. During our computations, we observed that a grouping of 100 hours yields a good algorithmic performance.

The BD-DO algorithm can now be used in conjunction with an NBD approach to solve (Sub:R:O). This NBD iteratively solves the $(\text{Sub:R:O})_{ht}$ problems. However, to obtain stronger Benders optimality cuts in step 5 of BD-DO, it is advantageous to dynamically update the linear overestimators for $\zeta_{ht}(\cdot) \equiv \zeta_{uht}(\cdot)$ while iterating in the NBD algorithm, instead of once in each iteration of BD-DO. This way, the resulting algorithm solves the subproblem (Sub:R:O) up to some user-defined tolerance. We merge steps 2 and 3 in BD-DO and solve the following, what we call "Inner NBD" (for some trial $\bar{\gamma}$ and initial state variable level $\bar{q}^{\mathrm{L}}_{i00}$ and tolerance TOL).

Inner NBD to Solve (Sub:R:O)

1. *Initialize*. Empty cut sets $\tilde{\mathbb{J}}_{ht} = \varnothing\ \forall\, h \in \mathbb{H}$, $t \in \mathbb{T}$, initialize lower bound $\widetilde{LB} = -\infty$, and initialize trial state variable levels $\bar{q}^{\mathrm{L}}_{iht}$.

2. *Backward step*. For years $\tilde{t} = |\mathbb{T}|, \ldots, 1$ and hours $\tilde{h} = |\mathbb{H}|, \ldots, 1$,

2.1 *Solve* $(\mathit{Sub}\text{:}R\text{:}O)_{\tilde{h}\tilde{t}}$. Solve $(\text{Sub:R:O})_{\tilde{h}\tilde{t}}$ to obtain dual solutions $\tilde{\lambda}^*$ and optimal objective function value $\tilde{W}^*_{\tilde{h}\tilde{t}}(\bar{\gamma}, \bar{q}^{\mathrm{L}}_{ih't'})$.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

2.2 *Cut*. Construct optimality cut (38) using $\tilde{\lambda}^*$ and $\tilde{W}^*_{\tilde{h}\tilde{t}}(\bar{\gamma}, \bar{q}^{\mathrm{L}}_{ih't'})$; add cut to cut set $\tilde{\mathbb{J}}_{h't'}$ for $(h', t') \in \mathcal{A}_{\tilde{h}\tilde{t}}$.

3. *Update upper bound*. $\widetilde{UB} = \tilde{W}^*_{11}(\bar{\gamma}, \bar{q}^{\mathrm{L}}_{i00})$.

4. *Forward step*. For years $\tilde{t} = 1, \ldots, |\mathbb{T}|$ and hours $\tilde{h} = 1, \ldots, |\mathbb{H}|$,

4.1 *Solve* $(Sub{:}R{:}O)_{\tilde{h}\tilde{t}}$. Solve $(\mathrm{Sub{:}R{:}O})_{\tilde{h}\tilde{t}}$ to obtain overestimator value ϕ^*_{ht}, quantity q^*_{iht}, and start-up variable δ^*_{iht}.

4.2 *Update duals and equilibrium quantity*. Update dual solutions λ^* associated with constraints (35) and (36) and equilibrium quantity Q^*_{ht} (returned if terminated in step 6).

4.3 *Update trial state variables*. Update trial state variable values $\bar{q}^{\mathrm{L}}_{i\tilde{h}\tilde{t}} = q^{\mathrm{L},*}_{i\tilde{h}\tilde{t}}$.

5. *Update overestimator*. If

$$\sum_{t\in\mathbb{T}}\sum_{h\in\mathbb{H}}(\beta)^t\left(\phi^*_{ht} - \zeta_{ht}(Q^*_{ht})\right) > \frac{\epsilon}{2},$$

construct overestimating cut (34) at $y = Q^*_{ht}$ and add additional break point index to $\mathbb{K}_{ht}$ for all hours $h \in \mathbb{H}$ and all years $t \in \mathbb{T}$.

6. *Update lower bound*. Calculate

$$\underline{\tilde{w}} = \sum_{t\in\mathbb{T}}\sum_{h\in\mathbb{H}}(\beta)^t\left(\phi^*_{ht} - \sum_{i\in\mathbb{I}} c_{it}q^*_{iht} - \sum_{i\in\mathbb{I}^{\mathrm{R}}}\hat{c}_{it}K_i\delta^*_{iht}\right),$$

and update lower bound $\widetilde{LB} = \max\{\widetilde{LB}, \underline{\tilde{w}}\}$.

7. *Check convergence*. If $\widetilde{UB} - \widetilde{LB} \le \mathrm{TOL}$, STOP (return λ^*, Q^*_{ht}, and $\widetilde{LB}$); otherwise, go to step 2.

Proposition 3. *The BD-DO in conjunction with the "Inner Nested Benders" above is correct and converges after finitely many iterations with an ϵ optimal solution to* $(Core{:}R)$.

Proof. We need to show that for the *Inner NBD* algorithm, the optimality cuts (38) derived from solving $(\mathrm{Sub{:}R{:}O})_{h't'}$ for $(h', t') \in \mathcal{A}_{\tilde{h}\tilde{t}}$ are valid and that finitely many such optimality cuts suffice to reach convergence. Both the correctness and the finiteness follow from Theorem 2 and NBD algorithm theory.

For the BD-DO algorithm, we need to argue that the optimality cuts (24) are valid and that finitely many optimality cuts suffice to reach convergence. Correctness and finiteness follow both from linear programming theory, by recognizing that the solution of the Inner NBD algorithm yields an optimal basis for $(\mathrm{Sub{:}R{:}O})$. □

Remark 3. Note that the cuts in the Inner NBD algorithm derived from the linearized subproblems $(\mathrm{Sub{:}R{:}O})_{ht}$ might not be tight in the earlier iterations. More importantly, the Inner Nested Benders algorithm must converge to guarantee the construction of a valid cut (24) with respect to the overall convergence tolerance ϵ.

3.2.4. Efficient Cut Calculation for (Core).

Since (Core) is a special case of (Core:T) with $|\mathbb{U}| = 1$, we defer from describing the detailed algorithm for (Core) in this section. Instead, we introduce the general idea of cut calculation and make a remark in Section 3.2.5 how to extract the algorithm for (Core). As shown in Section 3.2.1, (Sub) decomposes into one-hour subproblems for (Core). Each of these independent problems represents the market clearing condition described in Section 2.3 and visualized in Figures 1–3. Through sorting all generators by their marginal costs c_{it} in ascending order, the optimal primal solution values in a given hour can be obtained by calculating the intersection of the supply curve, i.e., the generators with their effective capacities and variable generator costs, and the demand curve. Additionally, once the optimal primal solution values are obtained, we can use Equation (19) to calculate the associated market clearing price and make the following observation. The dual solutions on each generator's capacity (i.e., constraints (2) and (3)) can be expressed by using the market clearing price:

$$\lambda_{iht} = \begin{cases}(\beta)^t(P^*_{ht} - c_{it}) & c_{it} < P^*_{ht},\\ 0 & \text{otherwise,}\end{cases} \quad \forall\, i \in \mathbb{I}, h \in \mathbb{H}, t \in \mathbb{T}. \tag{41}$$

Figure 1 visualizes the concept. All generators with a variable cost c_{it} greater than or equal to the market price have a shadow price of zero in that hour. Generators with a marginal cost lower than the market price have a positive dual variable, because they would increase the welfare. The dual variable is the discounted difference of market price and marginal cost. Thus, we can calculate all necessary solution values without solving a single LP for the subproblem. All algorithms presented in Sections 3.2.4–3.2.6 build on this very idea, and, for notational convenience, we do not distinguish between existing and new generators in the following sections. We substitute $K_i := \bar{\gamma}^N_i K_i \;\forall\, i \in \mathbb{I}^{\mathrm{N}}$ and $K_i := \bar{\gamma}^D_i K_i \;\forall\, i \in \mathbb{I}^{\mathrm{EX}}$.

3.2.5. Hybrid Cut Calculation for the Core Model with Transmission Constraints.

If transmission constraints (9)–(11) are added to (Sub:Core), the simple demand–supply intersection calculation of Section 3.2.4 used to obtain price and system load becomes significantly more challenging. Zones u are linked and each zone has its own market price and load. However, one can prove the following proposition if there are no transmission costs:

Proposition 4. *If transmission constraints* (10) *are nonbinding at optimality, the market prices in all zones are identical.*

If the market prices are not identical across all zones, it is favorable to transmit electricity from a zone with a lower market price to a zone with a higher market price until they are equal, given that there are no transmis-

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

sion costs. If there is sufficient capacity on the transmission lines, Proposition 4 holds true. In Section 4 we show that for our data set, transmission constraints are binding in 10% to 15% of the hours in the planning horizon. In this case, the market price, load, export, and import in each zone can be calculated as shown in Algorithm 1. (Generators are ordered by their marginal cost c_i in ascending order.)

Algorithm 1 (Find Equilibria Under Nonbinding Transmission Constraints)

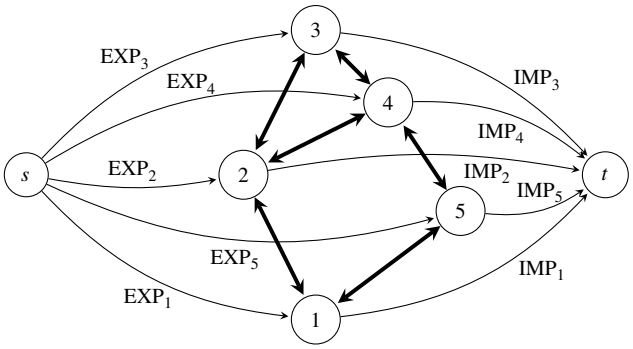
1: Given are demand functions $f_u \ \forall\, u \in \mathbb{U}$, the set of all generators $\mathbb{I}$, the generator capacities K_i, capacity utilization factors C_i, variable generator cost c_i, and generator to zone mapping $\mathcal{J}_u$.
2: $\bar{\mathcal{K}}_u = 0 \ \forall\, u \in \mathbb{U}$
3: **for** $i = 1 \rightarrow |\mathbb{I}|$ **do**
4: $\quad \bar{\mathcal{K}}_{\mathcal{J}_u} = \bar{\mathcal{K}}_{\mathcal{J}_u} + C_i K_i$
5: $\quad Q_u = f_u(c_i) \ \forall\, u \in \mathbb{U}$
6: $\quad \Sigma_Q = \sum_{u \in \mathbb{U}} Q_u$, $\Sigma_{\bar{\mathcal{K}}} = \sum_{u \in \mathbb{U}} \bar{\mathcal{K}}_u$
7: $\quad q_i = C_i K_i$
8: $\quad$ **if** $\Sigma_Q < \Sigma_{\bar{\mathcal{K}}}$ **then**
9: $\quad\quad$ **if** $\Sigma_{\bar{\mathcal{K}}} - C_i K_i > \Sigma_Q$ **then**
10: $\quad\quad\quad$ Solve system (42) to obtain P^*_u and $Q^*_u \ \forall\, u \in \mathbb{U}$
11: $\quad\quad\quad q_i = 0$
12: $\quad\quad$ **else**
13: $\quad\quad\quad P^*_u = c_i$, $Q^*_u = Q_u$, $\forall\, u \in \mathbb{U}$
14: $\quad\quad\quad q_i = \bar{\mathcal{K}}_{\mathcal{J}_u} - Q^*_{\mathcal{J}_u}$
15: $\quad\quad$ **end if**
16: $\quad\quad$ $\mathrm{EXP}^*_u = \max(\bar{\mathcal{K}}_u - Q^*_u, 0)$
17: $\quad\quad$ $\mathrm{IMP}^*_u = \max(Q^*_u - \bar{\mathcal{K}}_u, 0)$
18: $\quad\quad$ **return** load Q^*_u, generator levels q^*_i, import IMP^*_u, and export EXP^*_u for each zone u
19: $\quad$ **end if**
20: **end for**

We first assume that the transmission constraints are nonbinding and check later if this is indeed satisfied. With this assumption, the market prices P^*_u for all zones are identical; i.e., $P^*_u = p^* \ \forall\, u \in \mathbb{U}$. Therefore, transmission can be ignored, and we iterate over all generators in ascending order to find the generator for which the total supply in the system up to its marginal cost c_i is higher than the total demand at each zone u under $p^* = c_i$. Then, there are two cases: (i) either the market price is identical with c_i, or (ii) i is not running and we have the case shown in Figure 2. If the latter is true, a system with $|\mathbb{U}| + 1$ unknowns and $|\mathbb{U}| + 1$ equations can be solved to obtain p^* and Q^*_u:

$$\begin{aligned} p^* &= P_1(Q^*_1), \\ p^* &= P_2(Q^*_2), \\ &\vdots \\ p^* &= P_{|\mathbb{U}|}(Q^*_{|\mathbb{U}|}), \\ \Sigma_{\bar{\mathcal{K}}} - C_{i^*} K_{i^*} &= Q^*_1 + Q^*_2 + \cdots + Q^*_{|\mathbb{U}|}, \end{aligned} \tag{42}$$

where i^* is the cheapest generator not in the dispatch.

Figure 4. Maximum Flow For a System with Five Buses



The special case exists that all generators in a given zone u' are transmitting their electricity to the other zones; i.e., $Q^*_{u'} = 0$ and $P^*_{u'} = P_{u'}(Q^*_{u'})$. The column and row associated with u' are removed from system (42). We omitted this case in Algorithm 2, shown above for convenience purposes.

Export and import quantities for each bus can be obtained by comparing the supply and demand in each zone at equilibrium. At this point, we can solve a maximum flow problem to check if the assumption that all transmission capacities are nonbinding is indeed satisfied. Figure 4 displays the corresponding directed maximum flow problem for a system with five zones. Source node s and target node t are artificial nodes and connected to all zones in the system via artificial arcs. The capacities on these arcs correspond to the import and export values obtained for each zone using Algorithm 2. The bold bidirectional arcs represent the original system's transmission lines between zones.

Remark 4. Note that Algorithm 1 directly reduces to an algorithm for (Core) if $|\mathbb{U}| = 1$. Instead of solving system (42) on line 10, one can directly assign $Q^* = \Sigma_{\bar{\mathcal{K}}} - C_i K_i$ and calculate P^* through (19).

The maximum flow algorithm selected depends on the type of transmission constraints. In the case of constraints (9) and (10), the Edmonds–Karp algorithm can be used. If the resulting maximum flow is equal to the sum of exports across all zones, all transmission constraints are nonbinding and we can use the hourly market price obtained in Algorithm 1 to calculate the dual variables as in (41). If this is not the case, the demand in each node becomes a function of the supply curves in the other zones and standard network algorithms cannot be used. Thus, we propose to solve the corresponding one-hour subproblem to obtain the dual variables.

In the GB-DO decomposition approach described in Sections 3.1 and 3.2.2, step 2 is modified as follows.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

2.1. *Run Algorithm* 1. Run Algorithm 1 for $\bar{\gamma}$ and obtain equilibrium quantities Q^*_{uht}, equilibrium prices P^*_{uht}, and import/export quantities $\mathrm{IMP}^*_{uht}/\mathrm{EXP}^*_{uht}$.

2.2. *Solve maximum flow*. Solve the corresponding maximum flow problem shown in Figure 4 for all hours $h \in \mathbb{H}$ and all years $t \in \mathbb{T}$. Denote the maximum flow values by f^{flow}_{ht}.

2.3 *Solve subproblem*. For all hours $h \in \mathbb{H}$ and all years $t \in \mathbb{T}$,

If $f^{\mathrm{flow}}_{ht} = \sum_{u \in \mathbb{U}} \mathrm{EXP}_{uht}$, calculate dual variables λ^*_{iht} as described in Section 3.2.4 and define $\phi^*_{uht} = \zeta_{uht}(Q^*_{uht})$.

Else solve subproblem $(\mathrm{Sub{:}T{:}O})_{ht}$ for trial $\bar{\gamma}$ to obtain dual variables λ^*_{iht}, equilibrium quantities Q^*_{uht}, and overestimator value ϕ^*_{uht}.

Step 3 in BD-DO must only be carried out for hours in which the dual variables cannot be calculated. In the unlikely case that the transmission constraints are never binding, overestimating the NLP subproblem becomes entirely obsolete. In fact, we could calculate the optimal solution for (Sub:T), and thus the resulting Benders optimality cut (24) would be tight. This is always the case for $|\mathbb{U}| = 1$, i.e., for (Core) in Section 3.2.4, and it follows that the lower bound in step 4 can be calculated by using $W^*(\bar{\gamma})$; the same applies to step 5 when constructing the Benders optimality cut.

The following interpretation is valid: for hours $h \in \mathbb{H}$ and years $t \in \mathbb{T}$ when the transmission constraints are nonbinding, we calculate the dual variables of the linearized NLPs with break points at, and sufficiently close to, the optimal quantity Q^*_{uht}. (GBD theory also provides the correctness of this cut calculation algorithm.) For all other hours and years, we solve the corresponding linearized subproblem, as in BD-DO. This guarantees the convergence of the resulting decomposition algorithm.

If we were able to calculate the primal solution in the case of binding transmission constraints, the dual variables for constraints (2) and (3) are the discounted difference of a generator's marginal cost and the zonal market price it is located in. The dual variables on constraints (10) are the discounted difference in zonal market prices.

3.2.6. Hybrid Cut Calculation for the Core Model with Start-up Restrictions. In the previous two sections, the dual variables on constraints (2) and (3) could be obtained by calculating the market price and load in each hour independently of other hours. If start-up constraints are present, the one-hour subproblems are not independent anymore and the equilibrium price and quantity may depend on previous and future hours. We use the NBD framework presented in Section 3.2.3 to decompose $(\mathrm{Sub{:}R{:}O})$ into $|\mathbb{H}| \cdot |\mathbb{T}|$ problems of form $(\mathrm{Sub{:}R{:}O})_{ht}$. The link between a given hour and its previous hour is state variables $\bar{q}^{\mathrm{L}}$, whereas the link to the next hour is the Benders optimality cuts (38). This decomposition allows us to treat each hour individually.

In the following, the core idea of cut calculation with start-up constraints is drafted. We do not present computational results for this algorithm because of reasons discussed below and in Section 4. The main challenge is that dual variables on constraints (37) cannot be calculated for all one-hour subproblems, and we therefore obtain a hybrid algorithm in which dual variables are calculated if possible and the corresponding NLP is solved if not.

The idea of the dual variable calculation is to split each generator into multiple generators based on whether the generator (1) has start-up restriction, (2) was running in the previous hour, and (3) is affected by a cut coefficient.

We define splitting a generator as replacing a generator with two or more new generators. Their capacities add to the replaced generator's capacity, but the new generators may have different marginal costs.

These criteria result in the following five cases for a given hour h and year t, and previous hour $h't'$.

(a) If $[i \notin \mathbb{I}^{\mathrm{R}}]$ or $[\bar{q}^{\mathrm{L}}_{ih't'} = 1$ and $\forall\, j \in \mathbb{J}_{ht}\colon \tilde{\omega}^{\mathrm{s}}_{ijht} = 0]$, do not split generator i:

$$c^1_{it} := c_{it}, \quad (C_{ih}K_i)^1 := C_{ih}K_i.$$

(b) If $[i \in \mathbb{I}^{\mathrm{R}}]$ and $[\bar{q}^{\mathrm{L}}_{ih't'} = 1]$ and $[\exists j \in \mathbb{J}_{ht}\colon \tilde{\omega}^{\mathrm{s}}_{ijht} > 0]$, split generator i in two generators:

$$c^1_{it} := c_{it} - \frac{\tilde{\omega}^{\mathrm{s}}_{ijht}}{P^{\min}_i \cdot \beta_t C_{ih}K_i}, \quad (C_{ih}K_i)^1 := P^{\min}_i \cdot C_{ih}K_i;$$
$$c^2_{it} := c_{it}, \quad (C_{ih}K_i)^2 := (P^{\max}_i - P^{\min}_i) \cdot C_{ih}K_i.$$

(c) If $[i \in \mathbb{I}^{\mathrm{R}}]$ and $[\bar{q}^{\mathrm{L}}_{ih't'} < 1]$ and $[\forall\, j \in \mathbb{J}_{ht}\colon \tilde{\omega}^{\mathrm{s}}_{ijht} = 0]$, split generator i in two generators:

$$c^1_{it} := c_{it}, \quad (C_{ih}K_i)^1 := \bar{q}^{\mathrm{L}}_{ih't'} C_{ih}K_i;$$
$$c^2_{it} := c_{it} + \frac{\hat{c}_{it}}{C_{ih}}, \quad (C_{ih}K_i)^2 := (1 - \bar{q}^{\mathrm{L}}_{ih't'}) C_{ih}K_i.$$

(d) If $[i \in \mathbb{I}^{\mathrm{R}}]$ and $[\bar{q}^{\mathrm{L}}_{ih't'} < 1]$ and $[\exists j \in \mathbb{J}_{ht}\colon \tilde{\omega}^{\mathrm{s}}_{ijht} = 0]$ and $[\hat{c}_{it}K_i \geq \tilde{\omega}^{\mathrm{s}}_{ijht}]$, split generator i in three generators:

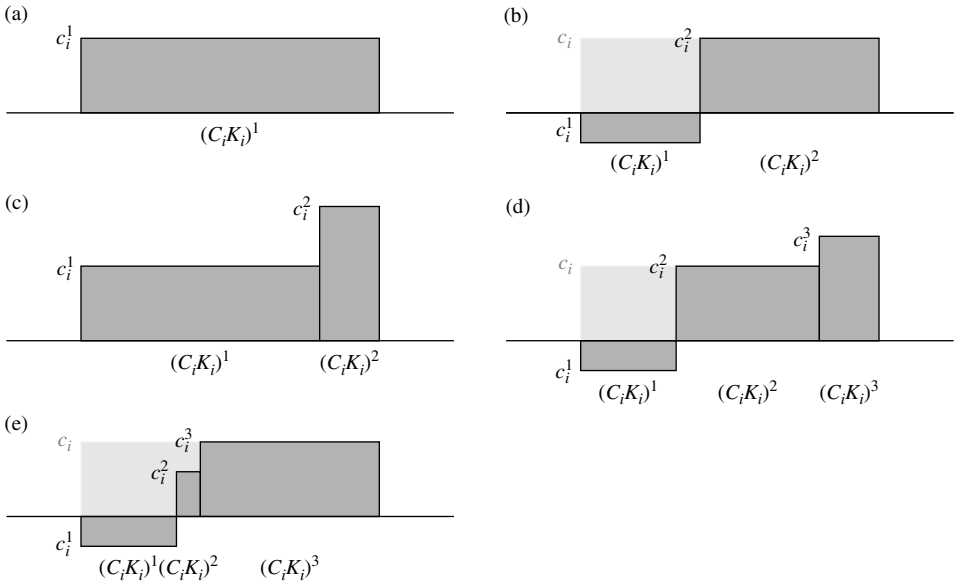
$$c^1_{it} := c_{it} - \frac{\tilde{\omega}^{\mathrm{s}}_{ijht}}{P^{\min}_i \cdot \beta_t C_{ih}K_i}, \quad (C_{ih}K_i)^1 := P^{\min}_i \cdot \bar{q}^{\mathrm{L}}_{ih't'} C_{ih}K_i;$$
$$c^2_{it} := \frac{\hat{c}_{it}}{C_{ih}} - \frac{\tilde{\omega}^{\mathrm{s}}_{ijht}}{\beta_t C_{ih}K_i}, \quad (C_{ih}K_i)^2 := (P^{\max}_i - P^{\min}_i) \cdot \bar{q}^{\mathrm{L}}_{ih't'} C_{ih}K_i;$$
$$c^3_{it} := c_{it}, \quad (C_{ih}K_i)^3 := (1 - \bar{q}^{\mathrm{L}}_{ih't'}) C_{ih}K_i.$$

(e) If $[i \in \mathbb{I}^{\mathrm{R}}]$ and $[q^{\mathrm{L}}_{ih't'} < 1]$ and $[\exists j \in \mathbb{J}_{ht}\colon \tilde{\omega}^{\mathrm{s}}_{ijht} = 0]$ and $[\hat{c}_{it}K_i < \tilde{\omega}^{\mathrm{s}}_{ijht}]$, split generator i in three generators:

$$c^1_{it} := c_{it} - \frac{\tilde{\omega}^{\mathrm{s}}_{ijht}}{P^{\min}_i \cdot \beta_t C_{ih}K_i}, \quad (C_{ih}K_i)^1 := P^{\min}_i \cdot \bar{q}^{\mathrm{L}}_{ih't'} C_{ih}K_i;$$

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Figure 5. Five Cases for Splitting Generators



$$c_{it}^2 := c_{it} + \frac{\hat{c}_{it}}{P_i^{\min} \cdot C_{ih}} - \frac{\tilde{\omega}_{ijht}^{s}}{P_i^{\min} \cdot \beta_t C_{ih} K_i},$$
$$(C_{ih}K_i)^2 := P_i^{\min} \cdot (1 - \bar{q}_{ih't'}^{L}) C_{ih} K_i;$$
$$c_{it}^3 := c_{it}, \quad (C_{ih}K_i)^3 := (P_i^{\max} - P_i^{\min}) \cdot C_{ih} K_i.$$

These five cases are visualized in Figure 5. If state variable $\bar{q}_{ih't'}^{L}$ is equal to one or if the generator has no start-up restriction, we can use it at full capacity with its regular costs (case (a)). However, if generator i is affected by a positive cut coefficient, we receive an additional benefit. This can be expressed as a decreased generator cost (case (b)). If state variable $\bar{q}_{ih't'}^{L}$ is less than one, there are three possible scenarios. If generator i is not affected by a cut coefficient, it can be used up to $q_{ih't'}^{L} C_{ih} K_i$ at its regular cost c_{it} because of the linearized start-up constraints. Generation above this level is penalized with start-up costs $\hat{c}_{it}$ (case (c)). If the generator is affected by a cut, the question is whether the additional benefit of the cut is higher than the generator's start-up costs. If this is the case, we might consider to run the generator at $q_{iht}^{L} = 1$ to be awarded this difference (case (e)). Case (d) is the complementary case.

Because multiple Benders cuts with nonzero coefficients can occur after the first iteration of the NBD, the splitting scheme for cases (b), (d), and (e) only works in the first iteration. The condition $[\forall\, j \in \mathbb{J}_{ht} \colon \tilde{\omega}_{ijht}^{s} = 0]$ therefore determines if the dual variables on constraints (37) can be calculated or if an NLP has to be solved to obtain them.

In case condition $[\forall\, j \in \mathbb{J}_{ht} \colon \tilde{\omega}_{ijht}^{s} = 0]$ is true or we are in the first iteration ($j = 1$) of the NBD, the same calculation as in Section 3.2.4 can be used to clear the market and obtain equilibrium quantity Q_{ht}, equilibrium price P_{ht}, and split generators' levels $\tilde{q}_{iht}$. The dual variables on constraints (37) to construct Benders optimality cut (38) during the backward step of the decomposition algorithm can then be calculated as

$$\tilde{\omega}_{ijh't'}^{s} = \beta_t \cdot \max(P_{ht} - c_{it}, 0) \cdot C_{ih} K_i + \tilde{\omega}_{ijht}^{s},$$

where $(h', t') \in \mathcal{A}_{ht}$ describes the previous hour.

Calculating the cut constant $\tilde{\omega}_{jh't'}^{c}$ (see (40)), state variables $\bar{q}_{iht}^{L}$, and objective function value $W_{ht}^{*}(\bar{\gamma}, \bar{q}_{ih't'}^{L})$ for each hour in the nested Benders decomposition is achieved by mapping the splitted generators' levels $\tilde{q}_{iht}$ back to their original counterparts q_{iht}. We omit a detailed description at this point.

4. Computational Results

We compare the performance of the algorithms from Section 3 with different settings to a monolithic approach, i.e., solving (PGEP)'s special cases as one model, and a state-of-the-art GBD implementation (Guan and Philpott 2011). The monolithic approach requires a considerable amount of memory, and only small instances can be solved within our prescribed time limit of 48 hours. Additionally, we turn all investment decisions into continuous variables; decommission decisions remain binary to avoid partial retirements. This simplification has several reasons. First, it makes the monolithic approach more tractable and we are able to solve larger instances within the time constraints. Second, it allows us to focus our analysis in this section on the efficient solution of the large subproblem. Our results show that the master problem's share of solution time becomes negligible. Third, and this goes back to Section 1, the purpose of (PGEP) is to show investment trends in the presence of electricity policies. The policy analysis we built the model for does not demand binary investment decisions.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

4.1. Case Study Description and Model Instances

We model the ERCOT power system and represent the zonal wholesale market in Texas using data from 2008. The power market in Texas can be seen as generally deregulated (although a few regions are still regulated, e.g., the cities Austin and San Antonio) and fairly isolated from other states. We model 380 existing generators in four load zones (Houston, North, South, West). The total capacity in the system is approximately 82 GW, excluding generators that are used for cogeneration, not connected to the grid, or retired.

Investments into eight different technologies (or aggregated generators) are possible: coal, integrated gasification combined cycle (IGCC), natural gas combined cycle (NGCC), natural gas combustion turbine (NGCT), nuclear, solar, and two wind types. All technologies but wind can be built in each of the four zones (wind can only be built in South, North, and West). We choose capacity K_i and K^{T}_{uv} sufficiently large in order for constraints (13) and (14) to be nonbinding.

The purpose of this section is not to carry out a detailed case study but to present computational results of the algorithms from Section 3. We therefore omit a more detailed description of the data and refer the reader to Fell and Linn (2013) for the underlying assumptions; the authors restrict their analysis to the case of a linear demand function. To demonstrate our dynamic overestimation approach, we also fit a partial-log demand function (see Table 1 and Bushnell 2010) using the same 2008 price and load data and elasticity assumption as in Fell and Linn (2013). We model scheduled and forced outage percentages of plants by reducing their capacity evenly across one year. This is a standard approach because no information is available when a downtime occurs in the (distant) future. Additionally, we use start-up costs of \$59/MW for coal generators (Kumar et al. 2012), which reflect the average hot start costs of a generator. We use \$200/MW as a proxy for nuclear generators, which essentially keeps them from ramping at all. Unfortunately, we do not have access to the transmission network data. However, ERCOT provides data for commercially significant constraints (CSC) between the four zones, which represent both a physical and commercial restriction on transmission in the system. Any transmission between zones that exceeds the CSC is charged with a cost for the corresponding power producer. We average the publicly available 2008 CSC data to obtain annual transmission constraints between the four load zones. We assume \$950 per MW and mile for transmission line extensions and use the distance between the following cities to calculate the cost F^{T}_{uvt}: Dallas (North), Houston (Houston), San Antonio (South), and Odessa (West).

The following instances of (PGEP)'s special cases are then constructed. Table 2 shows the number of hours and years modeled, and the size of the resulting instance if we solve the instance as a monolith. The investment costs for new generators and transmission lines are scaled proportionally to make investment in all instances profitable. Furthermore, the annual operation and maintenance costs are scaled to avoid decommissioning for the smaller instances. Otherwise, the algorithms of Section 3 would converge in the first iteration. That being said, it proved difficult to generate instances that had homogeneous convergence performance and the number of iterations varies significantly.

All data and GAMS files for the monolithic models are provided in the supplemental material.

4.2. Results

We compare our approaches from Section 3 for problems (Core), (Core:T), and (Core:R) to a monolithic approach and a classical GBD; i.e., the subproblem is solved as a (series of) NLPs. Table 3 presents all solution configurations tested. All algorithms are implemented in GAMS 24.3.3. Mixed-integer quadratic programming (MIQP) problems are solved with

Table 2. Instance Characteristics: Number of Rows (R), Number of Columns (C), Number of Nonzeroes (NZ), and Integer Variables (Int)

			(Core)			(Core:T)			(Core:R)		
No.	Hours	Int	R	C	NZ	R	C	NZ	R	C	NZ
1	100	389	0.04	0.04	0.16	0.04	0.04	0.17	0.04	0.04	0.16
2	200	389	0.08	0.08	0.31	0.09	0.09	0.35	0.10	0.10	0.37
3	400	389	0.16	0.16	0.63	0.17	0.17	0.69	0.20	0.20	0.74
4	1,000	389	0.39	0.39	1.57	0.43	0.43	1.72	0.49	0.49	1.84
5	2,000	389	0.78	0.78	3.13	0.86	0.86	3.44	0.99	0.99	3.68
6	4,000	389	1.57	1.57	6.26	1.73	1.73	6.87	1.98	1.98	7.35
7	6,000	389	2.35	2.35	9.39	2.59	2.59	10.31	2.96	2.96	11.03
8	8,760	389	3.43	3.43	13.70	3.79	3.79	15.05	4.33	4.33	16.10
9	70,080	3,112	27.47	27.47	112.07	30.28	30.28	131.20	34.62	34.62	131.27
10	140,160	6,224	54.95	54.95	229.74	60.56	60.56	287.07	69.25	69.25	268.15
11	201,480	8,947	78.99	78.99	337.30	87.05	87.05	443.69	99.54	99.54	392.52

Note. All numbers in R, C, and NZ are in millions.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Table 3. Solution Approaches Compared

		Linear		Partial-log	
Model	Method	Sequential	Parallel	Sequential	Parallel
(Core)	Monolith	MIQP	MIQP	MINLP	MINLP
	GBD	MILP + QP	MILP + QP	MILP + NLP	MILP + NLP
	Algorithm 3.2.1	MILP + LP	MILP + LP	MILP + LP	MILP + LP
	Algorithm 3.2.4	MILP[a]	MILP[a]	MILP[a]	MILP[a]
(Core:T)	Monolith	—	—	MINLP	MINLP
	Algorithm 3.2.2	—	—	MILP + LP	MILP + LP
	Algorithm 3.2.5	—	—	MILP + LP[a]	MILP + LP[a]
(Core:R)	Monolith	—	—	MINLP	—
	Algorithm 3.2.3	—	—	MILP + LP	—
	Algorithm 3.2.6	—	—	—	—

Note. Cells show model classes for the monolith and Master + Subproblem for the algorithms.
[a]Dual variables in the subproblem are explicitly calculated.

Gurobi 5.6.2 and BONMINH 1.7, and the fastest solution time among the two is reported. MINLPs are solved with BONMINH and its outer-approximation-based branch-and-cut algorithm. We experimented with all other commercial MINLP solvers in GAMS, but they were significantly slower than BONMINH. MILPs, LPs, and QPs are solved with Gurobi; NLPs in the GBD are solved with Conopt 3.16C. We use default settings for all solvers. The runs are carried out on a Windows 7 64 bit machine with an Intel Core i7 CPU at 2.93 GHz and 12 GB memory. CPLEX was also tested with default settings but had numerical issues both in the monolithic approaches as well as algorithms and was therefore dropped. In general, the models show numerical instability due to the large integral terms and small generator costs in the objective function. Gurobi is able to handle these issues the best with default settings, although we still observe varying iteration numbers for different numbers of threads used.

Tables 4 and 5 present the results for (Core) on one and four threads, respectively. As we expected, the LP approximation of the subproblems significantly outperforms the NLP approach. It must be noted that we use parts of the market clearing algorithms presented in Sections 3.2.4 and 3.2.5 (assuming transmission constraints are nonbinding) to obtain the equilibrium quantities in advance and improve the linear overestimator before solving the subproblem for a specific $\bar{\gamma}$. What may be still surprising is that the iteration count of the GBD approach does not differ from the LP approach on average. We argue that this is due to the superior numerical stability of the LP solver. In fact, several of the larger instances were not solved to the correct objective function value using GBD. Instance 11 does not converge at all because of an invalid cut in iteration 19.

The LP-based algorithms significantly outperform the monolithic approaches, especially for the larger instances. The cut calculation approaches are faster than the LP-based approaches by a factor of 10–20 on average and by up to a factor of 28. Furthermore, they are unaffected by the numerical instability. The average time per iteration when using a linear demand function compared to a partial-log demand function in the LP-based approaches is not significantly different, demonstrating the viability of the overestimation approach. The parallelization factor for all algorithms and instances is close to three on average. The monolithic approach does not parallelize well in most instances, although exceptions exist.

The algorithms for (Core:T) perform similarly to their (Core) counterparts; see Table 6. Note the increased average iteration count due to an increased number of investment decisions. The hybrid cut calculation approach discussed in Section 3.2.5 reduces the solution time by up to 75%, and approximately 10% of the hours remain to be solved as an LP. The parallelization factor is generally lower than for (Core), which is caused by an increased overhead in communication between GAMS and C# as well as GAMS's model updating facility.

The algorithm for (Core:R) is able to solve all but the largest instance within 48 hours; see Table 7. Instance 11 is in iteration 6 with a gap of $5.4 \cdot 10^{-2}$. The algorithm's performance strongly depends on the number of subsequent hours batched together, because fewer NB iterations are necessary. This is due to fewer nested Benders cuts and fewer overestimator updates. However, a larger model batch takes longer to update and, more importantly, longer to solve than the equivalent number of small models. The cut calculation approach in Section 3.2.6 describes an extreme case in this regard. Because we can only calculate all necessary primal and dual variable values if we consider one hour at a time, we receive the combined effect of many nested Benders cuts and frequent overestimator updates. In addition, the nested Benders cuts are not tight until the overestimator is sufficiently refined. Furthermore, the calculation of variable values is mostly

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Table 4. Sequential Results for (Core): Solution Time (Solve), Number of Iterations Until Convergence (It), Time Spent Solving the Subproblem (Sub)

	Linear demand function											Partial-log demand function									
	Monolith		GBD			Algorithm 3.2.1			Algorithm 3.2.4			Monolith	GBD			Algorithm 3.2.1			Algorithm 3.2.4		
	Gurobi	BONMINH										BONMINH									
No.	Solve	Solve	Solve	It	Sub	Solve	It	Sub	Solve	It	Sub	Solve	Solve	It	Sub	Solve	It	Sub	Solve	It	Sub
1	5.2	17.7	7.5	18	6.6	8.8	19	6.3	0.7	20	0.1	41.2	128.4	29	126.7	17.1	30	10.0	7.0	27	0.1
2	64.8	50.7	11.4	19	10.8	9.8	18	7.2	0.9	22	0.1	157.4	228.3	31	226.4	15.9	30	11.3	2.5	30	0.2
3	31.5	113.7	20.5	15	20.0	10.6	15	8.2	0.6	15	0.2	8,714.4	492.6	43	488.1	28.1	37	20.4	6.5	38	0.5
4	1,820.9	367.4	39.1	12	38.4	14.3	12	11.7	0.7	12	0.3	1,315.6	1,006.7	33	1,003.9	46.4	37	37.2	3.7	34	1.1
5	2,841.6	970.0	124.9	19	123.3	39.6	19	33.3	1.6	19	1.0	5,994.4	2,054.5	40	2,049.4	81.1	37	67.4	6.1	40	2.7
6	5,659.3	3,210.5	137.2	10	135.8	37.8	10	32.3	1.5	10	1.1	8,157.2	1,577.8	10	1,576.1	38.3	10	32.5	2.4	10	1.5
7	29,930.2	[b]	210.0	11	207.9	62.5	11	53.9	2.4	11	1.9	[b]	4,320.0	17	4,316.3	95.1	17	81.4	4.9	16	3.7
8	157,506.0	[b]	286.2	10	283.4	80.4	10	69.1	3.0	10	2.5	[b]	3,891.8[c]	11	3,888.5	79.0	10	67.8	4.0	10	3.1
9	[a]	[a]	2,450.8	11	2,446.2	644.3	11	619.6	24.7	11	19.9	[a]	49,149.2	22	49,134.9	1,093.8	19	1,045.2	72.6	20	54.3
10	[a]	[a]	4,535.9	10	4,529.7	1,191.7	10	1154.8	49.6	10	39.8	[a]	52,201.2[c]	9	52,195.5	1,258.0	11	1,214.3	82.0	11	61.4
11	[a]	[a]	6,867.5	11	6,859.1	1,866.9	11	1,816.7	71.4	11	57.3	[a]	[b]	19	—	9,598.8	54	9,228.4	640.8	54	437.6

Notes. All times are in seconds. Convergence tolerance for all algorithms and relative gap of monolithic models is 10^{-6}.
[a]Memory limit of 12 GB reached.
[b]Time limit of 48 hours reached.
[c]Algorithm converged with wrong objective function value because of numerical instability in nonlinear subproblems.

Table 5. Parallel Results for (Core) with Four Threads

	Linear demand function											Partial-log demand function									
	Monolith		GBD			Algorithm 3.2.1			Algorithm 3.2.4			Monolith	GBD			Algorithm 3.2.1			Algorithm 3.2.4		
	Gurobi	BONMINH										BONMINH									
No.	Solve	Solve	Solve	It	Sub	Solve	It	Sub	Solve	It	Sub	Solve	Solve	It	Sub	Solve	It	Sub	Solve	It	Sub
1	3.9	21.3	7.8	17	7.2	12.9	19	8.9	0.7	17	0.0	48.4	30.8	29	28.1	20.2	30	12.8	4.1	26	0.1
2	11.3	54.4	8.9	19	8.1	10.5	18	7.7	1.0	22	0.1	227.1	87.4	31	85.3	18.3	31	13.2	2.9	30	0.1
3	23.5	118.2	12.8	15	12.2	10.0	15	7.6	0.6	15	0.1	8,901.7	451.4	43	447.0	25.3	37	17.4	6.5	42	0.3
4	1,071.7	386.1	19.7	12	19.0	13.0	12	10.4	0.5	12	0.1	1,339.8	1,686.2	33	1,683.2	40.4	37	31.2	3.3	31	0.4
5	2,056.9	993.0	54.2	19	52.6	28.0	19	21.8	1.1	19	0.4	6,170.6	1,321.4	40	1,316.0	57.1	40	42.9	5.3	38	1.1
6	14,787.7	3,156.6	54.5	10	52.9	21.7	10	16.4	0.8	10	0.5	8,133.8	868.2	10	866.5	21.9	10	16.5	1.6	10	0.7
7	28,766.0	[b]	88.7	11	86.4	32.4	11	24.2	1.2	11	0.8	[b]	2,254.7	17	2,250.7	48.7	17	36.1	2.9	16	1.6
8	166,382.1	[b]	118.1	10	115.1	41.2	10	31.0	1.5	10	1.1	[b]	1,992.6[c]	11	1,989.2	39.0	10	29.2	2.2	10	1.3
9	[a]	[a]	1,025.0	11	1,019.8	236.0	11	220.0	12.1	11	8.3	[a]	23,524.2	22	23,509.6	421.7	19	388.8	37.8	20	19.7
10	[a]	[a]	1,818.3	10	1,812.1	424.8	10	404.1	24.4	10	16.7	[a]	24,005.9[c]	9	24,000.3	467.2	11	443.8	39.4	11	22.7
11	[a]	[a]	2,903.4	11	2,894.3	674.4	11	647.5	34.2	11	23.3	[a]	[b]	19	—	3,769.9	54	3,527.0	339.2	54	149.9

Note. See Table 4 for column descriptions.
[a]Memory limit of 12 GB reached.
[b]Time limit of 48 hours reached.
[c]Algorithm converged with wrong objective function value because of numerical instability in nonlinear subproblems.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Table 6. Results for (Core:T) with One and Four Threads: Percentage of Hours That Had to Be Solved in Hybrid Algorithm 3.2.5 [% LP]

	Sequential								Parallel							
	Monolith	Algorithm 3.2.2			Algorithm 3.2.5				Monolith	Algorithm 3.2.2			Algorithm 3.2.5			
No.	Solve	Solve	It	Sub	Solve	It	Sub	% LP	Solve	Solve	It	Sub	Solve	It	Sub	% LP
1	87.3	36.4	41	32.0	25.2	38	21.1	12.5	224.8	30.0	41	25.4	29.7	40	25.4	12.3
2	527.6	61.1	51	54.0	41.2	53	33.8	11.8	1,114.5	43.7	47	37.3	48.7	53	41.7	11.8
3	952.0	105.8	60	96.9	58.2	60	49.3	11.6	5,709.4	80.4	58	71.6	65.0	64	54.4	11.7
4	1,884.9	152.8	49	147.1	60.3	46	55.0	9.0	1,600.9	87.0	46	81.2	45.3	41	40.2	8.9
5	5,802.1	289.5	51	282.6	98.8	53	91.5	8.0	4,848.2	158.7	54	150.6	80.5	56	73.0	7.8
6	23,167.1	175.5	20	173.6	51.0	20	49.2	7.1	17,996.0	82.2	21	79.8	36.6	21	34.6	7.1
7	[b]	367.9	25	365.4	110.7	29	107.9	7.8	[b]	142.4	25	139.6	68.2	29	65.4	7.8
8	[b]	404.6	19	402.8	88.1	19	86.6	7.6	[b]	143.0	19	140.8	54.8	19	52.9	7.6
9	[a]	6,183.0	33	6,159.0	1,235.9	33	1,211.7	10.2	[a]	2,359.9	33	2,335.1	486.0	33	462.0	10.2
10	[a]	7,276.9	21	7,262.6	1,626.0	21	1,612.8	11.7	[a]	2,739.5	21	2,724.5	601.3	21	588.0	11.7
11	[a]	71,070.2	97	70,398.6	16,412.1	97	15,788.7	14.5	[a]	28,323.9	97	27,638.2	6,613.4	97	5,986.7	14.5

Note. See Table 4 for other column descriptions.
[a]Memory limit of 12 GB reached.
[b]Time limit of 48 hours reached.

Table 7. Results for (Core:R): Total Number of Nested Benders Iterations (NB It) and Total Time Spent in Nested Benders Decomposition (NB time)

	Monolith	Algorithm 3.2.3			
No.	Solve	Solve	It	NB It	NB time
1	45.5	59.5	35	230	54.8
2	772.6	136.7	35	282	133.2
3	14,260.3	363.1	48	382	358.1
4	1,507.8	816.5	39	312	812.9
5	4,779.4	6,388.1	39	420	6,383.8
6	27,017.9	2,321.1	10	56	2,319.8
7	43,241.8	1,667.3	16	109	1,664.8
8	[b]	4,970.4	10	113	4,968.4
9	[a]	86,612.1	18	96	86,574.7
10	[a]	83,012.1	11	60	82,974.3
11	[a]	[b]	—	—	

Notes. The nested Benders convergence tolerance is set to 10^{-7}. See Table 4 for other column descriptions.
[a]Memory limit of 12 GB reached.
[b]Time limit of 48 hours reached.

Table 8. Best Algorithms for All Special Cases: Fastest Reported Monolith Across All Solvers and Number of Threads (Solve) and Fastest Reported Algorithm Compared to the Fastest Monolith (Factor)

	(Core)					(Core:T)		(Core:R)	
	Linear		Partial-log						
No.	Solver	Solve	Factor	Solve	Factor	Solve	Factor	Solve	Factor
1	Gurobi	3.9(4)	5.7	41.2(1)	10.0	87.3(1)	3.5	45.5(1)	0.8
2	Gurobi	11.3(4)	12.4	157.4(1)	63.8	527.6(1)	12.8	772.6(1)	5.7
3	Gurobi	23.5(4)	40.9	8,714.4(1)	1,341.3	952.0(1)	16.4	14,260.3(1)	39.3
4	BONMINH	367.4(1)	713.4	1,315.6(1)	401.0	1,600.9(4)	35.3	1,507.8(1)	1.8
5	BONMINH	970.0(1)	909.9	5,994.4(1)	1,137.2	4,848.2(4)	60.2	4,779.4(1)	4.8
6	BONMINH	3,156.6(4)	3,740.0	8,133.8(4)	5,055.2	17,996.0(4)	492.2	27,017.9(1)	27.0
7	Gurobi	28,766.0(4)	23,179.7	[a]	—	[a]	—	43,241.8(1)	43.2
8	Gurobi	157,506.0(1)	107,074.1	[a]	—	[a]	—	[a]	—

Notes. All times are in seconds. The number of threads appears in parentheses under Solve.
[a]Memory or time limit reached.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Table 9. Best Algorithms for (Core) with Linear and Partial-Log Demand Functions: Solve Time in Seconds for GBD (Solve) and Factors Compared to GBD for the Other Columns (Factor)

	Linear						Partial-log					
	Sequential			Parallel			Sequential			Parallel		
	GBD	Alg. 3.2.1	Alg. 3.2.4	GBD	Alg. 3.2.1	Alg. 3.2.4	GBD	Alg. 3.2.1	Alg. 3.2.4	GBD	Alg. 3.2.1	Alg. 3.2.4
No.	Solve	Factor	Factor	Solve	Factor	Factor	Solve	Factor	Factor	Solve	Factor	Factor
1	7.5	0.9	10.2	7.8	0.6	11.6	128.4	7.5	18.3	30.8	1.5	7.5
2	11.4	1.2	12.5	8.9	0.9	8.7	228.3	14.3	92.5	87.4	4.8	29.8
3	20.5	1.9	33.7	12.8	1.3	22.3	492.6	17.5	75.6	451.4	17.8	69.5
4	39.1	2.7	58.0	19.7	1.5	38.3	1,006.7	21.7	268.9	1,686.2	41.7	513.9
5	124.9	3.2	77.9	54.2	1.9	50.9	2,054.5	25.3	337.8	1,321.4	23.1	250.7
6	137.2	3.6	90.6	54.5	2.5	64.6	1,577.8	41.2	666.0	868.2	39.6	539.6
7	210.0	3.4	89.1	88.7	2.7	71.4	4,320.0	45.4	876.8	2,254.7	46.3	783.2
8	286.2	3.6	94.5	118.1	2.9	80.3	3,891.8[a]	49.3	963.3	1,992.6[a]	51.0	925.5
9	2,450.8	3.8	99.3	1,025.0	4.3	84.8	49,149.2	44.9	677.1	23,524.2	55.8	621.9
10	4,535.9	3.8	91.4	1,818.3	4.3	74.4	52,201.2[a]	41.5	636.3	24,005.9[a]	51.4	609.8
11	6,867.5	3.7	96.1	2,903.4	4.3	84.9	[b]	—	—	[b]	—	—

[a]Algorithm converged with wrong objective function value due to numerical instability in nonlinear subproblems.
[b]Time limit of 48 hours reached.

restricted to the first iteration. Our experiments show that Algorithm 3.2.6 does work in general but performs poorly in cases of heavy ramping activity. However, it has the potential to outperform the LP-based algorithm if we can handle multiple cuts. In this case, no overestimator is required anymore and tight cuts are obtained in every iteration, drastically reducing the number of iterations.

A summary of the best monolith and the best algorithm for each instance can be found in Table 8. The direct cut calculation approaches clearly outperform the purely LP-based and monolithic approaches. This becomes evident in (Core) instance 8, for which Algorithm 3.2.4 on four threads is faster than the best monolith by a factor of 107,074. Table 9 compares the state-of-the-art GBD implementation to the LP-based and the cut calculation approaches. The GBD is outperformed across all instances and numbers of threads. In general, the fact that most Benders optimality cuts are not tight in the early iterations does not have a strong impact on the convergence performance. The time used for solving the master problems is not explicitly reported in Tables 4–7 but tends to be negligible. For instance, the time spent in the subproblem of the largest instance of (Core:T), which requires 97 iterations, varies between 91% and 99%. The remaining time is spent in the master problems as well as data distribution, which is considerable for the larger instances. The non-subproblem time tends to represent a larger fraction for the cut calculation approaches since the subproblem, especially in the case of (Core), is solved extremely efficiently.

5. Conclusion

The tailored algorithms we present in this work outperform the monolithic approaches by one to five orders of magnitude and solve all but one model instance to optimality within 48 hours. In contrast, the largest instance a monolithic approach is able to solve spans one year. Our efficient cut calculation approach, which makes the solution of mathematical programming problems in the Benders subproblem obsolete, outperforms the monolith in this instance by a factor of more than 100,000. The same algorithm is able to solve the overall largest instance, consisting of 79 million variables and 79 million constraints, in 34.2 seconds on a standard desktop computer. To our knowledge, we have therefore solved, for our specific problem, the largest reported convex MINLP model within one minute of CPU time.

A comparison of our algorithms to classical generalized Benders, in which the nonlinear subproblems are solved as NLPs, shows tremendous improvements, both with respect to computational speed as well as numerical stability. This proves the viability of overestimating the NLP subproblem instead of solving it directly, although the Benders optimality cuts may not be tight in early Benders iterations.

References

Albadi MH, El-Saadany EF (2008) A summary of demand response in electricity markets. *Electric Power Systems Res.* 78(11): 1989–1996.

Anderson D (1972) Models for determining least-cost investments in electricity supply. *Bell J. Econom. Management Sci.* 3(1):267–299.

Arroyo JM, Conejo AJ (2004) Modeling of start-up and shut-down power trajectories of thermal units. *IEEE Trans. Power Systems* 19(3):1562–1568.

Baringo L, Conejo AJ (2011) Wind power investment: A Benders decomposition approach. *IEEE Trans. Power Systems* 27(1): 433–441.

Benders JF (1962) Partitioning procedures for solving mixed variables programming problems. *Numerische Mathematik* 4(1): 238–252.

Downloaded from informs.org by [2a00:1398:4:6c02:dcd6:f884:3020:56e3] on 09 September 2024, at 04:16 . For personal use only, all rights reserved.

Benders JF (2005) Partitioning procedures for solving mixed-variables programming problems. *Comput. Management Sci.* 2(1): 3–19.

Birge JR, Louveaux F (2011) *Introduction to Stochastic Programming*, Operations Research and Financial Engineering, 2nd ed. (Springer, New York).

Blanford GJ, Merrick JH, Young D (2014) A clean energy standard analysis with the US-REGEN model. *Energy J.* 35(Special Issue).

Bloom JA (1982) Long-range generation planning using decomposition and probabilistic simulation. *IEEE Trans. Power Apparatus Systems* 101(4):797–802.

Bloom JA (1983) Solving an electricity generating capacity expansion planning problem by generalized benders' decomposition. *Oper. Res.* 31(1):84–100.

Bloom JA, Caramanis M, Charny L (1984) Long-range generation planning using generalized benders' decomposition: Implementation and experience. *Oper. Res.* 32(2):290–313.

Borenstein S (2005) The long-run efficiency of real-time electricity pricing. *Energy J.* 26(3):93–116.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Bushnell J (2003) A mixed complementarity model of hydrothermal electricity competition in the Western United States. *Oper. Res.* 51(1):80–93.

Bushnell J (2010) Building blocks: Investment in renewable and non-renewable technologies. Moselle B, Padilla J, Schmalansee R, eds. *Harnessing Renewable Energy in Electric Power Systems* (Resources for the Future Press, Washington, DC).

Cappers P, Goldman C, Kathan D (2010) Demand response in U.S. electricity markets: Empirical evidence. *Energy* 35(4):1526–1535.

Castillo E, Conejo AJ, Pedregal P, García R, Alguacil N (2002) *Building and Solving Mathematical Programming Models in Engineering and Science* (Wiley, Hoboken, NJ).

DeJonghe C, Hobbs BF, Belmans R (2011a) Integrating short-term demand response into long-term investment planning. Cambridge Working Papers in Economics 1132, Faculty of Economics, University of Cambridge, Cambridge, UK.

DeJonghe C, Hobbs BF, Belmans R (2012) Optimal generation mix with short-term demand response and wind penetration. *IEEE Trans. Power Systems* 27(2):830–839.

DeJonghe C, Delarue E, Belmans R, D'haeseleer W (2011b) Determining optimal electricity technology mix with high level of wind power penetration. *Appl. Energy* 88(6):2231–2238.

De Wolf D, Smeers Y (1996) Optimal dimensioning of pipe networks with application to gas transmission networks. *Oper. Res.* 44(4):596–608.

EPIS (2015) AuroraXMP. Accessed May 5, 2016, http://epis.com/aurora_xmp/.

Fell H, Linn J (2013) Renewable electricity policies, heterogeneity, and cost-effectiveness. *J. Environ. Econom. Management* 66(3):688–707.

Frank S, Steponavice I, Rebennack S (2012) Optimal power flow: A bibliographic survey I—Formulations and deterministic methods. *Energy Systems* 3(3):221–258.

García J, Román J, Barquín J, González A (1999) Strategic bidding in deregulated power systems. *Proc. 13th Power Systems Comput. Conf.*, Vol. 1 (IEEE, New York), 258–264.

Geoffrion AM (1972) Generalized benders decomposition. *J. Optim. Theory Appl.* 10(4):237–260.

Gollmer R, Nowak MP, Römisch W, Schultz R (2000) Unit commitment in power generation: A basic model and some extensions. *Ann. Oper. Res.* 96(1–4):167–189.

Gomez-Exposito A, Conejo AJ, Canizares C (2008) *Electric Energy Systems: Analysis and Operation* (CRC Press, Boca Raton, FL).

Gross G, Finlay D (2000) Generation supply bidding in perfectly competitive electricity markets. *Comput. Math. Organ. Theory* 6(1): 83–98.

Guan X, Luh PB, Yan H (1992) An optimization-based method for unit commitment. *Eletric Power Energy Systems* 14(1):9–17.

Guan Z, Philpott AB (2011) A multistage stochastic programming model for the New Zealand dairy industry. *Internat. J. Production Econom.* 134(2):289–299.

Han XS, Gooi HB, Kirschen DS (2001) Dynamic economic dispatch: Feasible and optimal solutions. *IEEE Trans. Power Systems* 16(1):22–28.

Ho JK, Manne AS (1974) Nested benders decomposition for dynamic models. *Math. Programming* 6(1):121–140.

Hobbs BF (1995) Optimization methods for electric utility resource planning. *Eur. J. Oper. Res.* 83(1):1–20.

Hobbs BF, Rothkopf MH, O'Neill RP, Chao H-P (2001) *The Next Generation of Electric Power Unit Commitment Models* (Springer, New York).

Jin S, Ryan SM, Watson J-P, Woodruff DL (2011) Modeling and solving a large-scale generation expansion planning problem under uncertainty. *Energy Systems* 2(3–4):209–242.

Kazerooni AK, Mutale J (2010) Network investment planning for high penetration of wind energy under demand response program. *IEEE 11th Internat. Conf. Probabilistic Methods Appl. Power Systems (PMAPS)* (IEEE, New York), 238–243.

Kim H, Sohn H-S, Bricker DL (2011) Generation expansion planning using Benders' decomposition and generalized networks. *Internat. J. Indust. Engrg.* 18(1):25–39.

Kumar N, Besuner PM, Lefton SA, Agan DD, Hilleman DD (2012) Power plant cycling costs. Technical report, National Renewable Energy Laboratory, Golden, CO. http://wind.nrel.gov/public/wwis/aptechfinalv2.pdf, prepared for NREL.

Lindsay J, Dragoon K (2010) Summary report on coal plant dynamic performance capability. Technical report, Renewable Northwest Project, Portland, OR.

Martinez A, Eurek K, Mai T, Perry A (2013) Integrated Canada-U.S. power sector modeling with the Regional Energy Deployment System (ReEDS). Report No. TP-6A20-56724, National Renewable Energy Laboratory, Golden, CO.

Massé P, Gibrat R (1957) Application of linear programming to investments in the electric power industry. *Management Sci.* 3(1):149–166.

Maurer L, Barroso L (2011) *Electricity Auctions: An Overview of Efficient Practices (World Bank Studies)* (World Bank Publications, Washington, DC).

Murphy FH, Smeers Y (2005) Generation capacity expansion in imperfectly competitive restructured electricity markets. *Oper. Res.* 53(4):646–661.

Nolden C, Schöfelder M, Eßer-Frey A, Bertsch V, Fichtner W (2013) Network constraints in techno-economic energy system models: Towards more accurate modeling of power flows in long-term energy system models. *Energy Systems* 4(3):267–287.

Paul A, Burtraw D, Palmer K (2009) Haiku documentation: RFF's electricity market, model version 2.0. Report, Resources for the Future, Washington, DC.

Pereira MVF, Pinto LMVG (1991) Multi-stage stochastic optimization applied to energy planning. *Math. Programming* 52(1–3):359–375.

Stoft S (2002) *Power System Economics* (Wiley-IEEE Press, Hoboken, NJ).

Tseng CL, Li CA, Oren SS (2000) Solving the unit commitment problem by a unit decommitment method. *J. Optim. Theory Appl.* 105(3):707–730.

U.S. Energy Information Administration (2015) Assumptions to AEO2015. Accessed May 5, 2016, http://www.eia.gov/forecasts/aeo/assumptions.

Vavasis SA (1991) *Nonlinear Optimization: Complexity Issues* (Oxford University Press, Oxford, UK).

Wang C, Shahidehpour SM (1995) Optimal generation scheduling with ramping costs. *IEEE Trans. Power Systems* 10(1):60–67.

Warland G, Haugstad A, Huse ES (2008) Including thermal unit start-up costs in a long-term hydro-thermal scheduling model. *Proc. 16th Power System Comput. Conf.* (IEEE, New York), 808–814.

Xia X, Elaiw AM (2010) Optimal dynamic economic dispatch of generation: A review. *Electric Power Systems Res.* 80(8):975–986.