OPTIMIZING THE ECONOMIC EFFICIENCY OF WHOLESALE ELECTRICITY MARKETS

by

Yanchao Liu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Industrial and Systems Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: July 14, 2014

The dissertation is approved by the following members of the Final Oral Committee:
Michael C. Ferris, Professor, Computer Sciences, Industrial and Systems Engineering
Jeffrey T. Linderoth, Professor, Industrial and Systems Engineering
James R. Luedtke, Associate Professor, Industrial and Systems Engineering
Bernard Lesieutre, Professor, Electrical and Computer Engineering
Thomas F. Rutherford, Professor, Agricultural and Applied Economics

To my parents Liu Tse and Zhang Fengmin, and to my wife Wang Qilu.

ACKNOWLEDGMENTS

I would like to thank Professor Michael C. Ferris for being a great advisor. He has been tremendously helpful in advancing my knowledge, skills and research vision. While fostering independence in my research, he has offered invaluable ideas, insights and directions that have guided me to overcome numerous difficulties along the journey. I appreciate all his training, advice and encouragement that have made my graduate study productive and enjoyable. My dissertation would not have been completed without his sagacious advice and continual support.

I would like to thank Professors Bernard Lesieutre, Jeffrey T. Linderoth, James R. Luedtke and Thomas F. Rutherford for their time and effort to serve on my doctoral committee and their generous help, advice and encouragement in many aspects over the years.

My interest in power systems was sparked by a handwritten note by Professor Christopher L. DeMarco. The writing of Chapter 1 was primarily motivated by his example. I would like to acknowledge his enlightenment for me in this exciting field. The work on stochastic unit commitment in Chapter 5 and on multi-stage security-constrained economic dispatch in Chapter 6 would not have been possible without the knowledge obtained from ISO New England, Inc. I am grateful for Eugene Litvinov and Tongxin Zheng for offering me the internship opportunity and I would like to thank them, along with Feng Zhao, my direct supervisor, and Jinye Zhao at the ISO, for making the experience so rewarding and for their continual support in my future research.

I would like to thank Professor Leyuan Shi for admitting me to the University of Wisconsin-Madison and her guidance in the first two years of my graduate study. I would like to thank Professor Stephen M. Robinson for his genuine care, advice and support of my professional growth.

Many thanks to my friends, in particular Jesse Holzer, Lisa Tang, Youngdae Kim and Taedong Kim, at the Wisconsin Institutes for Discovery for being great company through the years. Special thanks go to Herman Stampfli for his help with logistics.

CONTENTS

Contents iii

List of Tables vi

List of Figures viii

Abstract x

1 Overview of Power Economics Models 1
- *1.1 The Power Flow Equation* 1
- *1.2 Computing the Bus Admittance Matrix* Y_{bus} 4
- *1.3 Polar Coordinates Formulation of the Power Flow* 8
- *1.4 DC Approximation* 13
- *1.5 Economic Dispatch* 15
- *1.6 Unit Commitment* 17
- *1.7 Market Clearing Price and Locational Marginal Price* 21
- *1.8 Data Sources and Formats* 22
- *1.9 GAMS Model Suite* 24
- *1.10 A Case Study at ISO New England* 26

2 Modeling Demand Response for FERC Order 745 36
- *2.1 Introduction* 36
- *2.2 Modeling the Demand Response* 45
- *2.3 Alternative Approaches* 53
- *2.4 Numerical Experiments* 60
- *2.5 Extensions* 70
- *2.6 Conclusion* 75

3 Payment Rules for Unit Commitment Dispatch 77
- *3.1 Introduction* 77

3.2 The Problematic Payment Rule 78
3.3 The Imperfect Two-sided Electricity Market 80
3.4 Justification of Pay-as-bid in the UCED Context 81
3.5 Suppliers' Response under Pay-as-bid 88
3.6 Conclusion 94

4 Extended Bidding Structure for Demand Response 95
4.1 Introduction 95
4.2 Demand Types and Behavioral Models 99
*4.3 Bidding and Central Dispatch Model*104
*4.4 Implementation and Experiments*109
*4.5 Conclusion*116

5 Stochastic Unit Commitment with Derand Sampling Method118
*5.1 Introduction*118
*5.2 Problem Formulation*121
*5.3 Derandomization Sampling Method*121
*5.4 Performance Evaluation*129
*5.5 Conclusion*136

6 Multi-stage Security-constrained Economic Dispatch138
*6.1 Introduction*138
*6.2 The Model and its Structure*142
*6.3 Benders' Decomposition*148
*6.4 Computational Enhancements*151
*6.5 Numerical Experiments*161
*6.6 Conclusion*167

7 Security-constrained Economic Dispatch using Semidefinite Programming169
*7.1 Introduction*169
*7.2 Benders' Decomposition with SDP Subproblems*173

7.3 Numerical Experiments 178
7.4 Conclusion 182

8 Conclusion 183

References 187

LIST OF TABLES

1.1 Notations for the Complex Formulation 4
1.2 Symbols for the Branch data 6
1.3 Electrical Line Characteristics Measures 8
1.4 Notations for the Polar Formulation 9
1.5 More Notations for the Polar Formulation 11
1.7 Counts of Different Modeling Elements in the ISO's System 28
1.6 List of Files in the Model Suite 35

2.1 Notations for the Economic Dispatch Model 46
2.2 Cost Parameters of a Two-generator Example 55
2.3 DR1 Results for Cost-ineffective Demand Levels 63
2.4 DR1 Results for Cost-effective Demand Levels with Different C_1 Values 63
2.5 Comparison of DR1 and LS1 Solutions for $C_1 = 73$ 65
2.6 Comparison of DR1 and LS1 Solutions for Different C_1 Levels 66
2.7 Setting and Solution of IEEE Test Cases 67
2.8 Solution Time (in seconds) of Different Formulations and Solvers . . . 67
2.9 DR Test Results on Polish Networks 69
2.10 Settings and Bounding Results on Polish Networks 69

3.1 Summary of the Generators by Fuel Types 89
3.2 Characteristics of GEN6 90
3.3 Marginal Costs of GEN6 90

4.1 Bidding Parameters and Decision Variables 105
4.2 Cost Results 114

5.1 Stochastic RAA UC Instance Size v.s. Number of Scenarios 120
5.2 Stochastic UC Performance with 3 Scenarios 133
5.3 Stochastic UC Performance with 5 Scenarios 134

6.1 Time (seconds) Spent in Sequentially Solving 100 Subproblems using Different LP Methods, on a Dell Laptop[6] 159
6.2 Solution Statistics of Different Formulations 163
6.3 Time (seconds) Spent to Pre-screen for Different Cases. The LPs Are Solved Sequentially. 164
6.4 Solution for Big Cases, 80 threads, $L^{fc} = 5$ 165
6.5 Active Contingencies at Optimum 166
6.6 Contingency-to-line Mapping for Active Contingencies at Optimum . . 166
6.7 SCED Solution Considering Different Post-contingency Stages 167

7.1 Solution Comparison of Three SCED Models 181

LIST OF FIGURES

1.1 Power engineering view of a transmission line 8
1.2 Relations of different modeling elements 28

2.1 Electricity supply and demand curves 39
2.2 Market inefficiency caused by imperfect demand information 40
2.3 LMP curve in a 3-generator example 54
2.4 An LMP curve with a jump 56
2.5 DR cost-effectiveness test on an LMP curve 57
2.6 Heuristic framework for finding a DR solution 59
2.7 LMP curve for the 14-bus case without line limits 62
2.8 Optimal solutions for decrementing C_1 levels on different data cases . 71
2.9 Simulation results for the 300-bus case. 73
2.10 Comparison of DR model variants on the 14-bus case. 75

3.1 Profit curves of two generators with different commitment costs 91
3.2 Payoff of GEN5 under pay-as-bid 92
3.3 Payoff of GEN6 under pay-as-bid 93

4.1 Framework for demand-side participation 100
4.2 Day-ahead demand profile of FERC dataset 4012 111
4.3 Elastic demand bid for hour 1 112
4.4 Effect of extended bidding on LMP 114
4.5 LMP for different arbitrage levels 116
4.6 Profit of arbitrage for different penetration and efficiency 117

5.1 Relative forecast error distribution in 2011. 128
5.2 2011 average zonal share of the system load. 130
5.3 Load scenarios generated by Derand. 131
5.4 Load scenarios generated by SAA. 132
5.5 Load scenarios of SCENRED by cost. 133

5.6 Load scenarios of SCENRED by error. 134
5.7 Cost savings of 3-scenario stochastic model with Derand sampling. . . 135

6.1 Post-contingency line flow requirements 142
6.2 On the u_0 plane, the feasible region of an N-1 SCED is the intersection of N polyhedra. Since it would involve huge numbers of variables and constraints to represent all these polyhedra, we leave most of them out from the master model and use Benders' cuts to approximate the relevant pieces. 144
6.3 When L1 fails, flow on L2 will instantly rise to 150 MW. Ramping up G1 and G2 can meet STE requirement, while ramping up G1 and G3 can meet LTE requirement. However, STE and LTE can not be satisfied simultaneously. 147
6.4 Sparsity structure of the Jacobian matrix of a 6-bus case with 3 contingencies and 3 post-contingency checkpoints. 149
6.5 Performance of different formulations on an instance of 118-bus case with 20 contingencies (subproblems). 153
6.6 Contingency 2 is intrinsically infeasible. Either the corresponding subproblem is infeasible or its Benders' cuts will render the master problem infeasible. 154
6.7 Each individual contingency is feasible, but they are not simultaneously feasible. Their Benders' cuts will render the master problem infeasible. 155
6.8 Algorithm progress on the 118-bus case. 157

7.1 Benders' iterations of different models for the 118-bus case. 179

ABSTRACT

Improving the economic efficiency of today's wholesale energy markets has generated tremendous interest and actions among policy makers, market participants, and operations researchers. This dissertation aims to contribute by proposing solutions to some imperative issues at the heart of the deregulated market design, operations and their policy framework. Mathematical programming, hierarchical optimization modeling and parallel computing techniques constitute the methodological basis of my research. The dissertation is organized as follows.

Chapter 1 presents a pedagogical overview of the power economics models that exert direct impact on the electricity markets operations and efficiency. The exposition of the models enables a straightforward translation to computer programs in an algebraic modeling language such as GAMS and AMPL. This chapter also serves as a first documentation of the GAMS model suite for power systems and power economics models which are evolving into a "library" of data and models through a joint project with FERC.

Chapter 2 studies the FERC Order 745 regarding demand response compensation in organized wholesale energy markets and investigates different approaches to model and solve a compliant implementation of the Order. In an economic sense, demand response in the Order context is a trade of "consuming rights" instead of a sale of energy, therefore it must be traded separately from the energy market. In this chapter, a bi-level optimization model is developed to simultaneously clear the energy and demand response markets and a three-phase solution procedure is devised for large-scale instances.

Chapter 3 analyzes the efficiency and equity issues of the existing payment rules in the context of unit commitment economic dispatch, and justifies an alternative pay-as-bid rule for consideration by policy makers. The inefficiency of the existing payment rule is rooted in its pricing mechanism. This chapter argues that pricing only the power balance constraints and neglecting the marginal prices of other constraints lack justification and concludes, based on linear programming duality, that a theoretically correct pricing mechanism exactly corresponds to a pay-as-

bid payment rule. The effectiveness of this payment rule is then validated via a simulation of market participants' bidding behaviors in a realistic experiment setting.

Chapter 4 extends the demand response discussion and proposes a general bidding structure that clears obstacles for efficient demand-side participation. It is observed that the existing bid formats are all separable over time while a significant and growing segment of demand can be shifted across time and hence has no way to bid its true valuation of consumption. To meet the growing trend, this chapter proposes additional bid types that allow deferrable, adjustable and storage-type loads to better express their value, thus elicit demand response in the most natural way – direct participation in the market. It is then shown that these bid types are easily incorporated into the existing market with no technological barrier and that they preserve the market's efficiency and incentive-compatibility properties.

Chapter 5 presents a stochastic programming model for ISO New England's reserve adequacy analysis that manages the load uncertainty. Due to the large size of the ISO's system and the increasing net-load variability caused by increasing penetration of renewable resources, the problem is computationally challenging. This chapter develops an effective scenario reduction technique, Derandomization (or Derand), to identify a small number of scenarios that extract key and unbiased information from the distribution of random variables. Numerical testing results show that the stochastic model with only 3 or 5 scenarios outperforms its deterministic counterpart by a significant margin. Results also show that the Derand method outperforms several conventional scenario reduction methods, and the solution quality is comparable to the cost based scenario-reduction technique but with less computational efforts.

Chapter 6 deals with the security-constrained economic dispatch (SCED) which pivots economic efficiency and operational reliability of the power grid. Post-contingency corrective actions are modeled in SCED while multiple stages of rescheduling are considered to meet different security constraints. The resulting linear program is not solvable by traditional LP methods due to its large size. A series of algorithmic enhancements based on the Benders' decomposition method

are proposed to ameliorate the computational difficulty. In addition, a set of online measures are devised to diagnose and correct infeasibility issues encountered in the solution process. The overall solution approach, coded directly in GAMS, is able to process the "N-1" contingency list in ten minutes for all large network cases available for experiments.

Chapter 7 proposes a novel solution approach for the SCED problem in a nonlinear AC setting, in which the model is a large-scale nonconvex problem and is extremely difficult to solve. The proposed approach deals with the scale and nonconvexity issues separately and effectively. The key point is to approximate the nonconvex AC feasibility problem with its semidefinite programming (SDP) relaxation and use these SDP models as a convex subproblem within a Benders' decomposition framework. Numerical experiments demonstrate the superior solution quality of this approach and its tractability for IEEE test cases.

1 OVERVIEW OF POWER ECONOMICS MODELS

1.1 The Power Flow Equation

Operations researchers can construe an electrical grid as a directed graph, in which the nodes are called *buses* and the arcs *lines*. A bus is either a *generating unit*, generating electrical energy, or a *load*, consuming the energy. While the arcs represent the actual transmission lines, the directional nature of them should not be compared to that of the usual transportation networks in which the commodity flows only in the specified direction along the arcs. Rather, the direction of an arc merely indicates the sign of the parameters associated with the transmission line, which will be used to compute the *bus admittance matrix*. The bus admittance matrix is what governs the energy transmission characteristics of the network. We will come back to it shortly.

In the electricity realm, one cannot avoid dealing with complex matrices and the algebra. Simply put, the alternating current carries power in two forms, *active* and *reactive*. The active power is the rate of energy that gets delivered to and consumed by the loads (such as lighting a light bulb or running an electric motor, etc.) while the reactive power is a kind of energized wave swinging back and forth along the lines and does not transfer energy. Even though the reactive power does no work at the loads, abundant amount of it must be present at practical loads to account for the loads' reactance. Therefore, the energy demand of a load always comes in the form of a complex number, with the real part standing for the active power and the imaginary part for the reactive power. As a convention, both numbers use negative signs for demand and positive signs for generation. Similarly, the power supplied by a generator is also quantified by a complex number. Both the active and the reactive power can be measured in MVA (megavolt ampere) and the active power is most often measured in MW (megawatt, 1 MW = 1 MVA). In a transmission network, the total active power supplied should equal the total active power consumed plus the transmission losses. This is a basic equation in power systems.

Then how does the energy flow in the network? It turns out that the word "flow" is somewhat off target. Unlike in a transportation network where the flow balance is observed at every node (inbound equals outbound), the power grid simply "distributes" the roles (positive sign for generating units and negative sign for loads) and shares (how much power to contribute or absorb) among the buses as a whole, in which case the microscopic view of the flow becomes irrelevant. For a network of n buses and m lines, the bus admittance matrix is a n by n complex matrix, denoted by Y_{bus}. For power flow and related computations, Y_{bus} is usually known or can be computed from other physical parameters of the transmission system. With Y_{bus}, the power "distribution" can be computed as described below. Each bus k has a complex control variable called the *bus voltage phasor*

$$V_k = |V_k| e^{i\delta_k} \tag{1.1}$$

where $|V_k|$ is the voltage magnitude, δ_k the phase angle, and i the imaginary unit, i.e., $i = \text{sqrt}(-1)$. Alternatively, V_k can be written in the complex form as $V_k = V_k^{real} + iV_k^{imag}$ which gives

$$|V_k|^2 = (V_k^{real})^2 + (V_k^{imag})^2 \tag{1.2}$$

Collectively, the phasors form a vector $V \in \mathbb{C}^n$. The complex power vector $S \in \mathbb{C}^n$ is then calculated as

$$S = V. * (Y_{bus} V)^* \tag{1.3}$$

where $.*$ is element-by-element multiplication and the superscripted $*$ is the conjugate operator, all in the complex field. Let $P \in \mathbb{R}^n$ denote the real (active) power and $Q \in \mathbb{R}^n$ the reactive power, then one has $S = P + Qj$. As discussed above, P_k is positive if bus k is a generating bus and negative if it is a load bus.

There are four fundamental quantities related to a bus k, namely, V_k^{real}, V_k^{imag} (or alternatively $|V_k|$ and δ_k), P_k and Q_k. In the power flow model, for each bus two of the four quantities are given as parameters and the other two are unknown.

Specifically, for a generating bus, P_k and $|V_k|$ are known while Q_k and δ_k are unknown; for a load bus, P_k and Q_k are known while V_k^{real} and V_k^{imag} (or equivalently $|V_k|$ and δ_k) are unknown; and for a swing bus, V_k^{real} and V_k^{imag} are known and P_k and Q_k are unknown. So in a network of n buses, there are altogether $2n$ unknowns. On the other hand, equations (1.1) and (1.2), after plugging in the known parameter values, provide $2n$ nonlinear equations. The task of the power flow study is to solve this system of nonlinear equations.

For the ease of presenting complex equations in a real formulation, an intermediate variable I_k is introduced for each bus k to represent the complex current generated by the bus. It is defined as

$$I = Y_{bus} V \tag{1.4}$$

and it relates to the complex S and V via the following equation,

$$S = V. * I^* \tag{1.5}$$

or equivalently for each bus k,

$$S_k = V_k I_k^* \tag{1.6}$$

The power flow model is presented below, following the list of notations in Table 1.1.

$$I_k^{real} = \sum_{l \in B \cup S} (Y_{bus}^{real}(k,l) V_l^{real} - Y_{bus}^{imag}(k,l) V_l^{imag}), \qquad \forall k \in B \cup S \tag{1.7}$$

$$I_k^{imag} = \sum_{l \in B \cup S} (Y_{bus}^{real}(k,l) V_l^{imag} + Y_{bus}^{imag}(k,l) V_l^{real}), \qquad \forall k \in B \cup S \tag{1.8}$$

$$P_k = V_k^{real} I_k^{real} + V_k^{imag} I_k^{imag}, \qquad \forall k \in B \cup S \tag{1.9}$$

$$Q_k = V_k^{real}(-I_k^{imag}) + V_k^{imag} I_k^{real}, \qquad \forall k \in B \cup S \tag{1.10}$$

$$|V_k|^2 = (V_k^{real})^2 + (V_k^{imag})^2, \qquad \forall k \in B \cup S \tag{1.11}$$

Table 1.1: Notations for the Complex Formulation

BUS	The set of buses, also denoted by $\mathcal{B}$
$k, l \in BUS$	The indices of the buses
$Y_{bus}^{real}(k,l)$	Real part of the (k,l)-th element of the Y_{bus} matrix
$Y_{bus}^{imag}(k,l)$	Imaginary part of the (k,l)-th element of the Y_{bus} matrix
V_k^{real}	Real part of the voltage phasor of bus k
V_k^{imag}	Imaginary part of the voltage phasor of bus k
$\|V_k\|$	Voltage magnitude of the bus k
I_k^{real}	Real part of the current generated by bus k
I_k^{imag}	Imaginary part of the current generated by bus k
P_k	Real power output of bus k
Q_k	Reactive power output of bus k

One can do a simple counting to verify that the model is a square system. The lower part of Table 1 lists seven variables for each bus, among which two have fixed values, so there are altogether $5n$ unknown variables. And equations (1.7) to (1.11) give $5n$ equations.

1.2 Computing the Bus Admittance Matrix Y_{bus}

Y_{bus} is where information about transfer capability of the network is encapsulated. From a modeler's standpoint, it is merely the raw data organized in a matrix form and this matrix is supposed to be provided for power flow and other related computations. Indeed, every modeler's life would have been easier if Y_{bus}'s were provided in the benchmark data sets, not because it is difficult to compute, but because there are multiple versions of formulas for Y_{bus}, or its equivalence in the polar format, being used in the literature depending on how much network detail an author chooses to include or neglect. This is a roadblock for comparison work that involves models and formulations from different sources, and is especially annoying when the specific formula being used is not articulated in the work. Therefore, the formation of the Y_{bus} matrix is not something that a responsible modeler could tolerate to remain mysterious. Dobson et al. (2001) provide an excellent engineering

tutorial for computing and analyzing the power transfer characteristics of a network.

This section gives a relatively comprehensive formula for computing Y_{bus}, which involves all the relevant quantities in the IEEE Common Data Format (CDF) documented by Christie (1993). This formula is consistent with that used in the *makeYbus.m* routine in the Matpower package (Zimmerman et al., 2011), but the differences between the two are also worth noting. In some cases, there are multiple physical lines, called circuits, running between two buses, so in the IEEE CDF each line is identified by a triple, e.g., (k, l, c), meaning that the line is the c-th circuit between bus k and bus l and the positive sign direction is along (k, l). However, the MatPower case format (see *caseformat.m*) has obsoleted the circuit identifier, i.e., the third element in the triple, and allowed the data set to contain duplicate (k, l) entries. The drawback is apparent. For one, the *makeYbus.m* routine based on this format heavily uses Matlab specific matrix manipulation functions which obscures physical interpretations; for another, allowing for duplicate data entries is not a common practice in the mathematical modeling context, hence limits the usability of the case format outside the Matlab platform.

Let CIR be the set of circuit numbers and more notations are summarized in Table 1.2. The third column indicates the column ranges that the quantity occupy in IEEE CDF data files.

For each line (k, l, c), the conductance and susceptance are

$$G_{(k,l,c)} = \frac{r_{(k,l,c)}}{r^2_{(k,l,c)} + x^2_{(k,l,c)}} \tag{1.12}$$

and

$$B_{(k,l,c)} = \frac{-x_{(k,l,c)}}{r^2_{(k,l,c)} + x^2_{(k,l,c)}} \tag{1.13}$$

respectively. And for each bus k, define the shunt admittance Y_k by

$$Y_k^{real} = \frac{Gs_k}{baseMVA} \tag{1.14}$$

$$Y_k^{imag} = \frac{Bs_k}{baseMVA} \tag{1.15}$$

Table 1.2: Symbols for the Branch data

Symbol	Quantity	In CDF
$k \in BUS$	From bus number	BR 1-4
$l \in BUS$	To bus number	BR 6-9
$c \in CIR$	Circuit number	BR 17
$r_{(k,l,c)}$	Branch resistance p.u.	BR 20-29
$x_{(k,l,c)}$	Branch reactance p.u.	BR 30-40
$bb_{(k,l,c)}$	Line charging susceptance p.u.	BR 41-50
$t_{(k,l,c)}$	Transformer ratio	BR 77-82
$\alpha_{(k,l,c)}$	Transformer angle (deg)	BR 84-90
$k \in BUS$	Bus number	BU 1-4
Gs_k	Shunt conductance p.u.	BU 107-114
Bs_k	Shunt susceptance p.u.	BU 115-122
$baseMVA$	MVA Base	TI 32-37

Then the diagonal entries of the Y_{bus} can be computed as

$$\begin{aligned} Y_{bus}(k,k) = &\sum_{\substack{j\in BUS \\ c\in CIR}} G_{(j,k,c)} + i(B_{(j,k,c)} + \frac{bb_{(j,k,c)}}{2}) \\ &+ \sum_{\substack{j\in BUS \\ c\in CIR}} \frac{G_{(k,j,c)} + i(B_{(k,j,c)} + \frac{bb_{(k,j,c)}}{2})}{t^2_{(k,j,c)}} \\ &+ (Y_k^{real} + iY_k^{imag}) \end{aligned} \tag{1.16}$$

and the off-diagonal entries are

$$Y_{bus}(k,l) = \begin{cases} -\sum_{c\in CIR} \frac{G_{(k,l,c)} + iB_{(k,l,c)}}{t_{(k,l,c)} e^{-i\alpha_{(k,l,c)}}}, & \text{if } (k,l) \text{ exists} \\ -\sum_{c\in CIR} \frac{G_{(l,k,c)} + iB_{(l,k,c)}}{t_{(l,k,c)} e^{i\alpha_{(l,k,c)}}}, & \text{if } (l,k) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{1.17}$$

Since equation (1.17) contains complex quantities, a step further is needed to

"realize" them. Re-write the transformer tap by

$$tap^{real}_{(k,l,c)} = t_{(k,l,c)} \cos(\alpha_{(k,l,c)}) \tag{1.18}$$

$$tap^{imag}_{(k,l,c)} = t_{(k,l,c)} \sin(\alpha_{(k,l,c)}) \tag{1.19}$$

then the real and imaginary parts of Y_{bus} can be expressed separately using real quantities, as follows.

$$Y^{real}_{bus}(k,l) = \begin{cases} \sum\limits_{\substack{j\in BUS \\ c\in CIR}} G_{(j,l,c)} + \sum\limits_{\substack{j\in BUS \\ c\in CIR}} \frac{G_{(k,j,c)}}{t^2_{(k,j,c)}} + Y^{real}_k, & \text{if } k = l \\ \sum\limits_{c\in CIR} \frac{-G_{(k,l,c)} tap^{real}_{(k,l,c)} + B_{(k,l,c)} tap^{imag}_{(k,l,c)}}{t^2_{(k,l,c)}}, & \text{if } (k,l,c) \text{ exists} \\ \sum\limits_{c\in CIR} \frac{-G_{(l,k,c)} tap^{real}_{(l,k,c)} - B_{(l,k,c)} tap^{imag}_{(l,k,c)}}{t^2_{(l,k,c)}}, & \text{if } (l,k,c) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{1.20}$$

$$Y^{imag}_{bus}(k,l) = \begin{cases} \sum\limits_{\substack{j\in BUS \\ c\in CIR}} (B_{(j,l,c)} + \frac{bb_{(j,l,c)}}{2}) + \sum\limits_{\substack{j\in BUS \\ c\in CIR}} \frac{(B_{(k,j,c)} + \frac{bb_{(k,j,c)}}{2})}{t^2_{(k,j,c)}} + Y^{imag}_k, & \text{if } k = l \\ \sum\limits_{c\in CIR} \frac{-G_{(k,l,c)} tap^{imag}_{(k,l,c)} - B_{(k,l,c)} tap^{real}_{(k,l,c)}}{t^2_{(k,l,c)}}, & \text{if } (k,l,c) \text{ exists} \\ \sum\limits_{c\in CIR} \frac{G_{(l,k,c)} tap^{imag}_{(l,k,c)} - B_{(l,k,c)} tap^{real}_{(l,k,c)}}{t^2_{(l,k,c)}}, & \text{if } (l,k,c) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{1.21}$$

Table 1.3: Electrical Line Characteristics Measures

Term	Symbol	A measure of ...
Resistance	R	the opposition to the passage of an electric current
Reactance	X	the opposition to a change of current
Impedance	Z	the opposition to alternating current (AC)
Conductance	G	the ease of electricity to flow along the line
Susceptance	B	the ease of polarization of the line
Admittance	Y	the ease to allow an AC to flow

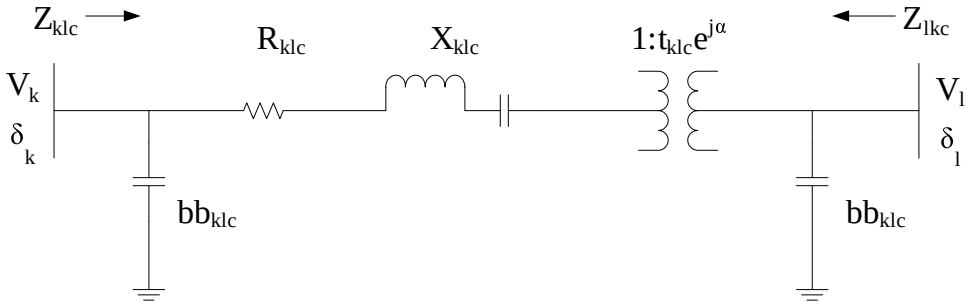


Figure 1.1: Power engineering view of a transmission line

1.3 Polar Coordinates Formulation of the Power Flow

Another prevalent formulation uses the trigonometric form of complex numbers and more expressively involves the physical characteristics of the lines and buses in the equations, instead of encapsuling them in the Y_{bus} matrix. In this formulation, the power flow is calculated line by line and the flow balance constraints are enforced at each bus, thus the traditional "network flow" point of view becomes relevant.

In order to make sense of this formulation, some electrical terms are symbolized in Table 1.3 and a graphical depiction of a typical transmission line is given in Figure 1.1. Details are abundant in any introductory electrical engineering textbook, hence are omitted here for succinctness.

For a single transmission line, the following relations hold,

$$Y = G + jB,\ Z = R + jX,\ Y = \frac{1}{Z}, \tag{1.22}$$

Table 1.4: Notations for the Polar Formulation

V_i	The voltage magnitude of bus i
δ_i	The voltage angle of bus i relative to the swing bus
t_k	Ideal transformer tap ratio on line k
α_k	Ideal transfomer phase angle shift on linek k
G_k	$= R_k/(R_k^2 + X_k^2)$, conductance of line k
Ω_k	$= X_k/(R_k^2 + X_k^2)$, negative susceptance of line k
B_k^{cap}	Line charging capacitance of line k
Z_{ijk}^P	Real power flowing from bus i to j along line k between them
Z_{ijk}^Q	Reactive power flowing from bus i to j along line k between them
y_i^P	Net real power injection into the network at bus i
y_i^Q	Net reactive power injection into the network at bus i
$I(i)$	Set of buses which are linked to bus i by a line

and therefore,

$$G = \frac{R}{R^2 + X^2},\ B = -\frac{X}{R^2 + X^2} \tag{1.23}$$

As the formation of the Y_{bus} matrix is clear, it is tempting to expand the basic power equations (1.3) to (1.6) a little bit and write

$$\begin{aligned} S_k &= V_k(\sum_{l\in BUS} Y_{bus}(k,l)V_l)^* \\ &= \sum_{l\in BUS} V_k(Y_{bus}(k,l)V_l)^* \\ &= \sum_{l\in BUS} S_{(k,l)} \\ &= \sum_{l\in BUS} P_{(k,l)} + iQ_{(k,l)} \\ &= \sum_{l\in BUS\setminus\{k\}} (\sum_{c\in CIR} P_{(k,l,c)} + i\sum_{c\in CIR} Q_{(k,l,c)}) + (P_{(k,k)} + iQ_{(k,k)}) \end{aligned} \tag{1.24}$$

where $S_{(k,l)}$ represents the complex power flowing from bus k to bus l, with $P_{(k,l)}$ being its real part and $Q_{(k,l)}$ its imaginary part. The last line of the above equations further split the power flows in circuits. Since there is no circuit between a bus k

and itself, $P_{(k,k)} + iQ_{(k,k)}$ is used to denote the flow from bus k to itself. Then using equation (1.1) and the $Y_{bus}(k,l)$ formula given in (1.16) and (1.17), the algebraic formulas for $P_{(k,l,c)}$, $Q_{(k,l,c)}$, $P_{(k,k)}$ and $Q_{(k,k)}$ can be derived.

$$P_{(k,l,c)} = \begin{cases} (|V_k||V_l|/t_{(k,l,c)})(-G_{(k,l,c)}\cos(\delta_k - \delta_l - \alpha_{(k,l,c)}) \\ \qquad - B_{(k,l,c)}\sin(\delta_k - \delta_l - \alpha_{(k,l,c)})), & \text{if } (k,l,c) \text{ exists} \\ (|V_k||V_l|/t_{(l,k,c)})(-G_{(l,k,c)}\cos(\delta_k - \delta_l + \alpha_{(l,k,c)}) \\ \qquad - B_{(l,k,c)}\sin(\delta_k - \delta_l + \alpha_{(l,k,c)})), & \text{if } (l,k,c) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{1.25}$$

$$Q_{(k,l,c)} = \begin{cases} (|V_k||V_l|/t_{(k,l,c)})(B_{(k,l,c)}\cos(\delta_k - \delta_l - \alpha_{(k,l,c)}) \\ \qquad - G_{(k,l,c)}\sin(\delta_k - \delta_l - \alpha_{(k,l,c)})), & \text{if } (k,l,c) \text{ exists} \\ (|V_k||V_l|/t_{(l,k,c)})(B_{(l,k,c)}\cos(\delta_k - \delta_l + \alpha_{(l,k,c)}) \\ \qquad - G_{(l,k,c)}\sin(\delta_k - \delta_l + \alpha_{(l,k,c)})), & \text{if } (l,k,c) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{1.26}$$

$$P_{(k,k)} = |V_k|^2\Big(\sum_{\substack{j\in BUS \\ c\in CIR}} G_{(j,k,c)} + \sum_{\substack{j\in BUS \\ c\in CIR}} \frac{G_{(k,j,c)}}{t^2_{(k,j,c)}} + Y_k^{real}\Big) \tag{1.27}$$

$$\begin{aligned} Q_{(k,k)} = -|V_k|^2\Big(&\sum_{\substack{j\in BUS \\ c\in CIR}} (B_{(j,k,c)} + bb_{(j,k,c)}/2) \\ &+ \sum_{\substack{j\in BUS \\ c\in CIR}} \frac{B_{(k,j,c)} + bb_{(k,j,c)}/2}{t^2_{(k,j,c)}} + Y_k^{imag}\Big) \end{aligned} \tag{1.28}$$

Note that the symbols P and Q are somewhat overused (overloaded) in the

Table 1.5: More Notations for the Polar Formulation

Line	Set of lines
$Z^P_{(k,l,c)}$	Real power flowing from bus k to l on the c-th circuit between them
$Z^Q_{(k,l,c)}$	Reactive power flowing from bus k to l on the c-th circuit between them
W^P_k	Real power flowing from bus k to bus k
W^Q_k	Reactive power flowing from bus k to bus k
P_k	Net real power injection into the network at bus k
Q_k	Net reactive power injection into the network at bus k

above derivations. Their meanings depend on the number of subscripts that go with them. Take P for example, P_k represents the real power output (injected into the network) from bus k, $P_{(k,l)}$ is the real power flowing from bus k to bus l from all circuits between them (or no circuit if $k = l$), and $P_{(k,l,c)}, k \neq l$ is the real power flowing from bus k to bus l along the c-th circuit between k and l. The following relations hold,

$$P_k = \sum_{l \in \mathit{BUS}} P_{(k,l)} = \sum_{\substack{l \in \mathit{BUS} \setminus \{k\} \\ c \in \mathit{CIR}}} P_{(k,l,c)} + P_{(k,k)} \tag{1.29}$$

The same interpretation goes to Q. Symbol overloading is not a standard feature for modeling languages such as GAMS, and the bypass is to use more symbols. Table 1.5 and the following equations will do this.

For each line $(k, l, c) \in \mathit{Line}$,

$$\begin{aligned} Z^P_{(k,l,c)} = (|V_k||V_l|/t_{(k,l,c)})(&-G_{(k,l,c)} \cos(\delta_k - \delta_l - \alpha_{(k,l,c)}) \\ &- B_{(k,l,c)} \sin(\delta_k - \delta_l - \alpha_{(k,l,c)})) \end{aligned} \tag{1.30}$$

$$\begin{aligned} Z^P_{(l,k,c)} = (|V_l||V_k|/t_{(k,l,c)})(&-G_{(k,l,c)} \cos(\delta_l - \delta_k + \alpha_{(k,l,c)}) \\ &- B_{(k,l,c)} \sin(\delta_l - \delta_k + \alpha_{(k,l,c)})) \end{aligned} \tag{1.31}$$

$$Z^{Q}_{(k,l,c)} = (|V_k||V_l|/t_{(k,l,c)})(B_{(k,l,c)}\cos(\delta_k - \delta_l - \alpha_{(k,l,c)}) - G_{(k,l,c)}\sin(\delta_k - \delta_l - \alpha_{(k,l,c)})) \tag{1.32}$$

$$Z^{Q}_{(l,k,c)} = (|V_l||V_k|/t_{(k,l,c)})(B_{(k,l,c)}\cos(\delta_l - \delta_k + \alpha_{(k,l,c)}) - G_{(k,l,c)}\sin(\delta_l - \delta_k + \alpha_{(k,l,c)})) \tag{1.33}$$

For each bus $k \in BUS$,

$$W^{P}_{k} = |V_k|^2(\sum_{\substack{j\in BUS \\ c\in CIR}} G_{(j,k,c)} + \sum_{\substack{j\in BUS \\ c\in CIR}} \frac{G_{(k,j,c)}}{t^2_{(k,j,c)}} + Y^{real}_{k}) \tag{1.34}$$

$$W^{Q}_{k} = -|V_k|^2(\sum_{\substack{j\in BUS \\ c\in CIR}} (B_{(j,k,c)} + bb_{(j,k,c)}/2) + \sum_{\substack{j\in BUS \\ c\in CIR}} \frac{B_{(k,j,c)} + bb_{(k,j,c)}/2}{t^2_{(k,j,c)}} + Y^{imag}_{k}) \tag{1.35}$$

$$P_k = \sum_{(k,l,c)\in Line} Z^{P}_{(k,l,c)} + \sum_{(l,k,c)\in Line} Z^{P}_{(k,l,c)} + W^{P}_{k} \tag{1.36}$$

$$Q_k = \sum_{(k,l,c)\in Line} Z^{Q}_{(k,l,c)} + \sum_{(l,k,c)\in Line} Z^{Q}_{(k,l,c)} + W^{Q}_{k} \tag{1.37}$$

Suppose that there are m lines and n buses, i.e., $|Line| = m$ and $|BUS| = n$, equations (28) to (35) represent a total of $4m + 4n$ equations and contain $4m + 6n$ variables. Remember that for each bus k, two of the four quantities $|V_k|$, δ_k, P_k and Q_k are fixed, which reduces the number of unknowns by $2n$. Again, we have a square system.

Line losses are readily computed. Let $l^{P}_{(k,l,c)}$ and $l^{Q}_{(k,l,c)}$ be the losses of real and reactive power, respectively, on line (k, l, c). The following relations hold,

$$l^{P}_{(k,l,c)} = Z^{P}_{(k,l,c)} + Z^{P}_{(l,k,c)} \tag{1.38}$$

$$l^{Q}_{(k,l,c)} = Z^{Q}_{(k,l,c)} + Z^{Q}_{(l,k,c)} \tag{1.39}$$

1.4 DC Approximation

Nonlinear nonconvex models are hard to solve, intractable, and the local solution usually depends on the starting point. In many circumstances such as in the real-time dispatch of generating resources, power flow models or other models built upon the power flow equations need to be solved frequently and reliably. Such requirements call for a set of linear equations that approximate the nonlinear behavior of the power system so as to substitute for the nonlinear equations in the power flow model. Models of this type are called DC power flow models, whereas in contrast the original nonlinear models are often called AC power flow models. Stott et al. (2009) provide a review of DC power flow models. Common approximations and assumptions made in DC power flow models include:

- Completely ignore the power balance equations for reactive power.
- Assume that all voltage magnitudes are identically one per unit, i.e., set all $|V_k|$'s to one.
- Ignore all line losses, i.e., set the resistance $r_{(k,l,c)}$ to zero for each line (k, l, c).
- Ignore tap dependence in the transformer reactance, i.e., set $t_{(k,l,c)} = 1$ and $\alpha_{(k,l,c)} = 0$ for each line (k, l, c).
- Assume that the voltage angle difference $\delta_k - \delta_l$ across any line (k, l, c) is sufficiently small so that $\cos(\delta_k - \delta_l) \approx 1$ and $\sin(\delta_k - \delta_l) \approx \delta_k - \delta_l$.

These assumptions eliminate many variables and equations compared to the AC power flow models, resulting in a set of strikingly simple linear equations.

For each line $(k, l, c) \in \mathtt{Line}$,

$$Z^P_{(k,l,c)} = B_{(k,l,c)}(\delta_l - \delta_k) \tag{1.40}$$

where $B_{(k,l,c)} = -1/(x_{k,l,c}t_{k,l,c})$ (the line susceptance taking account of the effect of the transformer ratio, see Table 1.2 for notation; set $t_{k,l,c} = 1$ if transformer tap is ignored), and for each bus $k \in \mathtt{BUS}$,

$$P_k = \sum_{(k,l,c)\in Line} Z^P_{(k,l,c)} + \sum_{(l,k,c)\in Line} Z^P_{(k,l,c)} \tag{1.41}$$

Remember that for a bus k, either δ_k or P_k has a fixed value. Therefore, the DC model remains a square system with $(m + n)$ equations and the same number of unknowns.

A vector format of the DC power flow equations is often used in the literature. Denote the set of network arcs by $\mathcal{A} = \{(k, l, c)|(k, l, c) \in \text{LINE or } (l, k, c) \in \text{LINE}\}$. Let A be the arc-bus incidence matrix of dimension $|\mathcal{A}| \times |\mathcal{B}|$. Nonzero entries of A are given by

$$A_{in} = \begin{cases} -1, & \text{if arc } i \text{ originates from bus } n \\ 1, & \text{if arc } i \text{ points to bus } n \end{cases}$$

Let B be an $|\mathcal{A}| \times |\mathcal{A}|$ diagonal matrix, the B_{ii} entry of which is the susceptance of arc $i \in \mathcal{A}$ (for example, if $i = (k, l, c)$, then $B_{ii} = B_{k,l,c}$). Then the power flow equations (1.40) and (1.41) can be respectively rewritten as

$$Z = BA\delta \tag{1.42}$$

$$P = A^T Z \tag{1.43}$$

The incidence matrix A of a connected network has a rank number of one less than the number of nodes, i.e., rank(A) = $|\mathcal{B}| - 1$, which would make the above linear system under-determined. This issue can be mended by fixing the voltage angle at the swing bus to zero by

$$e_1^T \delta = 0 \tag{1.44}$$

where e_1 is a $|\mathcal{B}|$-by-1 vector with the first element (assuming the first bus is the swing bus) equal to 1 and all other elements equal to 0, i.e., $e_1 = [1\,0\,0\ldots0]^T$.

1.5 Economic Dispatch

Although power flow models may appear in various forms, they are essentially square systems of equations. In these systems, two of the four fundamental quantities relating to each bus are fixed and the other two are unknown. The goal of the power flow study is to solve these system of equations to obtain the unknown quantities. In contrast, the essence of the economic dispatch is to relax (unfix and set bounds) certain quantities that are otherwise fixed in the power flow model, set an objective function, and then solve the resulting optimization model. The quantities being relaxed are usually the real power outputs of the generating buses, which then become variables to be determined in the lowest cost way. Economic dispatch problem is also called optimal power flow (OPF) problem. Cain et al. (2012) provide a historical review of the OPF problem and formulations, with an emphasis on the potential cost savings of increased efficiency of the dispatch.

The objective function in an economic dispatch model typically captures the cost of production. There are two prevalent forms for the cost function, quadratic and piece-wise linear. The quadratic form is rooted in the fact that the heat rate of a traditional coal-fueled generator is a quadratic function of its MW output level. Most of the power systems test cases have quadratic generation costs. Let $\mathcal{GEN} \subset \mathcal{BUS}$, for a generating unit $k \in \mathcal{GEN}$, the quadratic cost function is determined by three parameters α_k, β_k and γ_k, and takes the following form,

$$H_k(P_k) = \beta_k P_k + \gamma_k P_k^2 \tag{1.45}$$

The piece-wise linear form of the cost function is mainly used in the ISOs' market clearing practice, where the generators submit to the ISOs their generation offers by break points, i.e., the (dollar/MW, MW output) pairs, which serve as the parameters for a piece-wise linear cost function. Convex piece-wise linear cost functions can be formulated by linear functions and inequalities, in at least two ways.

1. Let $S(k)$ be the set of linear "pieces" in the cost function of unit k, with $c^1_{k,s}$ and $c^0_{k,s}$ being the linear coefficient and the constant term, respectively, of

piece s. Assume that $\{c^1_{k,s}, s \in S(k)\}$ is in increasing order which ensures the convexity of the cost function, then the cost function, denoted here by $H_k(P_k)$, could be characterized by

$$H_k \geqslant c^1_{k,s} P_k + c^0_{k,s} \qquad \forall o \in S(k) \tag{1.46}$$

$$H_k \geqslant 0 \tag{1.47}$$

2. Let $O(k)$ be the set of offer segments of unit k, and for each $o \in O(k)$ the offer is given by a price-quantity pair $(c_{k,o}, \bar{P}_{k,o})$, indicating that for the increment of $\bar{P}_{k,o}$ MW output, the marginal cost is $c_{k,o}$ dollars/MW. Assume that $\{c_{k,o}, o \in O(k)\}$ is in increasing order to ensure the convexity of the cost function, and let the decision variable $P_{k,o}$ be the MW dispatched in segment o, then the total cost $H_k(P_k)$ can be characterized by the following constraints,

$$H_k(P_k) = \sum_{o \in O(k)} c_{k,o} P_{k,o} \tag{1.48}$$

$$P_{k,o} \leqslant \bar{P}_{k,o} \qquad \forall o \in O(k) \tag{1.49}$$

$$P_k = \sum_{o \in O(k)} P_{k,o} \tag{1.50}$$

Note that both of the two formulations rely on the fact that the cost H_k is being *minimized* in the objective.

Each generating unit $k \in GEN$ has a lower limit P^{min}_k and an upper limit P^{max}_k on its real power output, which corresponds to inequalities of the following type,

$$P^{min}_k \leqslant P_k \leqslant P^{max}_k \tag{1.51}$$

Reactive power outputs may be subject to similar inequalities. Depending on circumstances, there might be restrictions on voltage angles, line flows, line losses and so forth.

In terms of the real power, the nodal balance of injection and withdrawal is implied in the power flow equations, e.g., in (1.9), (1.29) or (1.41). However, many

power economics models such as economic dispatch and unit commitment, are based on the DC approximation of the network or not considering the network at all. It is a convention to symbolize the distinction between the generation (injection) and demand (withdrawal). In such contexts, P_k is taken as the (positive) generation at bus k, whereas a parameter D_k is designated as the demand (or negative generation) at bus k, and the power balance constraint is expressed as

$$P_k = D_k \qquad \forall k \in GEN \tag{1.52}$$

1.6 Unit Commitment

Regarding the economic dispatch model discussed above, one might question the role of the constant term α_k in the cost parameters, as well as the validity of setting P_k^{min} to any value other than zero. In effect, the parameter α_k represents a fixed cost that is incurred regardless of the output of the generator, which is also termed "no-load" cost. A positive P_k^{min} parameter indicates the generator must maintain that level of output as long as it is in the *on* state, which is also termed "Economic Minimum" in a generator's bid data. These parameters are inputs to the unit commitment problem. The demand varies hour by hour, accordingly the optimal dispatch of generators will also vary hour by hour. However, it is impractical and uneconomical to turn on and off a generator too frequently. Unit commitment is a task of choosing among the available generating units the ones that are to be "up and running" for a contiguous blocks of time during the scheduling horizon, so a temporal dimension is needed in its model.

Suppose the planning horizon is discretized by a set $T = \{1, 2, \ldots, |T|\}$ of $|T|$ periods. For example, the day-ahead market concerns the hourly unit commitment for the next 24 hours, so each period is a hour. A binary variable $y_{k,t} \in \{0, 1\}$ is used to indicate whether a generator k is committed in period t. Then (1.51) can be rewritten as follows,

$$P_k^{min} y_{k,t} \leqslant P_{k,t} \leqslant P_k^{max} y_{k,t} \tag{1.53}$$

Let C_k^{noload} represent the no-load cost, which is given by

$$C_k^{noload} = \sum_{t \in T} \alpha_k y_{k,t} \tag{1.54}$$

A committed unit k in hour t may not be running at its full capacity, and the residual generating capability defined by

$$P_{k,t}^{res} = \max\{P_k^{max} y_{k,t} - P_{k,t}, 0\} \tag{1.55}$$

contributes to the system's spinning reserve, the generating capacity available to dispatch within a short period of time in case of supply disruption or demand surges. The system spinning reserve requirement is then written as

$$\sum_{k \in GEN} P_{k,t}^{res} \geqslant P_t^R \tag{1.56}$$

where P_t^R is a piece of data denoting the minimum required reserve amount during hour t.

Unit commitment problem is centered on the recognition that the startup and shutdown processes of most thermal based generating units are slow and costly, which deserves explicit treatment in a model. In general, there are minimum up/-down time constraints, ramping constraints and startup cost terms to be considered.

The minimum uptime (or downtime) constraints require that once a unit is committed (or uncommitted), it has to remain committed (or uncommitted) for a minimum of τ_k^{Up} (or τ_k^{Down}) hours. Suppose that it costs c_k^s dollars per startup. Two formulations are provided below.

1. If a unit k is turned on in hour t, then the set of hours during which it must be up and running is

$$T_{k,t}^{Up} = \{t_1 \in T | t + 1 \leqslant t_1 \leqslant t + \tau_k^{Up} - 1\} \tag{1.57}$$

Similarly, if a unit k is shut down in hour t, then the set of hours during which

it must remain offline is

$$T_{k,t}^{\text{Down}} = \{t_1 \in T | t+1 \leqslant t_1 \leqslant t + \tau_k^{\text{Down}} - 1\} \tag{1.58}$$

Note that $T_{k,t}^{\text{Up}}$ and $T_{k,t}^{\text{Down}}$ may be empty. With these two sets defined for each $k \in GEN$ and $t \in T$, the minimum up/down time constraints can be written as

$$y_{k,t} - y_{k,t-1} \leqslant y_{k,t_1} \qquad \forall t \in T\backslash\{1\}, t_1 \in T_{k,t}^{\text{Up}} \tag{1.59}$$

$$y_{k,t-1} - y_{k,t} \leqslant 1 - y_{k,t_1} \qquad \forall t \in T\backslash\{1\}, t_1 \in T_{k,t}^{\text{Down}} \tag{1.60}$$

The startup cost C_k^s of unit k during the entire horizon is

$$C_k^s = \sum_{t \in T} c_k^s (y_{k,t} - y_{k,t-1}) \tag{1.61}$$

2. Let binary variables $y_{k,t}^{\text{start}}$ and $y_{k,t}^{\text{shut}}$ indicate that unit k is started up and shut down in hour t, respectively. Their relations with $y_{k,t}$, as well as the minimum up and down time constraints are then expressed by the following constraints.

$$y_{k,t}^{\text{start}} \geqslant y_{k,t} - y_{k,t-1} \qquad \forall t \tag{1.62}$$

$$y_{k,t}^{\text{shut}} \geqslant y_{k,t-1} - y_{k,t} \qquad \forall t \tag{1.63}$$

$$-\sum_{t'=t-\tau_k^{\text{Up}}+1}^{t} y_{k,t'}^{\text{start}} + u_{k,t} \geqslant 0 \qquad \forall t \tag{1.64}$$

$$-\sum_{t'=t-\tau_k^{\text{Up}}+1}^{t} y_{k,t'}^{\text{shut}} - u_{k,t} \geqslant -1 \qquad \forall t \tag{1.65}$$

and the startup cost is

$$C_k^s = \sum_{t \in T} c_k^s y_{k,t}^{\text{start}} \tag{1.66}$$

Ramping constraints limit the magnitude of change in a unit's output level

between successive hours. Typical ramping constraints take the form

$$P_{k,t-1} - \Delta_k^{\text{Down}} \leqslant P_{k,t} \leqslant P_{k,t-1} + \Delta_k^{\text{Up}} \qquad \forall t \in T\backslash\{1\} \tag{1.67}$$

where Δ_k^{Up} and Δ_k^{Down} are the ramp-up and ramp-down rates (MW/hour), respectively, of the generating unit k.

The exclusion of the case $t = 1$ in the constraints (1.59), (1.60) and (1.67) hints a caveat: the model isolates the current planning horizon (e.g., a 24-hour period) from the chain of time, and assume a clean initial state, as well as a worry-free ending state. For example, the model could freely set $y_{k,1} = 1$ without considering whether the unit k has fully "cooled down" (been offline for τ_g^{Down} hours) since the last shutdown, or it could turn on a unit in the last hour of the day not worrying about the minimum uptime of constraints extending to the next day.

To mend this loophole, one could save the relevant information derived from the solution and use it to form the initial state constraints for the next run. In particular, the following data can be easily updated after each run:

- $U1$ and $D1$, subsets of GEN, the members of which must be Up and Down, respectively, in the first hour of the next run
- H_k, the number of hours the first hour commitment state (up or down, as indicated by $U1$ and $D1$) must last in the next run
- P_k^L, the last hour output level of unit k in the current run, i.e., $P_k^L = P_{k,24}$

With these extra data, the following constraints could be added to the model.

$$y_{k,t} = 1 \qquad \forall k \in U1 \tag{1.68}$$

$$t \leqslant H_k y_{k,t} = 0 \qquad \forall k \in D1, t \leqslant H_k \tag{1.69}$$

$$P_k^L - \Delta_k^{\text{Down}} \leqslant P_{k,1} \leqslant P_k^L + \Delta_k^{\text{Down}} \qquad \forall k \tag{1.70}$$

And the total start-up cost of unit k is updated as

$$C_k^S = \sum_{t \in T \setminus \{1\}} \alpha_k^S \max\{y_{k,t} - y_{k,t-1}, 0\} + 1_{\{k \notin U1\}} \alpha_k^S y_{k,1} \tag{1.71}$$

1.7 Market Clearing Price and Locational Marginal Price

The U.S. electricity market has undergone dramatic changes over the last two decades. Historically, electricity was supplied by vertically integrated utility companies subject to the cost-of-service regulation. Essentially, it was the suppliers who determined the price and the consumers who determined the transaction quantity. The contemporary wholesale market design is predicated on bid-based, competitive participation of both suppliers and demanders. Due to the characteristics of electric energy and its marketplace over the grid, centralized short-term (e.g., day-ahead) resource planning and real-time dispatch and control are indispensable to facilitate the competitive market. Independent System Operators (ISO) or Regional Transmission Organizations (RTO) are typically encharged of these responsibilities, under the regulation of Federal Energy Regulatory Commission (FERC). In its market orchestration, an ISO/RTO's statutory objective is maximizing the social welfare subject to resource and security constraints.

In the competitive market, the *market clearing price* (MCP), as well as the transaction quantity, is now determined by the market equilibrium, in economic terms, the intersection of the upward sloping supply curve and the downward sloping demand curve. When these two curves are accurate and unmanipulated, the resulting equilibrium maximizes the social welfare.

The *locational marginal price* (LMP) is defined as the cost of delivering the last unit (MW) of real power to a network node, at the current optimal dispatch solution. Mathematically, LMP is the optimal Lagrangian multiplier (also known as dual variable) corresponding to the real power balance equation, e.g., equation (1.9), (1.29), (1.41) or (1.52) depending on the formulation, in the economic dispatch

model. In an ideal case where the power flow on transmission lines were within the lines' thermal limits (no congestion) and the lines were conductive enough to have negligible resistance (no loss), the LMPs would be the same for all buses. But such ideal cases are uncommon in reality, which explains why the "location" matters. The congestion and loss components in the LMP differ by locations (buses) and reflect the relative difficulty and inefficiency to deliver power to a network node.

It is important to note that MCP and LMP are not equivalent by definition. However, in today's ISO- and RTO- run markets, LMP is widely used as a substitute for MCP. This gives rise to a myriad of efficiency, fairness and equity issues, some selected ones of which will be discussed in later chapters.

1.8 Data Sources and Formats

One of the most referenced data sources in power flow research is the Power Systems Test Case Archive at the University of Washington, see Christie (1993). The archive contains five power flow test cases named by the number of buses, including 14-bus, 30-bus, 57-bus, 118-bus and 300-bus cases. The data in these cases are believed to be representative of actual power systems of similar sizes. In particular, the cases of 14-bus up to 118-bus are portions of the American Electric Power System in the Midwestern U.S. as of the early 1960's, and the 300-bus case was developed in 1993 by an IEEE Test Systems Task Force. The data are stored in the IEEE Common Data Format (CDF), a text based file format inherited from the punch card era. Among other restrictions, CDF circumscribes the decimal length of each field or column, which to some extent limits the precision; furthermore, CDF contains neither the generator cost information nor the inter-temporal characteristics such as minimum up/down time and ramping rates, which are needed in many power economics models.

Matpower, a package of Matlab M-files for solving power flow and optimal power flow problems (Zimmerman et al., 2011), uses the Matlab way to store data. It encodes all the pertinent data of a case in a single struct and the users can edit the data as plain text. Some fields in the CDF are obsoleted and other fields such as

the generator cost parameters are added. Comparatively speaking, the Matpower case format is more advanced and more convenient to work with than CDF, but it is not an ultimate format without any limitation. The dynamics of the electricity market often require the model to handle demand forecasts in finer time scales (which means there needs to be a time axis in the demand data structure), or treat the demands as stochastic quantities rather than deterministic ones. Likewise, the generator cost parameters also come in various forms in practice. Different ancillary and security parameters and constraints may also need accommodations in unit commitment models. None of these are reflected in the Matpower case format.

To facilitate research, FERC publishes in its online *eLibrary* some sets of the market input data. Each set is of industrial scale (e.g., more than 1000 generators, spanning 24 hours) and includes generator offers, demand forecasts, demand bids, demand response offers and virtual bids. The data are derived from the PJM RTO's market operation data. These data, as well as a documentation, are available to the public and can be found on the FERC website at

http://www.ferc.gov/industries/electric/indus-act/market-planning/rto-commit-test.asp

However, the underlying network data are considered Critical Energy Infrastructure Information (CEII) and inaccessible to the public.

The design of a common data protocol is imperative for promoting an open and efficient research environment. Apart from the advantage in data exchange, a common protocol also helps present the problem inclusive of all facets, so researchers are aware of the big picture while focusing on certain aspects of the problem. At the same time, assumptions or simplifications made in a particular research work can be more explicitly exposed to other researchers. The idea of building hierarchical models to facilitate the planning and operation of power systems on multiple scales is elaborated in Ferris (2011).

1.9 GAMS Model Suite

The power flow models, economic dispatch (ACOPF) models and the unit commitment models described in the previous sections are implemented in GAMS. Table 4 lists the model files and auxiliary files currently in the suite.

Most models use the Matpower case data as the source data. Specifically, a Matpower case data file is a M-file with the name caseXX.m, where XX is a number indicating the number of buses in the case. For example, case14.m is the data file for the 14-bus case and case300.m is the data file for the 300 bus case. The following paragraphs will use the 14-bus case to illustrate the usage of files in the suite. Note that togdx.m uses the Matlab-GAMS interfacing utility called GDXMRW which requires separate installation. For how to install GDXMRW, refer to Ferris et al. (2011).

Unless the file case14.gdx has already been generated, the following steps are needed to generate it.

1. In Matlab, make sure case14.m and togdx.m are both included in the path, then execute

   ```
   >> togdx(case14,'case14raw');
   ```

   This will generate the case14raw.gdx in the current directory.

2. In the GAMS IDE or the DOS mode, run GAMS command

   ```
   gams raw2gdx --raw=case14raw --out=case14
   ```

   This will generate the case14.gdx, which is readable by the model files.

Once case14.gdx is generated, one can pick a model and solve it for this case. For instance, to solve the economic dispatch model with the Y_{bus} formulation, run the GAMS command

```
gams ed1_mp --case=case14
```

The GAMS default NLP solver (often CONOPT) will be used. To solve the case with a different solver, PATHNLP for example, run

```
gams ed1_mp --case=case14 --nlp=pathnlp
```

The solutions are reported in the GAMS listing file ed1_mp.lst.

Two unit commitment models, uced.gms and uced2.gms, use the FERC data sets. Each data set consists of two Excel files, one for generator data and the other for demand data. For example, the 4012gen.xls and 4012demand.xls in the suite are obtained from FERC eLibrary, Docket Number AD10-12, ACCNNUM20120222-4012. The following steps are needed to run the model.

1. Upon executing the following command-line commands, 4012gen.gdx and 4012demand.gdx will be created.

   ```
   gdxxrw 4012gen.xls @extract_gen.txt
   gdxxrw 4012demand.xls @extract_demand.txt
   ```

2. Run uc_raw2gdx.gms (uc2_raw2gdx.gms) to generate model-ready GDX files for the model uced.gms (uced2.gms), for example, run

   ```
   gams uc_raw2gdx --genfile=4012gen --demandfile=4012demand --outfile=uc_data
   ```

   to generate uc_data.gdx file, and then run the model by

   ```
   gams uced --datafile=uc_data
   ```

Communication is the key to the success of any joint project, especially of the multidisciplinary ones. The suite of models provides a common basis for communication between the problem definer and the problem solver in power system related projects, and more generally serves as a bridge to convey the power economics problems and the computational challenges to the OR community.

Compared to Matlab, GAMS separates the core mathematical formulation from the solution technique in a much cleaner fashion, and is independent of the data

processing chores and solver technologies. Via the GDX utilities, the models are able to read in raw data from various sources and formats, and easy to customize for different aspects or detail levels of the problem. Powered by a stable of high-performance solvers integrated with GAMS, these models are expected to provide off-the-shelf solutions for industrial applications. Furthermore, attributed to the flexibility of GAMS language, the models and the shared data sources could enable coherent testing of new ideas.

As a first documentation of the GAMS model suite on power economics, this chapter has presented the problems and their mathematical formulation in the subscripted format (as opposed to the vector format) that is directly implementable in algebraic modeling languages. The notations and models documented here also serve as a basis for the subsequent chapters. For instance, the demand response model in Chapter 2, the unit commitment model used in Chapter 3 and 5 and the economic dispatch models discussed in Chapter 4, 6 and 7 can all find their root in this chapter.

1.10 A Case Study at ISO New England

ISO New England Inc. has three key missions: (1) Developing and administering the region's competitive wholesale electricity markets; (2) Overseeing the day-to-day operation of new England's electric power generation and transmission system; (3) Managing comprehensive regional power system planning. The unit commitment and dispatch model is at the core of two of them, clearing the market and maintaining the power systems operation. This section briefly reviews how these missions are modeled and implemented.

Physically, a *bus* is a metal bar[1] where several power lines end and connect to each other. The other end of the line with one end tieing to a bus may connect to a nearby generator that injects power into the bus, or to a nearby load[2] that withdraws power from the bus, or to another faraway bus to transmit power over a

[1] Three bars, to be exact, one for each phase of the three-phase AC power.

[2] A load here physically indicates a stepping-down transformer.

long distance. A line that links two buses is called a branch. Each piece of power injection/withdrawal equipment (e.g., a generator or a load), or an equivalent of it, tied to a bus is called a *node*. Bus, node and branch are the main components of a power network from the engineering viewpoint.

The network configuration, or topology, can be modified by a number of controlling components, such as switches and breakers. For example, a node can be connected to one bus or another in its vicinity, or disconnected from any of them. A bus can be disconnected from the transmission line that leads to another bus, or it can be separated into two buses when a breaker in its local configuration is open. However, a node is never connected to more than one bus simultaneously. In general, the mapping between buses and nodes may change over time. At any moment, a bus can have any number (including zero) of nodes attached to it, while a node can connect to at most one bus.

Nodes are geographically grouped into different LMP locations. Locations are not mutually exclusive in terms of the nodes they contain. As subsets of nodes, LMP locations can intersect, and even contain one another. For example, a reserve zone as an LMP location may encompass many switchyard-based LMP locations. The mapping between nodes and LMP locations rarely changes over time. Typically, each location contains at least one node and each node can belong to zero, one or multiple locations.

Any participant that submits energy bids (either buy or sell) in the day-ahead market is treated as a *resource* in the market models. A resource can, but does not have to, be backed by a physical point of injection (generator or load), hence a mapping between resources and nodes is irrelevant. Instead, a resource must specify the LMP location in which it bids into the market. All cleared resources in an LMP location will be subject to the same LMP. An LMP location can have any number (including zero) of resources, whereas the number usually changes from day to day and hour to hour. On the other hand, a resource may bid in different LMP locations for different time periods, but it cannot bid in multiple locations simultaneously.

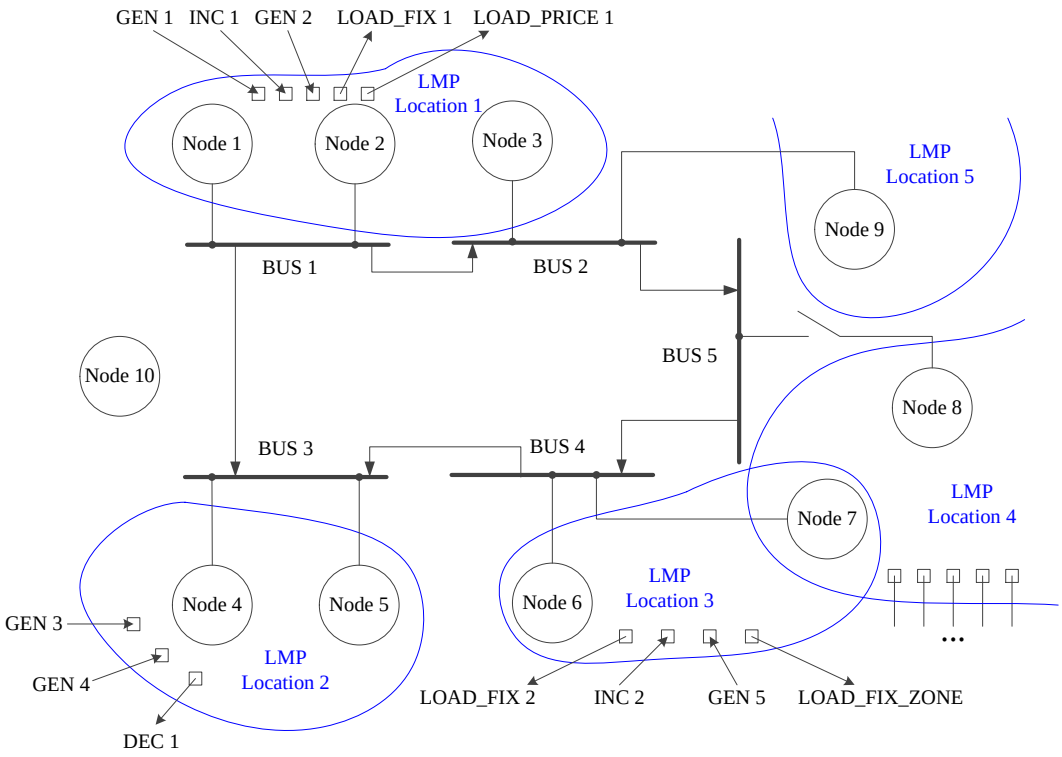


Figure 1.2: Relations of different modeling elements

Table 1.7: Counts of Different Modeling Elements in the ISO's System

Bus	Node	Resource	LMP Location	Branch	Bid Block	Hour
2,336	13,450	793	2,837	3,250	280	24

Resource bids typically come in two forms: fixed and price-sensitive. A fixed bid specifies the quantity of MWh energy to be supplied or demanded in the specified hour. Self-scheduled generation and fixed demand are examples of fixed bids. A price-sensitive bid consists of one or multiple blocks of MWh quantities, each accompanied by a bid price. Apart from the energy component, resources that are (verified to be) backed by physical generators can include unit commitment costs in their bids. Limited energy generation (LEG) requirements are also accepted in a generator's bid. Figure 1.2 illustrates the relation of different modeling elements

mentioned above and Table 1.7 demonstrates the size of the ISO's system in terms of component count.

Day-ahead Unit Commitment and Dispatch

The day-ahead (DA) market is a forward market for the delivery of energy in 24 hours. The market outcome consists of three parts: the cleared quantities of resources, the LMP and the day-ahead unit commitment schedules for generators.

The DA market clearing algorithm involves two optimization models, i.e., the security-constrained unit commitment (SCUC) model and the security-constrained economic dispatch (SCED) model. SCUC aims to find an optimal unit commitment solution that supports the market clearing, while SCED computes the optimal forward contracts (e.g., resources' buy/sell positions) and the LMPs. Both models take into account the transmission limitation as well as line contingencies. The transmission and contingency constraints in these models are supplied by the simultaneous feasibility test (SFT). We group the constraints for the ease of discussing the models.

UC constraints

These constraints only involve binary variables, primarily vIsOnline, vIsStartup and vIsShutdown. They include:

```
EQ_StartShut
EQ_MinUpTime
EQ_MinDownTime
EQ_StartUpState_1
EQ_StartUpState_2
EQ_StartUpState_3
EQ_PumpStorageCommitment
EQ_CC_ModeSelection
EQ_CC_ModeCalc_1
EQ_CC_ModeCalc_2
```

EQ_MaxStartups

Linking constraints

These constraints link a unit's commitment status and its energy output and reserve level. The ramping constraints belong to this category, as ramping also depends on the commitment status. When the linking constraints appear in the SCED model, the commitment variables are fixed according to the preceding UC solution.

EQ_RampUp
EQ_RampDown
EQ_RampUp_Startup
EQ_RampUp_DemandStartup
EQ_RampDown_DemandShutDown
EQ_EnergyMax
EQ_EnergyMin
EQ_TotalCapacity
EQ_ReserveCapacity

Market constraints

These constraints characterize the resource dispatch and market clearing mechanisms. The market clearing is conducted on the "transaction network" built on LMP locations, in lieu of the physical network of nodes and buses. The net energy sales in an LMP location (vNetInjection) is the signed sum of all resource energy outputs (vResourceEnergy) in this location, positive sign for supply resources and negative sign for demand resources[3]. The LMP is the shadow price of the constraint EQ_Energy_Balance. Constraints on the energy flow among LMP locations, which cause the LMPs to differ by location, are encapsulated in the proxy constraints. The transmission proxies are derived from the base-case network condition (i.e., no line

[3]Supply resources are GEN and INC, and demand resources include LOAD_FIX_ZONE, LOAD_FIX, LOAD_PRICE, LOAD_ARD, DEC and TRANSACTION

contingency) and the SFT proxies are derived from the N-1 contingency scenarios. The proxy constraints are identified by the SFT given the dispatch solution of the previous iteration. The use of proxy constraints greatly reduces the size and complexity of the SCUC and SCED models. Generic constraints are additional linear constraints derived from information not available in the data, oftentimes manually added by the system operator via the software interface.

```
**** Market Clearing ****
EQ_Resource_Output
EQ_Resource_Reserve
EQ_NetInjection
EQ_Energy_Balance
EQ_System_Losses
EQ_ReserveConstraints
EQ_TransactionsMW
EQ_MaxEnergy

**** Generic ****
EQ_GenericLHS
EQ_GenericConstraints_LE
EQ_GenericConstraints_GE
EQ_GenericConstraints_EQ

**** Transmission Proxy ****
EQ_TransmissionFlow
EQ_TransmissionConstraints

**** SFT Proxy ****
EQ_SFTFlow
EQ_SFTConstraintsP
EQ_SFTConstraintsN
```

DC power flow constraints

These constraints model the physical network using DC power flow equations. The net locational sales (vNetInjection) is allocated to nodes (as nodal injections) within the location by a predefined weight vector (pLMPFactors). Bus injection is the sum of the nodal injections of all nodes connected to that bus. These relations are instantiated in the equation EQ_BusBalance.

```
EQ_BusBalance
EQ_BranchFlow
EQ_ZBRBranch
EQ_DCFlowLimit_P
EQ_DCFlowLimit_N
```

The remaining equations define the cost segments and objective functions for the models.

```
EQ_Total_Objective
EQ_Objective_Cost
EQ_Objective_PenaltyCost
EQ_CalcStartUpCost
```

DA market clearing process

The DA market clearing algorithm is an iterative process. The day-ahead UC solution u is obtained by solving the SCUC model, which includes the UC, Linking, Market (in which the proxy constraints are empty) and DC power flow constraints. Given u, the SCED model is solved for each hour, individually and sequentially. The SCED solution P_t of hour t will be used as input to the intertemporal (e.g., ramping) constraints in the SCED model for hour $t + 1$. Below, we discuss in detail the solution process of SCED for an hour t. The time index t is present in all variables and equations in the model, and will be omitted for succinctness. The SCED model minimizes the total as-bid energy cost (supply cost minus demand

benefit) subject to the Linking constraints with fixed u, the Market constraints with a growing set of proxies, and the DC power flow constraints.

DC power flow constraints consists of the following equations and inequalities[4],

$$\text{EQ\_BranchFlow:} \qquad Z = BA\delta \tag{1.72}$$

$$\text{EQ\_BusBalance:} \qquad P = A^T Z \tag{1.73}$$

$$\text{EQ\_DCFlowLimit\_P:} \qquad Z \leqslant \bar{Z} \tag{1.74}$$

$$\text{EQ\_DCFlowLimit\_N:} \qquad Z \geqslant -\bar{Z} \tag{1.75}$$

In the above equations, variables P, δ and Z are the bus injection, bus voltage angle (vPARAngleUC) and branch flow (vBranchFlow), respectively. A is the branch-bus incidence matrix. Nonzero entries of A are given by

$$A_{in} = \begin{cases} -1, & \text{if branch } i \text{ originates from bus } n \\ 1, & \text{if branch } i \text{ points to bus } n \end{cases}$$

and B is a diagonal matrix, the B_{ii} entry of which is the susceptance of branch i. Suppose $(A^T BA)$ is invertible[5], we have

$$\delta = (A^T BA)^{-1} P$$

and subsequently

$$Z = BA(A^T BA)^{-1} P$$

The SCED solution process for a given hour begins by solving the SCED model to obtain the bus injection P^*, which satisfies the base-case DC power flow con-

[4]EQ_ZBRBranch handles a special case and will not be discussed here.

[5]The incidence matrix A of a connected network has a rank number one less than the number of buses, so $(A^T BA)$ is not invertible. This can be amended by replacing a row (corresponding to the swing bus) by $e = [1\,0\,0\,\ldots 0]$ and the corresponding element in P by 0. The resulting system is determined and $(A^T BA)$ invertible.

straints. This solution P^* is then fed into the SFT to identify additional transmission constraints in cases of a line contingency.

The SFT is a what-if analysis that answers the question: would P^* satisfy all the line limit constraints if a certain branch (in a predefined subset of branches, called the contingency list) were tripped? Denote the incidence and susceptance matrices under the contingency scenario by $\tilde{A}$ and $\tilde{B}$, respectively. Given the injection P^*, the new branch flow $\tilde{Z}$ is computed by

$$\tilde{Z} = \tilde{B}\tilde{A}(\tilde{A}^T\tilde{B}\tilde{A})^{-1}P^*$$

If all branch flows are within limits, i.e., $\tilde{Z}$ satisfies (1.74) and (1.75), then P^* is good for this contingency. Otherwise, suppose the positive-direction flow limits (1.74) are violated at a subset I of branches, and the negative-direction flow limits (1.75) are violated at a subset J of branches, then the following SFT proxy constraints will be generated for the SCED model.

$$[\tilde{B}\tilde{A}(\tilde{A}^T\tilde{B}\tilde{A})^{-1}]_{I.}P \leqslant \bar{Z}_I \tag{1.76}$$

$$[\tilde{B}\tilde{A}(\tilde{A}^T\tilde{B}\tilde{A})^{-1}]_{J.}P \geqslant -\bar{Z}_J \tag{1.77}$$

The above SFT analysis is iterated for all contingency scenarios, each having different $\tilde{A}$ and $\tilde{B}$, and each identifying possible proxy constraints (1.76) and (1.77). Once the SFT process is completed, all identified proxy constraints will be added back to the SCED model. The updated SCED model is then solved again to find a new P^*. The process repeats until a solution P^* that does not violate any contingency flow limit is found or a predefined iteration limit is reached. At the end, the dispatch solution and the LMPs are returned, for the hour t at discussion. Chapter 6 will discuss the SFT analysis in more detail.

The day-ahead market is settled once the SCED process is completed sequentially for all hours.

Table 1.6: List of Files in the Model Suite

File Name	Description	Data Source
pf1_mp.gms	Power flow model with rectangular-coordinate (Y_{bus}) formulation	Matpower
pf2_mp.gms	Power flow model with polar-coordinate formulation	Matpower
ed1_mp.gms	ACOPF model with rectangular-coordinate (Y_{bus}) formulation	Matpower
ed2_mp.gms	ACOPF model with polar-coordinate formulation	Matpower
dc_mp.gms	DCOPF model	Matpower
uced.gms	Unit commitment economic dispatch model 1	FERC eLibrary
uced2.gms	Unit commitment economic dispatch model 2	FERC eLibrary
togdx.m	Convert the Matpower case format to GDX format	
raw2gdx.gms	Convert the GDX from togdx.m to model-ready GDX file	
extract_gen.txt	GDXXRW script to convert generator data to GDX	
extract_dem.txt	GDXXRW script to convert demand data to GDX	
uc_raw2gdx.gms	Create model-ready GDX file for uced.gms	
uc2_raw2gdx.gms	Create model-ready GDX file for uced2.gms	

2 MODELING DEMAND RESPONSE FOR FERC ORDER 745

2.1 Introduction

On March 15, 2011, Federal Energy Regulatory Commission issued Order No. 745 (FERC, 2011), the Final Rule settling the yearlong rule-making debate on how to compensate demand response resources that participate in an organized wholesale energy market (i.e., the day-ahead and real-time energy markets) administered by a Regional Transmission Organization (RTO) or an Independent System Operator (ISO).

According to the Final Rule, demand response means a reduction in the consumption of electric energy by customers from their expected consumption in response to an increase in the price of electric energy, or to incentive payments designed to induce lower consumption of electric energy. A demand response resource means any dispatchable entity that is capable of providing demand response. For example, a manufacturing plant that is capable of suspending its energy-intensive process when called upon by the ISO during hours of high prices, can be considered as a demand response resource.

The Order requires that "when a demand response resource participating in an organized wholesale energy market administered by an RTO or ISO has the capability to balance supply and demand as an alternative to a generation resource and when dispatch of that demand response resource is cost-effective as determined by a net benefits test, that demand response resource must be compensated for the service it provides to the energy market at the market price for energy, referred to as the locational marginal price (LMP)".

The Order on one hand dictates very specifically that the demand response (DR) resource, when dispatched, be compensated at the LMP, and on the other hand leaves the determination of the cost-effectiveness condition and the implementation of the demand response dispatch to the RTOs and ISOs, to which end it outlines a two-stage action plan with clear time lines. The Order states:

> *By July 22, 2011, each RTO and ISO should develop a mechanism as an approximation to determine a price level at which the dispatch of demand response resources will be cost-effective. ...*
>
> *By September 21, 2012, each RTO and ISO should undertake a study examining the requirements for and impacts of implementing a dynamic approach which incorporates the billing unit effect in the dispatch algorithm to determine when paying demand response resources the LMP results in net benefits to customers in both the day-ahead and real-time energy markets. ...*

As a response to the call in FERC Order 745, this chapter investigates the above-mentioned dynamic dispatch approach that incorporates the demand response dispatch and compensation rules as described in the Order. The remainder of this section introduces the background of the subject, in particular, the motivation for promoting demand response in the wholesale market and the key elements of the Order that pose challenges for implementation thus motivate our work. Section 2.2 presents our main contribution: modeling the demand response problem as a bi-level optimization program. In Section 2.3, we develop two alternative methods that work under different conditions, and use them in Section 2.4 to validate the bi-level model via experiments. Section 2.5 presents some useful observations about the demand response market and the dispatching operations by applying the model and its variants to various data cases. Section 2.6 summarizes the chapter and gives the plan for future work.

All occurrences of the term ISO in the rest of the chapter should be taken as ISO/RTO. Electric power means the real (instead of the reactive or complex) power. Node, bus and location mean the same thing. We use the units MW and dollars/MW to measure the power and the price, which can be regarded as equivalent to using MWh and dollars/MWh.

Motivation of Demand Response

The motive, if not the action, of consumers' response to electricity prices has existed since spot pricing was adopted in the electricity market. In the book by Schweppe

et al. (1988), two essential types of consumers' response were identified: reduce usage if the price in a given hour is high, and reschedule usage if the price is high in some hours and low in other hours. Another early work by Daryanian et al. (1989) studied how a storage-type consumer could respond to the spot pricing of electricity by determining an optimal schedule of electricity usage given a predetermined electricity price schedule.

However, demand response is not an inherent element of competitive energy markets; rather, it is a recourse measure to account for the market imperfection caused by some unusual characteristics of the underlying commodity, electric energy. Specifically, electricity supply and demand over the grid must match closely at every instant in time, so the market must clear in real time. Such frequent market clearing cannot happen naturally (at the discretion of the "invisible hand"), but requires coordination of a central dispatcher, namely, an ISO. The ISO attempts to clear the market efficiently, i.e., maximize the social welfare, and therefore needs to know explicitly how much the suppliers and demanders value each increment of supply and demand (the supply and demand curves, as depicted in Figure 2.1). This information is conveyed to the ISO via supply offers and demand bids, see, for example, Arroyo and Conejo (2002) and Su and Kirschen (2009).

While the supply curve is usually easy to estimate, it is difficult for the majority of the demand-side to identify the marginal value of electricity and hence bid a meaningful demand curve, see Kirschen (2003). We observe that such difficulty is not unique to electricity, but is present in many other commodity markets, e.g., markets of farm produces, consumer products and the like. However, those markets do not require instantaneous clearing, therefore, consumers' response to price signals, an alternative expression of the demand curve, can have enough time to settle in and keep market equilibrium at the efficient point (Mankiw, 2011). This is not the case for the ISO-run energy markets over the grid. In the absence of an accurate demand curve, social welfare cannot be accurately characterized, let alone optimized.

It is commonly recognized that the demand elasticity, or the willingness and ability of the demand-side to reduce consumption in times of high prices, is actually

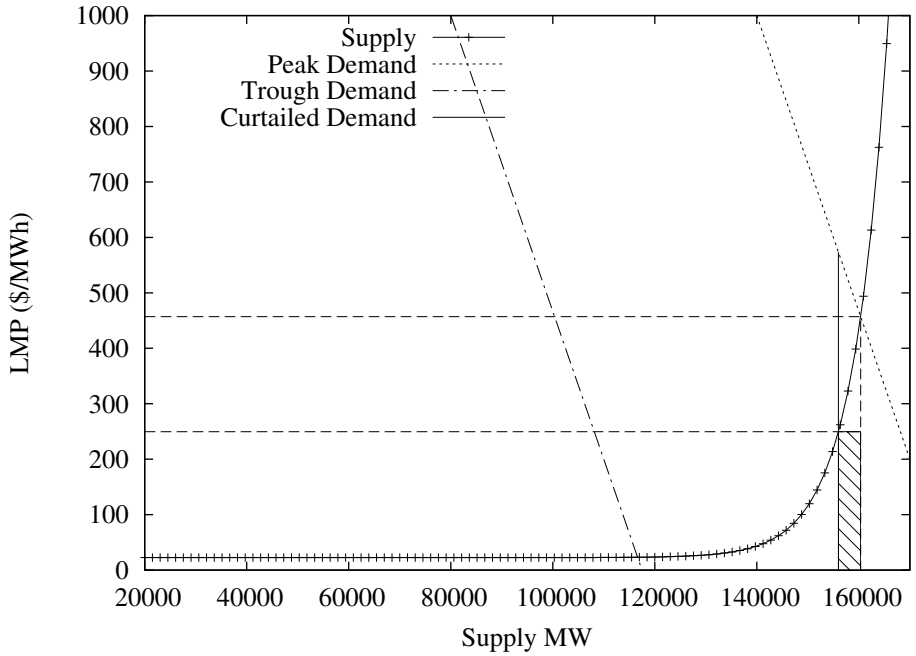


Figure 2.1: Electricity supply and demand curves

higher than what is perceived by the ISO from the demand bids, see, for example, Joskow (2001), Wellinghoff and Morenoff (2007) and Chao (2010). The consequence of under-perceiving the demand elasticity is illustrated in Figure 2.2. With the true demand and supply curves, the market equilibrium is at point E with supplier's surplus being the area of COE and the demander's surplus being the area of ACE. However, if the elasticity of the demand was under estimated, as represented by the Perceived Demand curve in the figure, the market would operate at point G, resulting in a supplier's surplus of BOG and a demander's surplus of ABD minus DHG. The net effect is a surplus transfer from the demander to the supplier by the amount BCEG and a social welfare loss of EHG.

For clarity, we make two technical points about the figure: (1) even though the elasticity varies along a linear demand curve, it is easy to see that at any given price level or demand level, the elasticity of the Perceived Demand curve is always

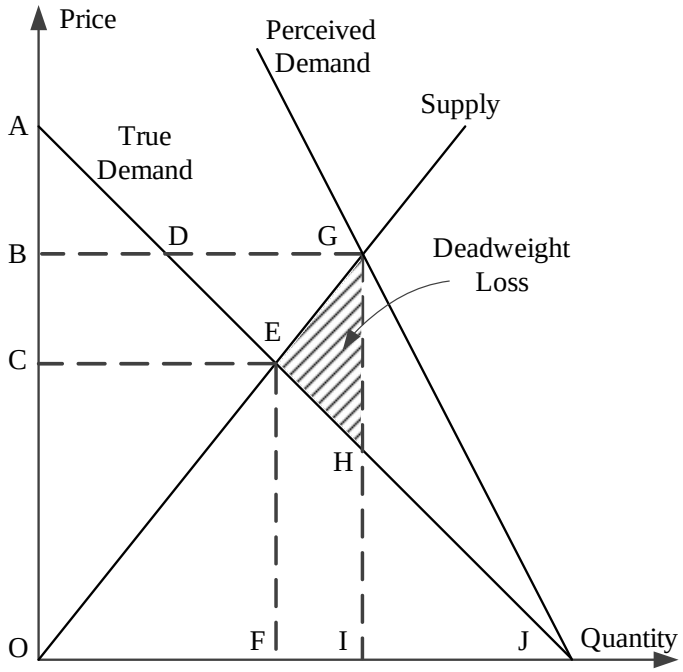


Figure 2.2: Market inefficiency caused by imperfect demand information

lower than the True Demand curve, therefore, using linear demand curves suffices for the comparison; (2) the two demand curves intersect at the zero price level, because when a commodity is free (at zero price), it is reasonable to assume that the consumption will be at the maximum consumption capacity or the most convenient level, which is the same for different demand curves.

A demand response mechanism, if appropriately designed and implemented, can serve to overcome the market inefficiency, by migrating the market equilibrium point from G to E in Figure 2.2. This is illustrated in Figure 2.1. The supply curve is plotted by the formula

$$\text{LMP} = 4.017731^{0.000107x - 12.7911} + 22.54387$$

where x is the supply quantity in MW. This formula is published by PJM as its Modeled Supply Function for June, 2012. There is not a single demand curve

because the demand for electricity is practically cyclical and the demand curve shifts from left to right depending on the time of day, day of the week or week of the year. For demonstration purposes, the two (dotted and dot-dashed) slanted lines are fictitious demand curves to represent the peak and trough demand scenarios, respectively. A similar depiction can be found in Kirschen (2003). The curtailed demand line (the vertical line that stems downward from the middle of the peak demand curve) represents a fictitious example of demand response, corresponding to an amount measured by the distance from the vertical dashed line to this line. The demand response yields an LMP reduction from around \$460/MWh to about \$250/MWh. The shaded area is the compensation received by the curtailed demand as per the FERC Order. It can be seen that the effect of demand response on the market equilibrium is aligned with what is needed to correct the inefficiency as depicted in Figure 2.2. As a side effect, demand response can also thwart the "all but irresistible temptation for generators to manipulate the market, sending prices soaring" as depicted by Spees and Lave (2007). The following paragraph in FERC Order 745-A (on page 15) briefly summarizes the above point:

> *A properly functioning market should reflect both the willingness of sellers to sell at a price and the willingness of buyers to purchase at a price. In an RTO- or ISO-run market, however, buyers are generally unable to directly express their willingness to pay for a product at the price offered. ... RTOs and ISOs cannot isolate individual buyers' willingness to pay which results in extremely inelastic demand. Including demand response as a resource in RTO and ISO markets provides a way for buyers to indicate the price at which they are willing to stop consumption.*

This constitutes the economic justification and motivation of eliciting demand response in the market.

Understanding the FERC Order

The essence of FERC Order 745 (referred to as "the Order" subsequently) is the requirement that "when a demand response resource participating in an organized wholesale energy market administered by an RTO or ISO has the capability to balance supply and demand as an alternative to a generation resource and when dispatch of that demand response resource is cost-effective as determined by a net benefits test, that demand response resource must be compensated for the service it provides to the energy market at the market price for energy, referred to as the locational marginal price (LMP)".

While the DR compensation level is dictated in the Order, there remains four key questions to answer to implement a compliant and efficient DR program: (1) who makes the decision about when, and how much, to reduce consumption, the DR provider or the ISO? (2) Should DR providers be treated as energy sellers, in the same way as are generators, in the market clearing and LMP calculation process? (3) What is meant by "cost-effective" and what is the net benefit test? (4) What measure is in place to ensure economic efficiency in the Order context?

Conventional wisdom would regard demand response as consumers' voluntary action to curtail consumption to cut down energy bills during periods of high prices. This is not the same notion of DR discussed in the Order. The Order clearly refers demand response as a service procured by, and a dispatchable resource of, the ISO, which means that the dispatch decision, i.e., when, how much and which resources to dispatch, is at the hand of the ISO, not of the DR providers. A DR provider, on the other hand, need to inform the ISO (via bids) its capability and willingness to follow the ISO's dispatch.

For the second question, abundant evidence in the literature suggests that the answer is no. In short, a DR provider is not entitled to sell (in the energy market, day-ahead or real-time) the energy curtailed from its baseline consumption, without physically or contractually owning the baseline amount of energy, see Ruff (2002), Chao (2010), Borlick (2011) and Hogan (2010a). Instead, DR can be treated as a sale of the "consuming right" from certain consumers (DR provider)

to other consumers (the remaining load). In particular, as implied by the Order, the remaining consumers pay the DR provider to reduce consumption. When the supply curve is steep, such trades among the demand-side can be beneficial to all consumers, including DR providers who get compensation from the remaining load, and the remaining load who enjoys lower LMP. This trading of consuming right is done outside the energy market so there is not an issue about energy entitlement. From a modeler's perspective, the simultaneous clearing of DR and energy requires either an iterative process, or a hierarchical model. This is the main subject of study in this chapter and will be elaborated in subsequent sections.

The answer to the third question has been indicated in the Order. Specifically, the Order recognizes and stresses the "billing unit effect", a phenomenon that, depending on the change in LMP relative to the size of the energy market, dispatching demand response resources may result in an increased cost per unit (dollars/MW) to the remaining wholesale load associated with the decreased amount of load paying the bill. The Order states that billing unit effect should be avoided when an ISO dispatches the demand response.

A simple example is given in the Order (footnote 119 in FERC (2011)) to illustrate a cost-effective scenario of paying the demand response LMP, quoted as: "assume a market of 100 MW, with a current LMP of $50/MW without demand response, and an LMP of $40/MW if 5 MW of demand response were dispatched. Total payments to generators and load would be $4,000 with demand response compared to the previous $5,000. Even though, the reduced LMP is now being paid by less load, only 95 MW compared to 100 MW, the price paid by each remaining customer would decrease from $50/MW to $42.11/MW ($4,000/95). Therefore, the payment of LMP to demand resources is cost-effective."

Following the same reasoning, a cost-ineffective (due to the billing unit effect) scenario can be cooked up easily. For example, if the 5 MW of demand response were only able to reduce the LMP to $49/MW, then the price paid by the remaining customers would be $51.58/MW ($4900/95), an increase compared to not dispatching the demand response.

The LMP at a location is defined to be the cost of providing the next unit amount

of power to this location. In the payment rule, LMP is the price at which the ISO pays to the generator to buy the dispatched amount of power, or to the demand response resource to compensate the dispatched amount of reduction in consumption. The total cost of buying power and compensating the DR resources are shared among the actual consuming loads. The average price, AvgPrice, is thus defined by,

$$\text{AvgPrice} = \frac{\sum_k (g_k + r_k)\lambda_k}{\sum_k (d_k - r_k)}, \tag{2.1}$$

where g_k is the generation in MW, d_k is the pre-DR demand in MW, r_k is the demand response amount in MW, and λ_k is the LMP in \$/MW, all for node k. This definition of the AvgPrice is consistent with the idea implied in the billing unit effect discussion in the Order, therefore, it enables the determination of the DR cost-effectiveness in the same way as in the Order. Specifically, if the post-DR AvgPrice is lower than the pre-DR AvgPrice, then there is no billing unit effect and the DR dispatch decision (quantified by the r_k's) is cost-effective and vice versa.

The fourth question is critical for an economically sound DR program. As pointed out by Hogan (2009), "if demand response is improperly compensated, hoped-for increases in efficiency may not materialize, as either too much or too little demand response may be developed." Better than nothing, the Order mentions a price level or threshold such that when the market price exceeds this level, the dispatch of demand response will be considered. Note that the "market price" here is meant to be a single price across all locations. We believe that this price is best defined as the demand weighted average LMP across all nodes (short for AvgLMP), calculated by the following formula,

$$\text{AvgLMP} = \frac{\sum_k d_k \lambda_k}{\sum_k d_k}. \tag{2.2}$$

Again, the d_k used in the formula is the demand before the demand response amount is deducted, if there is any at node k. By using this formula, we assume that $\sum_k d_k > 0$, that is, the total demand in the network is always positive.

The ideal level of AvgLMP should be that of the market clearing point resulted

from a perfect-information scenario, i.e., level C in Figure 2.2. The determination of such a point is of great importance to social welfare, and is not an easy task in practice. We refer the readers to Walawalkar et al. (2007) for a dedicated research and case study on this topic.

2.2 Modeling the Demand Response

The demand response problem arises from the ISO's practice of clearing the energy market, where economic dispatch is at the center of this practice. Research on this topic abounds in the power systems literature. In the development of this work, we find Monticelli et al. (1987b), Momoh et al. (1999), Andersson (2008), Wang et al. (2007) and Dommel and Tinney (1968) useful for understanding the subject matter. We briefly develop the economic dispatch model and then proceed to the demand response modeling.

Preliminaries

In the modeled power network, there is a set $\mathcal{B}$ of buses (or nodes), which are further distinguished by two subsets, $\mathsf{GEN} \subset \mathcal{B}$ for generating buses and $\mathsf{LOAD} \subset \mathcal{B}$ for load buses. A generating bus is one attached with a generating unit so that it may inject electricity into the network. A load bus is one that has no generating capability and can only withdraw electricity from the network. Buses are interconnected by transmission lines. In some cases there are more than one lines connecting two buses, and each line is called a circuit. Let CIR denote the set of circuit numbers, then every transmission line (or arc in graph theory terminology) in the network can be uniquely identified by the triple (k, l, c), where $k < l \in \mathcal{B}$, $c \in \mathsf{CIR}$. Let $\mathcal{A}$ denote the set of all arcs in the network, and use the symbol a as a substitute for the arc triple (k, l, c) in subscripts when context allows. More notations are listed in Table 1, of which the upper half lists the parameters and the lower half lists the decision variables.

Table 2.1: Notations for the Economic Dispatch Model

b_a	Susceptance of the arc a
d_k	Demand at bus k
$\underline{g}_k, \bar{g}_k$	Lower and upper generation limits at bus k
$\underline{z}_a, \bar{z}_a$	Lower and upper flow limits on arc a
α_k, β_k	Generation cost parameters of bus k
g_k	Generation at bus k
δ_k	Voltage angle of bus k
z_a	Power flow from k to l on arc a

Let $g \in \mathbb{R}^{|\mathcal{B}|}$, $z \in \mathbb{R}^{|\mathcal{A}|}$ and $\delta \in \mathbb{R}^{|\mathcal{B}|}$ be the vectors formed by the scalar variables g_k, z_a and δ_k, respectively. In the remainder of this chapter, undefined symbols without subscripts should be understood in the same way as the above ones. The Economic Dispatch model is presented below, we name it ED1.

$$\underset{g,z,\delta}{\text{Min}} \quad \sum_{k\in\mathcal{B}} \alpha_k g_k^2 + \beta_k g_k \tag{2.3}$$

$$\text{s.t.} \quad z_{(k,l,c)} - b_{(k,l,c)}(\delta_l - \delta_k) = 0, \qquad \forall (k,l,c) \in \mathcal{A} \tag{2.4}$$

$$g_k - \sum_{\substack{(l,c):\\(k,l,c)\in\mathcal{A}}} z_{(k,l,c)} + \sum_{\substack{(l,c):\\(l,k,c)\in\mathcal{A}}} z_{(l,k,c)} = d_k, \qquad \forall k \in \mathcal{B} \tag{2.5}$$

$$\underline{g}_k \leqslant g_k \leqslant \bar{g}_k, \qquad \forall k \in \mathcal{B} \tag{2.6}$$

$$\underline{z}_a \leqslant z_a \leqslant \bar{z}_a, \qquad \forall a \in \mathcal{A} \tag{2.7}$$

In ED1, the objective function (2.3) is the total generation cost, with $\alpha_k \geqslant 0, \forall k$ ensuring the convexity of the function. Constraints (2.4) are the defining equations for the power flow z_a. These power flow quantities participate in the constraints (2.5), the nodal power balance equations. The equations in (2.5) say that at each bus k, the net generation $(g_k - d_k)$ must equal to the sum of the outbound power flow from bus k along all lines adjacent to k. Constraints (2.6) are the lower and upper bounds on the power generation, with $\bar{g}_k \geqslant \underline{g}_k \geqslant 0$. A load node $k \in \text{LOAD}$ that

does not generate power is enforced by setting $\bar{g}_k = \underline{g}_k = 0$ in the data. Constraints (2.7) represent the thermal limits on the transmission lines, that is, the magnitude of the power flowing on an arc a should not exceed the arc's thermal limit $\bar{z}_a$ (whereas $\bar{z}_a = -\underline{z}_a$). Note that for a connected network which we assume here, the row rank of the linear system (2.4) to (2.5) was one less than full, which would leave an undesirable extra degree of freedom. For example, given g and z, we would be unable to determine δ. To overcome this issue, practitioners usually select a bus k at which the phase angle δ_k is artificially set to zero and serve as the reference to the angles at all other buses. This bus is called the swing bus. In ED1 and all the subsequent models, the variable fixing is not expressed in the model but will be handled at the solution stage.

An important by-product of solving ED1 is the LMP. Take the bus k for example. The LMP at node k, denoted by λ_k, is by definition the sensitivity of the optimal value of the objective function to the demand d_k. Since ED1 is a convex quadratic programming model for which the KKT conditions are both necessary and sufficient for optimality, it is not difficult to verify that λ_k are the optimal multipliers on the nodal power balance constraints (2.5).

Demand Response Model

In this section, we build a model to dispatch the DR and generation resources simultaneously, taking account of the LMP threshold and the DR cost-effectiveness conditions as required in the FERC Order.

We begin by defining some more variables and parameters. Let $r_k \geqslant 0, k \in \mathcal{B}$ be the amount of demand response to be dispatched at node k. It is a decision variable and is upper bounded by a parameter $\bar{r}_k \geqslant 0$. For a bus $l \in \mathcal{B}$ that is incapable of providing the DR service, setting $\bar{r}_k$ in the data could fix r_l to 0. Let C_1 be the AvgLMP threshold that the ISO tries to maintain via dispatching the DR resources. The resulting LMPs $\lambda_k, k \in \mathcal{B}$, should satisfy the following inequality:

$$\frac{\sum_{k\in\mathcal{B}} d_k\lambda_k}{\sum_{k\in\mathcal{B}} d_k} \leqslant C_1 \tag{2.8}$$

Let C_2 be the AvgPrice before dispatching any DR resources. It is a parameter that can be calculated from the results of ED1, prior to computing the DR dispatch (applying (2.1) with $r_k = 0, \forall k \in \mathcal{B}$). Then the DR cost-effectiveness condition could be expressed as

$$\frac{\sum_{k\in\mathcal{B}}(g_k + r_k)\lambda_k}{\sum_{k\in\mathcal{B}}(d_k - r_k)} \leqslant C_2 \tag{2.9}$$

Within the boundaries of the net benefit test and LMP threshold constraints, there is leeway regarding the dispatch decisions of DR (e.g., which DR provider to dispatch and how much to dispatch), which can also have a substantial impact on economic efficiency. Such decisions will be guided by the objective function of the ISO's DR dispatch algorithm, for which the Order does not have a specification. Since the intended price-suppressing goal of DR is fully represented in the constraints, the objective, on the other hand, should aim to discourage over-suppressing of the price, or equivalently over-dispatching of the demand response, so as to prevent uneconomic consequences (as an example, in Figure 2.2 if the price is suppressed to a level below C, the deadweight loss will emerge again). A myriad of functions can capture the "extent of DR dispatch", and the choice is up to the individual ISO. At present, we find no strong reason for the objective function to go beyond a linear form, so we will minimize a linear function $L(r)$ as the objective of the demand response dispatch. In subsequent analysis, we take $L(r) = \sum_{k\in\mathcal{B}} r_k$ to minimize the total amount of DR dispatch. Note that in cases where DR providers are allowed to bid a valuation, e.g., v_k, in addition to the upper bound $\bar{r}_k$, then $L(r) = \sum_{k\in\mathcal{B}} v_k r_k$ can be an appropriate objective function. This and other variants of the model will be demonstrated in Section 2.5.

For ease of analysis, we present the demand response model in vector format. First, let us make the following definitions. Q is a $|\mathcal{B}| \times |\mathcal{B}|$ diagonal matrix, with $Q_{kk} = 2\alpha_k$; c is a vector of size $|\mathcal{B}|$, with $c_k = \beta_k$; e is a vector of size $|\mathcal{B}|$ with all elements equal to 1; B is a $|\mathcal{A}| \times |\mathcal{A}|$ diagonal matrix, with $B_{aa} = b_a$, for $a \in \mathcal{A}$; A is a $|\mathcal{A}| \times |\mathcal{B}|$ arc-bus incidence matrix, and the element A_{ak}, where $a \in \mathcal{A}$ and $k \in \mathcal{B}$, is equal to -1 if $a = (k, l, c)$ for some (l, c), and equal to 1 if $a = (l, k, c)$ for some (l, c). An illustration is given below, followed by the demand response model DR1.

$$Q = \begin{bmatrix} 2\alpha_1 & & \\ & \ddots & \\ & & 2\alpha_{|\mathcal{B}|} \end{bmatrix} \quad c = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{|\mathcal{B}|} \end{bmatrix} \quad e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$B = \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_{|\mathcal{A}|} \end{bmatrix} \quad A = \begin{bmatrix} 1 & -1 & & \\ & & 1 & -1 \\ & \cdots & \cdots & \\ -1 & & & 1 \end{bmatrix}$$

$$\underset{g,z,\delta,r,\lambda}{\text{Min}} \quad L(r) \tag{2.10}$$

$$\text{s.t.} \quad d^T\lambda \leqslant C_1 e^T d \tag{2.11}$$

$$(g + r)^T\lambda + C_2 e^T r \leqslant C_2 e^T d \tag{2.12}$$

$$0 \leqslant r \leqslant \bar{r} \tag{2.13}$$

and (g, z, δ) solves

$$\underset{g,z,\delta}{\text{Min}} \quad 1/2 g^T Q g + c^T g \tag{2.14}$$

$$\text{s.t.} \quad z - BA\delta = 0 \qquad (\perp \lambda^z) \tag{2.15}$$

$$g - A^T z = d - r \qquad (\perp \lambda) \tag{2.16}$$

$$g \in [\underline{g}, \bar{g}] \qquad (\perp \eta^{lo}, \eta^{up} \geqslant 0) \tag{2.17}$$

$$z \in [\underline{z}, \bar{z}] \qquad (\perp \mu^{lo}, \mu^{up} \geqslant 0) \tag{2.18}$$

where λ is the multiplier of (2.16).

DR1 is a bi-level model. Readers could consult Bard (1998) for a thorough treatment of bi-level optimization models, whereas Colson et al. (2007) provides a useful survey on this subject. The lower level, (2.14) to (2.18), is an economic dispatch model (ED1) that takes the demand response variable r as a parameter. As discussed in Section 2.2, the LMP is the optimal multiplier λ on the nodal power balance constraint (2.16). The upper level minimizes the total MW amount of de-

mand response subject to the LMP threshold constraint (2.11), the cost-effectiveness constraint (2.12), DR bound constraints (2.13) and that (g, z, δ) solves the lower level problem so that λ represents the true LMPs.

An alternative model would be a single-level model formed by preserving all the constraints in DR1 and combining the two objectives in DR1 into one by summing them up. Let us name such a model DR1a. It is easy to generate examples where the solutions of DR1 and DR1a are different. By comparing the two models, we argue that the bi-level model is more appropriate for the problem at hand. First, in DR1a, it is not justifiable to simply take the multiplier λ of the constraint (2.16) as the LMP. By definition, LMP is the derivative of the Lagrangian function with respect to the demand (data) evaluated at a KKT point. The extra constraints (2.11) and (2.12) would complicate the expression of the derivative. In contrast, DR1 encapsulates the original ED model in its lower level and therefore the multiplier λ remains to represent the true LMP. Secondly, although the generation cost and the DR objective function both need to be minimized, they are not simply additive in a single objective function. In fact, minimization of the two objective functions is intrinsically hierarchical in that the core business remains to be the economic dispatch given the demand data as well as a particular DR decision, and on top of that, we seek a "minimal" dispatch of DR to satisfy the LMP threshold constraint and the net benefit test. The bi-level DR1 exactly serves this purpose.

Model Reformulation

The parameterized economic dispatch model in the lower level is a convex quadratic program, hence can be replaced by its KKT conditions, and therefore DR1 becomes an MPEC (mathematical program with equilibrium constraints) model. Specifically, the KKT conditions of the lower level problem include (2.15) to (2.18), as well as

the following equalities and inequalities,

$$A\lambda - \lambda^z - \mu^{lo} + \mu^{up} = 0 \tag{2.19}$$

$$Qg + c - \lambda - \eta^{lo} + \eta^{up} = 0 \tag{2.20}$$

$$(BA)^T\lambda^z = 0 \tag{2.21}$$

$$\eta^{lo}_k(g_k - \underline{g}_k) = 0,\ \eta^{lo}_k \geqslant 0,\ \forall k \in \mathcal{B} \tag{2.22}$$

$$\eta^{up}_k(\bar{g}_k - g_k) = 0,\ \eta^{up}_k \geqslant 0,\ \forall k \in \mathcal{B} \tag{2.23}$$

$$\mu^{lo}_a(z_a - \underline{z}_a) = 0,\ \mu^{lo}_a \geqslant 0,\ \forall a \in \mathcal{A} \tag{2.24}$$

$$\mu^{up}_a(\bar{z}_a - z_a) = 0,\ \mu^{up}_a \geqslant 0,\ \forall a \in \mathcal{A} \tag{2.25}$$

where λ's and η's are dual variables, and their correspondence to the primal constraints (2.15) to (2.18) is marked in the parentheses following the constraints in DR1.

Two difficulties remain for the global solution of DR1: the nonconvexity of the net benefit test constraint (2.12) and the nonconvexity of the complementarity conditions in (2.22) to (2.25). We will address them below.

Transforming constraint (2.12)

The bilinear term $(g + r)^T\lambda$ in the net benefit test constraint (2.12) can be converted into a linear expression of the dual variables, as follows.

$$\begin{aligned}
(g + r)^T\lambda &= (A^T z + d)^T\lambda && \text{by (2.16)} \\
&= z^T A\lambda + d^T\lambda \\
&= z^T(\lambda^z + \mu^{lo} - \mu^{up}) + d^T\lambda && \text{by (2.19)} \\
&= \delta^T(BA)^T\lambda^z + z^T\mu^{lo} - z^T\mu^{up} + d^T\lambda && \text{by (2.15)} \\
&= 0 + \underline{z}^T\mu^{lo} - \bar{z}^T\mu^{up} + d^T\lambda && \text{by (2.21),(2.24)-(2.25)}
\end{aligned}$$

Therefore, constraint (2.12) is reduced to a linear inequality:

$$\underline{z}^T \mu^{lo} - \bar{z}^T \mu^{up} + d^T \lambda + C_2 e^T r \leqslant C_2 e^T d \tag{2.26}$$

Implementing constraints (2.22)-(2.25)

We investigate three approaches to implement the bilinear equations in (2.22)-(2.25). The first approach is taking the bilinear equations "as-is" to form a nonlinear program (NLP), then using an NLP solver to obtain a local solution. The second approach involves linearizing them using binary variables. For instance, the relation

$$\eta_k^{lo}(g_k - \underline{g}_k) = 0 \tag{2.27}$$

is equivalent to

$$\eta_k^{lo} \leqslant \bar{\eta}_k^{lo} \nu_k^{lo} \text{ and } g_k - \underline{g}_k \leqslant (\bar{g}_k - \underline{g}_k)(1 - \nu_k^{lo})$$

where $\bar{\eta}_k^{lo}$ is the upper bound on η_k^{lo} and ν_k^{lo} is a binary variable. The third approach takes advantage of the special ordered sets (SOS) capability of MIP solvers such as CPLEX and Gurobi. For instance, for each generator k we define two positive variables $s_k^{lo} := g_k - \underline{g}_k$ and $s_k^{up} := \bar{g}_k - g_k$ and put the ordered quadruple $\{\eta_k^{lo}, s_k^{up}, s_k^{lo}, \eta_k^{up}\}$ in an SOS2 set (indicating that at most two members of the set can be positive and the positive members must be adjacent).

To obtain global solutions, the first two approaches require upper bounds (big-M) for the multipliers η^{lo}, η^{up}, μ^{lo} and μ^{up}, which must be set artificially in practice. The following method to set the big-M has been tested and shown to be effective in our experiments. First, it is observed that the dispatch of DR at any node would not result in an increase in the highest nodal LMP, and that the LMP usually does not drop below the marginal cost of the cheapest generator. Therefore, we set the upper bound for λ in the DR model as $\bar{\lambda} := \|\lambda^*\|_\infty e$, the highest nodal LMP resulted from the ED1 model, and set the lower bound by $\underline{\lambda} := \|Q\underline{g} + c\|_\infty e$, where e is a vector of 1's. From the bounds on λ, and by (2.20) and the fact that η^{lo} and η^{up} cannot be

positive simultaneously, we can set

$$\bar{\eta}_k^{lo} := (2\alpha_k \bar{g}_k + \beta_k) - \underline{\lambda}_k$$
$$\bar{\eta}_k^{up} := \bar{\lambda}_k - (2\alpha_k \underline{g}_k + \beta_k)$$

For the bound on μ^{lo} and μ^{up}, we use $\bar{\mu}^{lo} := \bar{\mu}^{up} := 2||\mu^*||_\infty e$, where μ^* is the optimal multiplier (a vector of size $2|\mathcal{A}|$) on constraint (2.7) in the ED1 solution and e is a $|\mathcal{A}|$-sized vector of ones.

2.3 Alternative Approaches

We develop two alternative approaches that address the demand response problem under different network assumptions and from different perspectives. They can be used to validate the demand response model.

In a Congestion Free Network

It is well-known that when there is no congestion and no losses in the network, the LMPs at all buses will be identical and equal to the marginal supply offer at the market clearing point. In particular, absent constraints (2.7), ED1 could be simplified to ED2, as follows.

$$\begin{aligned}
\underset{g}{\text{Min}} \quad & \sum_{k\in\mathcal{B}} \alpha_k g_k^2 + \beta_k g_k \\
\text{s.t.} \quad & \sum_{k\in\mathcal{B}} g_k = D \\
& \underline{g}_k \leqslant g_k \leqslant \bar{g}_k, \qquad \forall k \in \mathcal{B}
\end{aligned}$$

where $D = \sum_{k\in\mathcal{B}} d_k$, the total demand in the network. It is not difficult to show that ED2 is an equivalent model of ED1 without line limits – the solution of one implies the solution of the other.

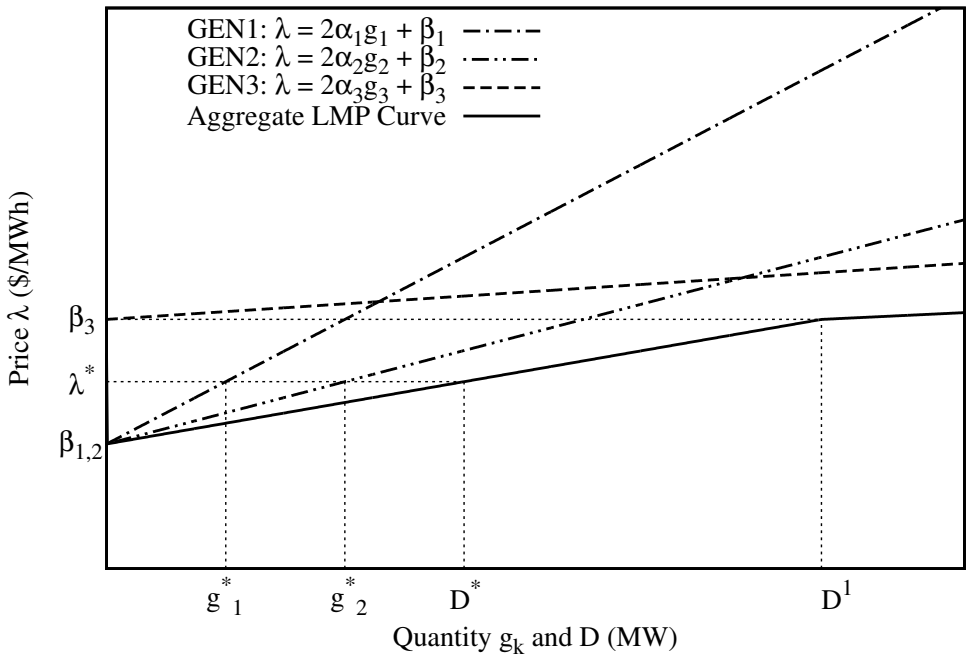


Figure 2.3: LMP curve in a 3-generator example

ED2 can be solved by a graphical method. We illustrate the solution process via a simple example, and then present a graphical criterion for the cost-effectiveness of the demand response.

In Figure 2.3, we draw the marginal cost lines of three generators, and the aggregate supply curve, which we call the LMP curve. In this example, β_1 and β_2 are equal, so we mark them by $\beta_{1,2}$ in the figure. This figure reveals the relationship among the demand, generator dispatch and price. For example, given a total demand D^*, we can read off from the figure the corresponding LMP, which is λ^*, and the generator dispatch solution, which is $(g_1^*, g_2^*, 0)$. It can be seen that the third generator will kick in when the demand is beyond the level of D^1, which marks a kink point on the piecewise linear LMP curve. We provide below a general procedure that constructs the aggregate LMP curve from generators' marginal cost curves.

The inputs to the algorithm include the index set GEN of the generators, the

Table 2.2: Cost Parameters of a Two-generator Example

k	α_k	β_k	$\bar{g}_k$
1	0.1	10	100
2	0.1	50	100

quadratic and linear term coefficients, α_i and β_i, and the capacity, $\bar{g}_i$, of each generator $i \in GEN$. The output is an ordered set $\mathcal{P}$ of break points on the curve. In the algorithm, we use the convention that the minimum over an empty set is infinity, formally, $\min_{s\in\mathcal{S}} s = \infty$ if $\mathcal{S} = \emptyset$.

Algorithm 1 Generating the LMP Curve

Initiate $\mathcal{P} = \emptyset, \mathcal{C} = GEN, \mathcal{U} = \emptyset, D_1 = 0$
Let $\lambda_1 = \min_{i\in\mathcal{C}} \beta_i, \mathcal{A}_1 = \{j \in \mathcal{C} : \beta_j = \lambda_1\}$
Add (D_1, λ_1) in $\mathcal{P}$
while $\mathcal{C} \neq \emptyset$ **do**
  $\lambda_2 = \min_{i\in\mathcal{C}\setminus\mathcal{A}_1} \beta_i$
  $\mathcal{A}_2 = \{j \in \mathcal{C}\setminus\mathcal{A}_1 : \beta_j = \lambda_2\}$
  For each $i \in \mathcal{A}_1$, let $a_i = 2\alpha_i\bar{g}_i + \beta_i$
  $\lambda_3 = \min_{i\in\mathcal{A}_1} a_i, \mathcal{A}_3 = \{j \in \mathcal{A}_1 : a_j = \lambda_3\}$
  if $\lambda_3 \leqslant \lambda_2$ **then**
    Let $d_i = (\lambda_3 - \beta_i)/\alpha_i$ for each $i \in \mathcal{A}_1$, and $D_1 = \sum_{i\in\mathcal{A}_1} d_i + \sum_{i\in\mathcal{U}} \bar{g}_i$
    $\mathcal{A}_1 = \mathcal{A}_1\setminus\mathcal{A}_3, \lambda_1 = \lambda_3, \mathcal{C} = \mathcal{C}\setminus\mathcal{A}_3, \mathcal{U} = \mathcal{U} \cup \mathcal{A}_3$
  else
    For each $i \in \mathcal{A}_1$, let $d_i = (\lambda_2 - \beta_i)/\alpha_i$
    $D_1 = \sum_{i\in\mathcal{A}_1} d_i + \sum_{i\in\mathcal{U}} \bar{g}_i$
    $\mathcal{A}_1 = \mathcal{A}_1 \cup \mathcal{A}_2, \lambda_1 = \lambda_2$
  end if
end while
Return $\mathcal{P}$

Once $\mathcal{P}$ is returned by Algorithm 1, the LMP curve can be plotted by connecting the points in $\mathcal{P}$ one by one in the same sequence as they are added. It is important to note that in the presence of the upper bounds on g_k, the LMP curve thus created may not represent a 1-to-1 mapping between D and λ. For instance, for the case given in

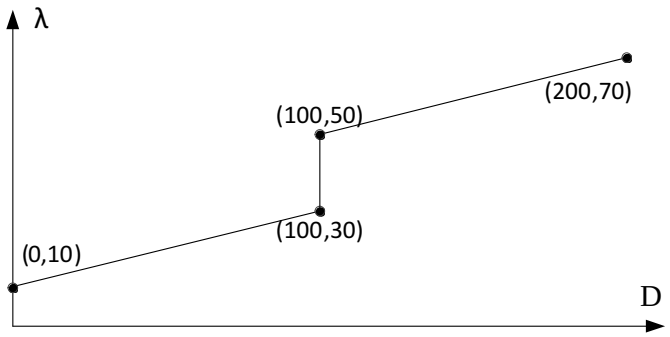


Figure 2.4: An LMP curve with a jump

Table 2.2, Algorithm 1 returns the ordered set $\mathcal{P}$ = {(0,10),(100,30),(100,50),(200,70)}. Connecting these points one by one, we get an LMP curve as shown in Figure 2.4. When $D = 100$, λ is indefinite, indicating that any value in the range [30, 50] would make λ satisfy the KKT conditions. This represents a case where the dual solution of ED2 is not unique.

For the subsequent analysis of demand response on an LMP curve, we use superscript "old" and "new" on a symbol to indicate that it is a quantity before and after the demand response, respectively. Figure 2.5 shows two points on an LMP curve following this superscripting convention.

Let p denote the average price, which is defined by Equation (2.1). The demand reduction of $\Delta D = D^{old} - D^{new}$ results in a reduction in LMP by $\Delta\lambda = \lambda^{old} - \lambda^{new}$, and the average price after the demand response is $p^{new} = \lambda^{new} D^{old}/D^{new}$, while the average price before the demand response is $p^{old} = \lambda^{old}$. The cost-effectiveness condition requires $p^{new} \leqslant p^{old}$, which gives $\lambda^{new} D^{old} \leqslant \lambda^{old} D^{new}$. Seen from Figure 2.5, this inequality is equivalent to

$$\lambda^{new} \Delta D \leqslant D^{new} \Delta\lambda, \tag{2.28}$$

the left and right hand sides of which are the lower-right and upper-left shaded areas in Figure 2.5. Since all the quantities involved in (2.28) are positive, we can

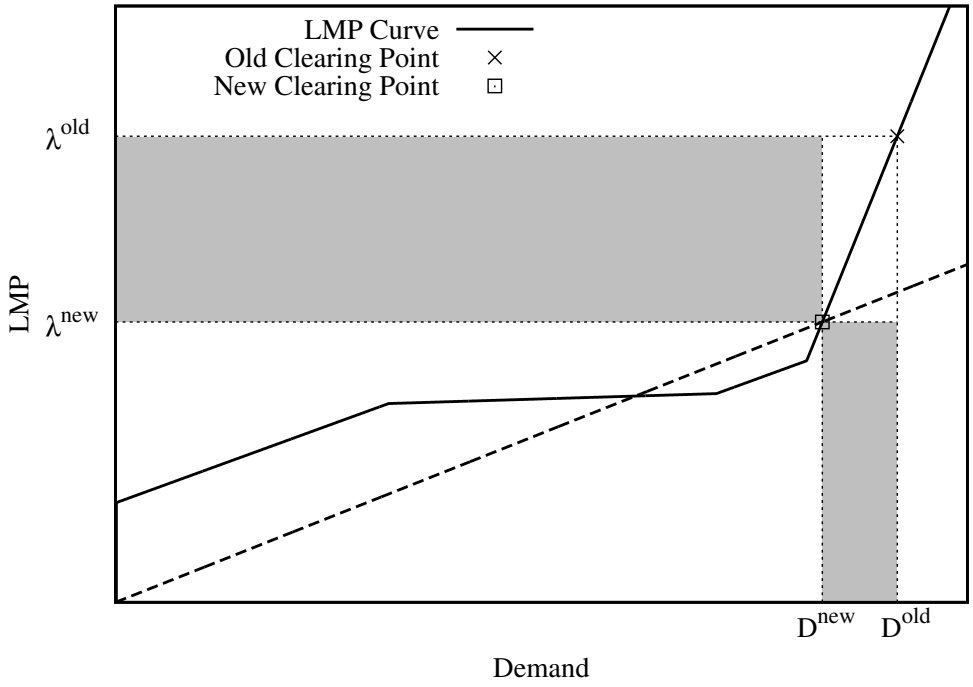


Figure 2.5: DR cost-effectiveness test on an LMP curve

equivalently write the inequality in the form

$$\frac{\lambda^{new}}{D^{new}} \leqslant \frac{\Delta\lambda}{\Delta D}. \tag{2.29}$$

In (2.29), the left hand side is the slope of the line passing through the origin and the point (D^{new}, λ^{new}), i.e., the dashed line in Figure 2.5, and the right hand side is the slope of the line segment connecting (D^{old}, λ^{old}) and (D^{new}, λ^{new}). We summarize this observation in the following rule.

Rule 1. *On the LMP curve, if the slope of the line connecting* (D^{old}, λ^{old}) *and* (D^{new}, λ^{new}) *is bigger than the slope of the line connecting (0,0) and* (D^{new}, λ^{new})*, then it is cost effective to reduce the demand from* D^{old} *to* D^{new}*.*

Applying Rule 1, we can see that Figure 2.5 shows a case where it is cost effective to dispatch the ΔD amount of demand response from the current demand level of

D^{old}.

It is also easy to derive the "local" cost-effectiveness condition for the demand response, in other words, whether the demand response is immediately cost effective as the DR amount ΔD increases from zero. As D^{new} approaches D^{old} from below, the left hand side of (2.29) becomes

$$\lim_{D^{new}\uparrow D^{old}} \frac{\lambda^{new}}{D^{new}} = \lim_{D^{new}\uparrow D^{old}} \frac{\lambda(D^{new})}{D^{new}} = \frac{\lambda(D^{old})}{D^{old}} = \frac{\lambda^{old}}{D^{old}},$$

which is the slope of the line passing through the origin and (D^{old}, λ^{old}), while the right hand side of (2.29) becomes

$$\lim_{D^{new}\uparrow D^{old}} \frac{\Delta\lambda}{\Delta D} = \lim_{D^{new}\uparrow D^{old}} \frac{\lambda^{old} - \lambda^{new}}{D^{old} - D^{new}} = \lim_{D^{new}\uparrow D^{old}} \frac{\lambda(D^{old}) - \lambda(D^{new})}{D^{old} - D^{new}} = \partial_{-}\lambda(D^{old}), \tag{2.30}$$

which is the left derivative of $\lambda(D)$ at D^{old}. If we make an convention that, at the demand level D^* which corresponds to a range of indefinite λ, let $\lambda(D^*)$ be the minimum value in the range, then $\lambda(D)$ becomes a function (1-to-1 mapping) of D, and (2.30) is just the slope of the LMP curve to the immediate left of the point (D^{old}, λ^{old}). Therefore, determining whether the demand response is locally cost effective at demand level D amounts to comparing the slopes of two lines that cut through the (D, λ) point in the LMP curve. This is noted in Rule 2.

Rule 2. *At a demand level* D*, if the left slope of the LMP curve at* (D, λ) *is bigger than the slope of the line connecting (0,0) and* (D, λ)*, then demand response is locally cost effective at the demand level* D*.*

In a General Network

In the presence of the line flow constraints, demand response decisions are no longer easy to make. Active line flow constraints would make the LMPs different for different locations, and there would be no simple mapping between the AvgLMP

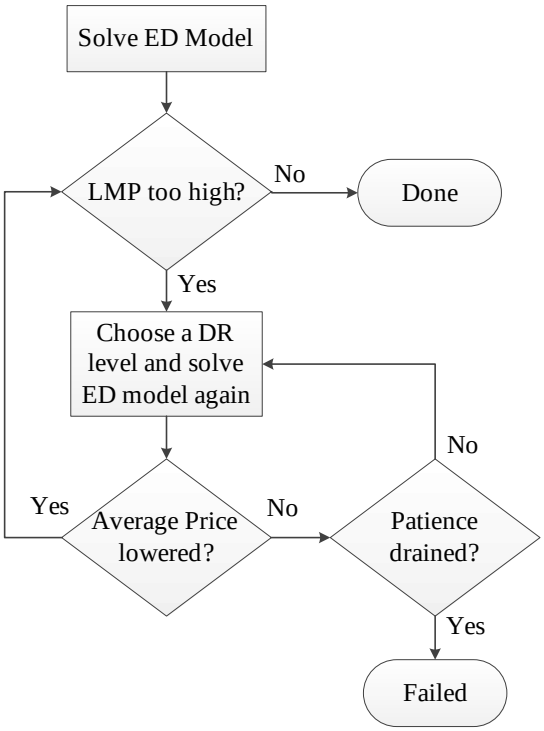


Figure 2.6: Heuristic framework for finding a DR solution

and the total demand D. An intuitive way of thinking would be viewing the demand response as data adjustments in the economic dispatch model. We thereby design a heuristic method following this idea.

From the data adjustment viewpoint, to test if a particular demand response proposition is feasible (both keeping the AvgLMP below the threshold and reducing the AvgPrice), one could adjust the demand data according to the DR proposition, solve the economic dispatch model with the adjusted demand data, compute the new AvgLMP and AvgPrice from the solution, and compare with the original AvgLMP and AvgPrice. This idea is illustrated in Figure 2.6.

The key point in this framework is how to choose an appropriate DR level in each iteration, so as to steer the process toward a feasible solution. A myriad of heuristics can be workable for this point. We design a simple line search method. Specifically,

in each iteration, try for each (DR qualified) bus one by one a small amount of demand reduction and select the bus that yields the lowest total cost to really make the small demand reduction. Repeat this process until the AvgLMP drops below the threshold and the AvgPrice is lower than the original price (succeeded) or no bus would yield any drop in the total cost by the small demand reduction (failed). We name such a line search heuristic method LS1.

Two merits of LS1 are worth noting. First, the search path progresses in small increments, which effectively reduces the possibility and extent of overdoing the demand reduction. The whole purpose of the demand response is keeping the AvgLMP under the threshold, not necessarily making the AvgLMP as low as possible. To this extent, small steps are safer than long shots, and simpler as well since no backtracking is needed. Second, the search path is locally optimized in terms of the total cost calculated in each step. This local optimality is implemented by evaluating the effect of the demand reduction step on each DR-qualified bus and picking the best one.

Admittedly, there is no guarantee that LS1 always terminates at a solution when one exists, however, if it does find a solution, the solution is feasible and is quite parsimonious in terms of the total DR amount dispatched. Apart from a hierarchical modeling approach as implemented in DR1, an iterative process in the form of Figure 2.6 is almost the only option to implement the requirements of the Order. In this sense, we believe it is acceptable to use LS1 to represent the class of the "second best" methods to compare to the DR1 model.

2.4 Numerical Experiments

We first validate the model DR1 by comparing its solutions to those obtained by the alternative approaches developed in Section 2.3. We will use the well-known 14-bus case (Christie, 1993) in the validation experiments, with the generator cost parameters coming from the Matpower (Zimmerman et al., 2011) data sets.

Without Line Limits

Applying the method developed in Section 2.3, the congestion-free LMP curve for the 14-bus case is given by

$$\lambda = \begin{cases} 0.073412D + 20, & D \in [0, 272.402] \\ 0.006112D + 38.33515, & D \in [272.402, 599.642] \\ 0.073421D - 2.02661, & D \in [599.642, 689.611] \\ 0.5D - 296.2, & D \in [689.611, 772.400] \\ \infty, & D \in [772.400, \infty] \end{cases}$$

At $D = 599.642$, $\lambda = 0.073421 \times 599.642 - 2.02661 = 42.00$. Since the slope $599.642 \div 42.00 = 0.070042$ falls in between 0.006112 and 0.073421, we can identify the demand level 599.642 as a threshold for the cost-effectiveness of DR. In other words, DR is locally cost-effective only when the demand level is higher than 599.642 MW, see Figure 2.7. With this knowledge, we can fabricate some scenarios on which the results are predictable and test if the solutions found by DR1 on these scenarios match our predictions.

For clarification, we note that the original 14-bus case has a total demand (sum of all nodal demand) of 259 MW. In the experiments, when we need to reset the demand to a particular level, we do this by multiplying a scale factor on all nodal demands. For example, the GAMS statement "d(k) = d(k)*2.5" scales up all nodal demands by 2.5, achieving a total demand level of $259 \times 2.5 = 647.5$ MW. Also note that for experimental purpose, we set $u_k^R = (1 - \epsilon)d_k$ for all $k \in BUS$, where $\epsilon = 0.01$, that is, the demand is allowed to freely decrease down to almost zero. Zero net demand is avoided to keep equation (2.1) well-defined.

Scenario 1:

Set the demand to 599.642 MW, the economic dispatch model ED1 gives the current AvgLMP λ of 42.00 and the current AvgPrice C_2 of 42.00. We know that at this demand level, any positive DR level would violate the cost-effectiveness condition.

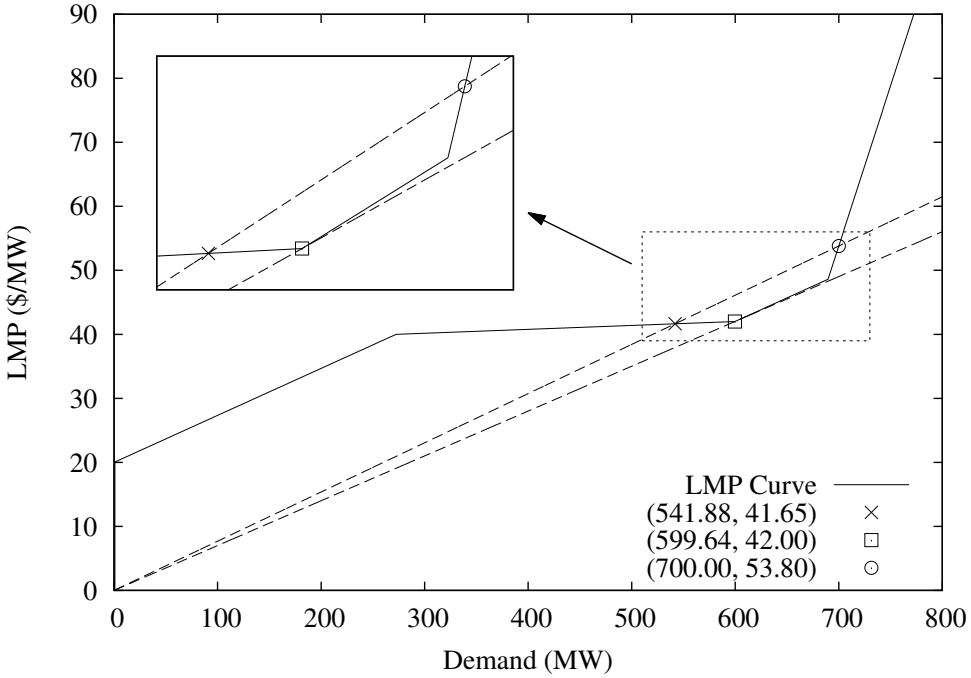


Figure 2.7: LMP curve for the 14-bus case without line limits

So, if the LMP threshold C_1 is set to 42.00, we would expect DR1 to be just feasible thus optimal at $R_k = 0, \forall k$; and for a slightly lower LMP threshold, e.g., $C_1 = 41.99$, DR1 would become infeasible. Experiments verified the above speculations. Furthermore, similar results are expected to occur for any demand level that is lower than 599.624, which is also confirmed by experiments on demand levels sampled within the range [0,599.624], as demonstrated in Table 2.3.

Scenario 2:

Set the demand to a level above 599.642 MW, for example, 600 MW, then ED1 gives $\lambda = C_2 = 42.03$. If we set $C_1 = 42.00$, an AvgLMP level corresponding to 599.642 MW demand, we would expect $600 - 599.642 = 0.358$ MW of demand response to be dispatched. The DR1 result confirmed this expectation, dispatching exactly this amount of DR at bus 2. We carry out a series of experiments on demand levels

Table 2.3: DR1 Results for Cost-ineffective Demand Levels

D	LMP (λ)	Price (C_2)	LMP Cap (C_1)	r_k
500	41.392	41.392	41.392	0
			41.391	infes
400	40.780	40.780	40.780	0
			40.779	infes
300	40.169	40.169	40.169	0
			40.168	infes
200	34.685	34.685	34.685	0
			34.684	infes

Table 2.4: DR1 Results for Cost-effective Demand Levels with Different C_1 Values

D	λ	C_2	C_1	$\sum R_k$	new λ	new Price
650	45.70	45.70	45.00	9.50	45.00	45.67
			44.00	23.12	44.00	45.62
			42.00	50.36	42.00	45.53
			41.986	52.65	41.986	45.687
			41.985	infes	N/A	N/A
700	53.80	53.80	53.00	2.60	53.00	53.12
			48.61	10.38	48.61	49.34
			42.00	102.79	42.00	49.21
			41.647	158.12	41.647	53.80
			41.646	infes	N/A	N/A
750	78.80	78.80	78.00	1.60	78.00	78.17
			48.61	60.38	48.61	52.87
			42.00	150.36	42.00	52.53
			40.703	362.58	40.703	78.795
			40.702	infes	N/A	N/A

above 599.642 coupled with various C_1 levels. The results are summarized in Table 2.4. All results produced by DR1 match those generated by the graphical approach.

The $D = 700$ case is also illustrated in Figure 2.7. The line connecting the origin and the point (700, 53.8) intersects the LMP curve at (541.88, 41.65), so the maximum amount of cost-effective demand reduction from 700 MW is 700 - 541.88 = 158.12

MW and the corresponding LMP is 41.65. These assertions are verified by DR1.

Conclusion 1. *DR1 works correctly in a congestion free network.*

Another important point could be observed. As shown above, the demand can be cost-effectively reduced from 700 to 541.88. However, we have noted in Scenario 1 (and also by examining Figure 2.7) that any demand reduction from a demand level below 599.64 would have been cost-ineffective. The rationale for these clashing observations lies in the fact that the cost-effectiveness judgement depends on the current (starting) demand level. For example, at $D = 700$ the AvgPrice is 53.80, so a demand reduction that could yield an average price no higher than 53.80 would be deemed cost-effective; but at $D = 599.64$, a demand reduction would have to yield an average price less than or equal to 42.00 in order to be cost-effective.

With Line Limits

We will experiment on the case D = 650 MW with a limit of 150 MW on every line. From the solution of ED1 and applying the formula (2.1) and (2.2), we first obtain AvgLMP = 73.13 and AvgPrice = 62.04. Then we run DR1 and LS1 on a few selected C_1 levels and compare the results. In each of the experiments below, the step length in LS1 is set to 0.1 MW.

Table 2.5 exhibits the DR1 and LS1 solutions for the case $C_1 = 73$. Both methods found the same solution. The solution is to dispatch 0.4 MW, or to be exact, 0.302082 MW of DR at bus 2. The numerical difference is due to the fact that LS1 is an approximate method whose precision is only up to 0.1 MW, while DR1 is an exact method. The decision that bus 2 is selected makes sense since the LMP at bus 2 is the highest, even after the demand response.

The reason why we choose the case $C_1 = 73$ to elaborate comes from practical considerations. In practice, the DR model or routine is executed whenever the AvgLMP rises above the LMP threshold, and since the demand usually does not change wildly in the time interval (5-minute or hourly) within which the ED model is executed and the AvgLMP is updated, the AvgLMP will not be much higher

Table 2.5: Comparison of DR1 and LS1 Solutions for $C_1 = 73$

	DR1			LS1		
k	g_k	r_k	λ_k	g_k	r_k	λ_k
1	230.3		39.819	230.3		39.819
2	119.4	0.302082	79.649	119.3	0.4	79.699
3	100		75.382	100		75.338
4			71.310			71.272
5			68.526			68.491
6	100		69.401	100		69.364
7			70.849			70.811
8	100		70.849	100		70.811
9			70.607			70.569
10			70.402			70.364
11			69.920			69.883
12			69.476			69.439
13			69.582			69.545
14			70.163			70.126

than the threshold at the time the DR model is triggered. In this sense, an LMP threshold of 73 is reasonable for the current AvgLMP of 73.13.

The results for other LMP threshold cases are summarized in Table 2.6. To save space, the third column lists the total DR amount (summed over all buses) in each solution. We can see that in terms of the total DR amount, the LS1 solution is a round-up of the DR1 solution in all of these cases, which matches our expectation knowing the nature of the two respective methods. However, the efficiency of the methods differs greatly. The solution time of LS1 grows significantly as C_1 is set lower and lower away from the AvgLMP of the initial ED1 solution, while the solution time of DR1 remains short regardless of the parameter.

Conclusion 2. *DR1 works correctly in a general network and significantly outperforms the heuristic method in terms of accuracy and efficiency.*

Table 2.6: Comparison of DR1 and LS1 Solutions for Different C_1 Levels

C_1	Model	$\sum r_k$	Gen Cost	DR Cost	Time (sec)
73.00	DR1	0.302082	25140.045	24.075	1.2
	LS1	0.4	25132.243	31.860	5.6
70.00	DR1	7.513285	24578.326	571.704	1.6
	LS1	7.6	24571.730	577.973	101
60.00	DR1	31.550629	22893.714	2021.568	2.2
	LS1	31.6	22890.551	2023.952	417
45.00	DR1	69.14031	20847.861	3175.374	2.3
	LS1	69.2	20845.194	3176.459	922

General Solvability

We experiment different formulations and solvers on five IEEE test cases (Christie, 1993) to demonstrate the general solvability of the model. In particular, we run the NLP formulation using CONOPT, BARON and GLOMIQO, and run the binary and SOS formulation using CPLEX and GUROBI. Two congestion conditions are examined for each test case: free and congested. In order to make feasible yet simple DR cases, we need to scale up the demand to certain levels and set appropriate line limits for the congested scenarios. The setting and solutions are presented in Table 2.7, in which λ^{ED} and λ^{DR} stand for the AvgLMP computed from the ED and DR solutions, and all C_1 levels were set to $0.9\lambda^{ED}$ for simplicity.

Table 6.2 lists the computation time (in seconds) for each solver to find the solution, where "-" indicates not finishing within an hour. The computer is a Dell R710 server with two 3.46G X5690 Xeon Chips, 12 Cores and 288GB Memory. For BARON, GLOMIQO and the binary formulation, we apply the big-M bounds discussed in Section 2.2 to pursue the global solution within the bounds. Note that the SOS formulation does not require artificial variable bounds, thus can be trusted to provide the true global solution. The fact that all solvers obtained the same solution is evidence for the validity of our choice of variable bounds. In all cases, CONOPT consistently provides a good local solution very quickly, which can serve as a starting point for other global solvers. We use this as part of a three-phase

Table 2.7: Setting and Solution of IEEE Test Cases

Bus	Setting			ED Soln.		DR Soln.		
	D	$\bar{z}_k$	C_1	λ^{ED}	C_2	λ^{DR}	AvgPrice	$\sum r_k$
14	700	∞	48.42	53.80	53.80	48.42	49.33	12.92
		180	69.42	77.13	64.76	69.42	61.59	19.95
30	320	∞	4.84	5.38	5.38	4.84	5.10	16.48
		42	5.50	6.11	5.89	5.50	5.47	3.65
57	1600	∞	54.23	60.26	60.26	54.23	56.01	50.93
		220	54.58	60.65	56.42	54.58	53.45	43.11
118	9500	∞	53.61	59.56	59.56	53.61	54.01	71.16
		390	156.55	173.94	135.01	156.55	122.91	0.85
300	31956	∞	68.79	76.43	76.43	68.79	69.74	437.50
		1680	252.95	281.05	270.15	252.95	243.61	11.24

Table 2.8: Solution Time (in seconds) of Different Formulations and Solvers

Bus	Status	NLP			Bin		SOS	
		Conopt	Baron	Glomiqo	Cplex	Gurobi	Cplex	Gurobi
14	free	0.12	0.10	0.12	0.17	0.19	0.18	0.16
	cong	0.12	0.16	0.12	0.28	0.13	0.13	0.15
30	free	0.12	202.76	1.10	0.16	0.16	0.29	0.15
	cong	0.13	82.71	2.66	0.29	0.31	0.17	0.15
57	free	0.17	16.88	3.66	0.17	0.17	0.26	0.17
	cong	0.15	-	11.21	0.29	0.29	0.82	0.28
118	free	0.13	-	9.54	0.28	0.25	9.68	5.29
	cong	0.13	-	226.62	2.91	2.68	8.40	5.68
300	free	0.25	-	7.35	0.42	0.49	4.30	1.81
	cong	0.14	-	833.44	2.51	2.70	4.22	2.56

solution strategy to be discussed below.

Solving Realistic Instances: A Three-phase Approach

We proceed to test DR1 on larger cases based on the Polish network. While the nodal demands are scaled up (by a factor between 1.05 to 1.2) to make feasible DR cases, we adopted the realistic line ratings given in the network data. It is observed

that in realistic cases most lines will never reach their thermal limits. We exploit this observation in a three-phase solution approach as outlined below.

1. Fast local solution: We first solve the NLP reformulation using CONOPT to obtain a local solution with objective value R^*. If CONOPT reports an infeasible solution, set $R^* = \sum_{k\in\mathcal{B}} \bar{r}_k$ (its maximum possible value) for use in the second phase.

2. Bound and fix: For each line $a \in \mathcal{A}$, we find the lowest/highest level that the flow z_a can possibly reach (let us call such a level an effective bound), by minimizing/maximizing z_a subject to (2.15), (2.16), (2.17), (2.18) and the inequality $\sum_{k\in\mathcal{B}} r_k \leqslant R^*$. If the effective lower bound of z_a is greater than $\underline{z}_a$, then μ_a^{lo} (which belongs to a SOS2 set) can be fixed to zero in the DR1 model; likewise, if the effective upper bound of z_a is less than $\bar{z}_a$, μ_a^{up} can be fixed to zero in the DR1 model. Such a "bound and fix" step could significantly reduce the effective number of discrete variables in the MIP (binary or SOS) formulation of DR1, making it easier to solve. Note that exploring the effective bounds requires solving $2|\mathcal{A}|$ linear programs, which is computationally inexpensive and is efficiently parallelizable (we used 40 parallel processes in the experiments).

3. Solving the MIP: After the variable fixing, we now solve the MIP (using either the binary or SOS2 formulation of Section 2.2) with CPLEX, taking the local solution from phase 1 as an initial integer feasible solution (i.e., enabling the *mipstart* option in CPLEX).

The performance of this approach is demonstrated in Table 2.9. The solution times of each step are listed in the last three columns of the table. We can see that for each of the five cases, CONOPT obtains the local solution within about 10 seconds and the bound strengthening time is well within 2.5 minutes. Both of the binary and SOS2 formulations obtained the same solution but the binary formulation solves much faster. Here are the settings used in the experiments: In order to realistically control the size of the instances, the qualified DR buses are set to be those having an original demand level within a certain interval, e.g., $[30, \infty]$ MW, and the DR upper bound r_k is set to 10% of the original demand D_k. Since the Polish data do not contain generators' quadratic cost coefficients, we artificially set

Table 2.9: DR Test Results on Polish Networks

Case	C_1	ED Soln.		DR Soln.			Soln. Time (seconds)			
		λ^{ED}	C_2	λ^{DR}	AvgPrice	$\sum r_k$	NLP	Bound	Bin	SOS
2383-bus	178.00	179.96	164.31	178.00	163.60	7.36	2.6	112.9	301.6	412.1
2736-bus	110.00	118.20	117.50	110.00	115.84	1161.64	10.1	124.2	16.4	67.1
2737-bus	113.00	115.21	114.74	113.00	113.68	146.85	4.4	100.2	4.90	27.1
2746-bus	112.00	112.82	111.98	112.00	111.38	117.76	4.5	95.8	73.1	1476.6
3012-bus	250.00	258.85	197.58	250.00	192.68	46.10	3.7	109.3	47.8	384.0

Table 2.10: Settings and Bounding Results on Polish Networks

Case	Lines	Demand	# DR Buses	Avail. DR MW	# Zmax	# Zmin
2383-bus	2896	25809.5	107	610.4	24	74
2736-bus	3269	19882.0	1305	1782.3	10	38
2737-bus	3269	13746.0	646	547.1	1	7
2746-bus	3279	26116.7	133	295.7	14	56
3012-bus	3572	29372.0	10	270.5	14	31

them to 0.1 for all generators in all cases. Table 2.10 summarizes the case-specific setting, i.e., total number of lines, total demand, number of DR qualified buses, total available MW of DR. The last two columns are the number of lines which can possibly reach their upper bound and lower bound, respectively, determined by the bounding procedure. It can be seen that most lines will never reach their bounds, hence the corresponding multipliers are fixed to zero in the subsequent MIP solve. We have also tested various cases in which CONOPT reported infeasible. For such cases, CPLEX was able to terminate with infeasibility quickly (in less than 1 minute), which globally verifies that the cases are indeed infeasible.

We acknowledge that the DR model contains unavoidable nonconvex constraints and there can be no general guarantee that a global solution is always obtainable within a reasonable amount of time – it is an extremely hard problem. However, realistic instances/data usually have exploitable characteristics such as the ones exploited above, and the use of SOS2 formulation (without a need of artificial bounds) or big-M formulations (for faster solution) are acceptable for practice. Case-specific data analysis and simplification are necessary complements to the

practical deployment of the model.

2.5 Extensions

In this section, we design and carry out more experiments on various data cases and obtain some insight on the model and the problem at large. For each of the cases involved below, the network data are obtained from Christie (1993) and the generator cost parameters are from Matpower. The experiments are performed on a Dell Precision laptop computer with Intel Core i7 CPU @ 1.87GHz and 8 GB memory, on which GAMS (ver 23.7.1) is run in a Windows 7 64-bit operating system.

On the Cost-effectiveness Condition

We design experiments on four data cases to shed some light on the DR cost-effectiveness condition. The results are presented in Figure 2.8. Each subplot in the figure represents a series of experiments on the same data scenario (as annotated by the subplot title) with different C_1 levels. Let us take the first subplot for example. The experiments are carried out on the 14-bus case with a total demand level of 650 MW (scaled up to this level to ensure that a positive demand response is locally feasible), and a 150 MW limit on every line (set to this level so that it is binding on at least one line). In the first run, C_1 is set to the current AvgLMP obtained from the economic dispatch run, a level that barely makes "$r_k = 0, \forall k \in \mathcal{B}$" feasible (thus optimal for DR1). After the run, the AvgLMP and the AvgPrice from the optimal solution is plotted as a solid and dashed dot, respectively. Then we reduce C_1 by a fixed interval (e.g., \$1/MW) and re-run the model to plot the next pair of dots, and so forth until we reach a C_1 level where DR1 could not find an optimal solution. The solid and dashed curves are obtained by connecting the dots. Roughly speaking, each point on the curves represents an optimal solution of DR1 given a certain C_1 level.

We make the following observations and remarks based on the experiments.

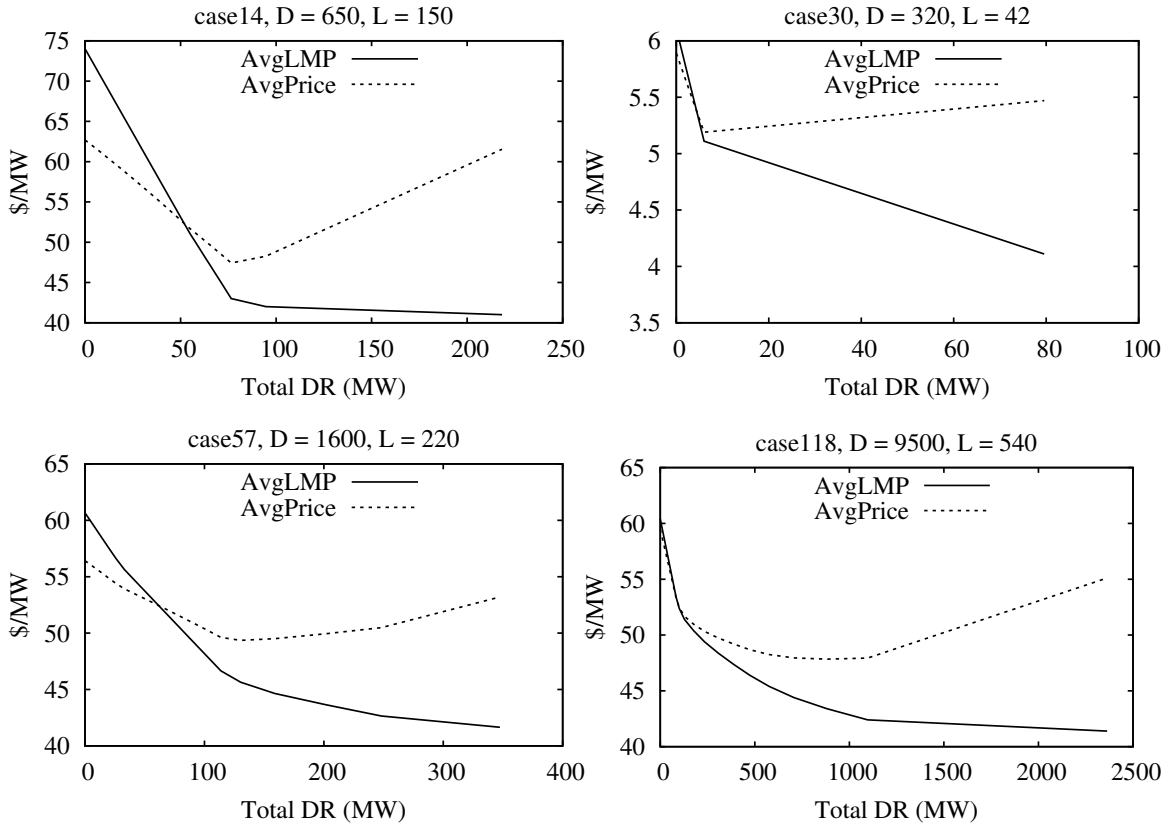


Figure 2.8: Optimal solutions for decrementing C_1 levels on different data cases

1. None of the maximum feasible Total DR levels reaches the total demand level D, and the problem becomes infeasible when the AvgPrice rises above its initial value at zero Total DR. This indicates that in the case when a demand response is cost-effective, it is only cost-effective within a certain interval. Beyond this interval, the demand response would make the AvgPrice higher than the original and thus violate the cost-effectiveness condition.

2. The AvgPrice curves are convex shaped. As the Total DR level increases, the AvgPrice first decreases and then increases, and as long as it has not surpassed the original AvgPrice level, a feasible solution exists. The rationale is explained at the end of Section 2.4. However, as practical advice, we would recommend

the ISO consider dispatching demand response when the AvgLMP is not too much above the LMP threshold C_1, or equivalently, set the threshold so that the resulting Total DR is substantially lower than the value that makes the AvgPrice start increasing.

3. Considering the shape of the AvgPrice curve, one could identify the C_1 level that corresponds to the minimum AvgPrice more efficiently by carrying out a strategic search method, such as the golden section search or Fibonacci search.

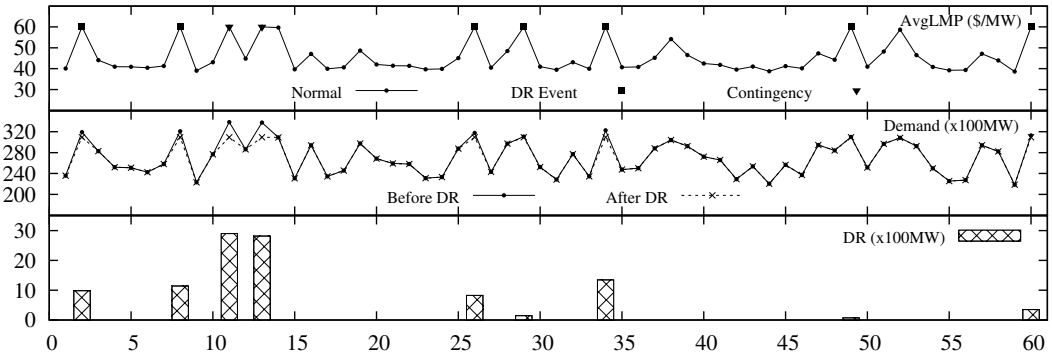
In all the experiments documented up to here, we have observed that the local solutions found by CONOPT coincide with the global solution found by BARON, although BARON took a longer time to terminate. For example, for each run of the 118-bus case, the time spent by CONOPT to obtain the initial local solution is less than 1 second, and the time spent by BARON to prove that the solution is indeed global is about 7 minutes. This observation provides evidence that the model can, in effect, be solved globally by CONOPT. In the subsequent simulation, we stop invoking BARON and directly report the CONOPT solution.

Simulation

We simulate an operating power system based on the 300-bus case to demonstrate the use of DR1 in the ISO's market clearing practice. Furthermore, we use demand response as a corrective measure to restore the normal operation when the economic dispatch is incapable of providing a feasible solution due to demand surges.

We do a series of 60 dispatch experiments each with a random demand profile. We randomize the demand by multiplying the base case demand by a random scale factor uniformly distributed between 0.90 and 1.42. The LMP threshold is set to \$60/MW as we regard this as a reasonable level for demonstration. Given a random demand profile, in each run we first execute ED1 and take one of the following three actions depending on the outcome of ED1. Specifically, if ED1 gives an optimal solution and the corresponding AvgLMP is below the threshold, then no demand response is needed thus DR1 is not executed. If ED1 gives an optimal

Figure 2.9: Simulation results for the 300-bus case.



solution and the AvgLMP is above the threshold, then DR1 is executed to find an optimal DR (and implicitly ED) solution with a satisfactory AvgLMP. Finally, if ED1 fails to give an optimal solution, which indicates that the demand has exceeded the generation capacity and we cannot compute a value for the AvgPrice, then DR1 is executed with $C_2 = \infty$. In the last case, by executing DR1 we hope to not only control the AvgLMP below the threshold but also restore a feasible ED solution, and in exchange for this ambitious goal, we compromise the cost-effectiveness requirement by setting C_2 to infinity.

The series of experiments can be seen as a simulation of the energy market over a certain period of time, the length of which depends on how frequent the dispatch is updated over its duration. For example, it could represent 60 hours within the day-ahead market with hourly dispatch, or 5 hours of the real time market with a 5-minute dispatch interval.

The simulation results, i.e., AvgLMP, demand and DR levels, are plotted in Figure 2.9. We mark three different system events: "Normal" if no DR is needed, "DR Event" if DR is dispatched to bring down the AvgLMP, and "Contingency" if DR is dispatched to restore the system feasibility. As shown in the experiments, DR1 is always successful to maintain the desired level of AvgLMP when a demand surge occurs (seven occurrences in the 300-bus case), and never fails to restore the system feasibility when needed (two occurrences). Furthermore, there is an apparent positive correlation between the demand level and the AvgLMP level,

which indicates that demand response is indeed an effective way to control the market prices.

Variants

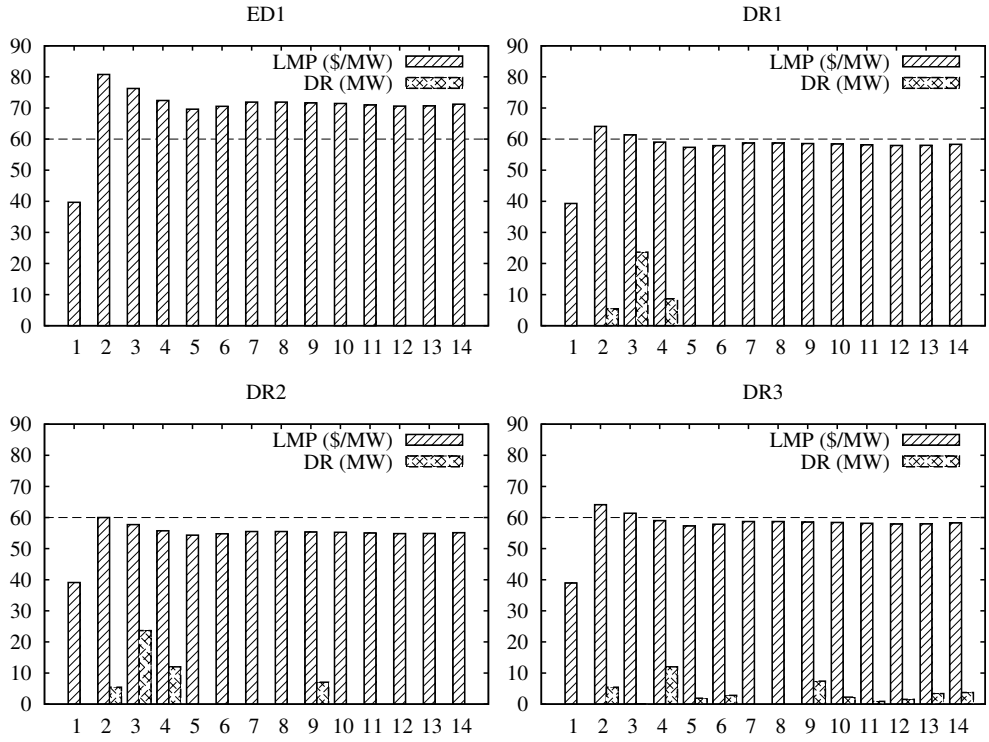
The bi-level structure of DR1 on one hand honors the original economic dispatch to its full extent, and on the other hand provides great flexibility for specifying various requirements on the demand response decisions. Users could modify the upper level objective function and constraints to achieve customized goals. As an example, we give two variants of DR1 as follows.

- DR2: Replace the constraint (2.11) by $\lambda_k \leqslant C_1, \forall k \in \mathcal{B}$ to impose the LMP threshold on every nodal LMP instead of on the AvgLMP.
- DR3: Set the objective function (2.10) by $L(\lambda) = \sum_{k\in\mathcal{B}} v_k r_k$ to account for the valuation v_k that the DR provider k places on a MW of demand reduction.

An illustrative experiment is performed on the 14-bus case with results presented in Figure 2.10. The total demand level is set to 650 MW and the line limit is 150 MW on every line. While ED1 gives an AvgLMP of \$74.01/MW, we set C_1 to \$60/MW, as depicted by the horizontal dotted lines in the subplots. For DR3, we set $v_3 = 200$ and $v_k = 100, \forall k \in \mathcal{B}/\{2\}$ to express a higher reluctance to dispatch demand response at node 3 compared to other nodes. For each node indicated on the horizonal axis, the bar on the left represents the LMP level and the bar on the right (if exists) represents the dispatched DR level at this node. Note that the LMP and DR levels share the same scale along the vertical axis but have different units, i.e., LMP is measured in \$/MW and DR in MW.

As seen in the figure, DR1 was able to reduce the AvgLMP by dispatching a total of 37.7 MW of DR at nodes 2, 3 and 4. DR2 dispatched more (totaling about 48.1 MW) demand response at nodes 2, 3, 4 and 9, thus was successful to keep the maximum nodal LMP under \$60/MW as intended. DR3 apparently took into account the higher valuation v_3, and as a result dispatched much less DR at node 3 (about 0.02

Figure 2.10: Comparison of DR model variants on the 14-bus case.



MW) but dispatched more at various other nodes. These variants show that the bi-level DR model behaves sensibly and is flexible for further customizations.

2.6 Conclusion

Since the enactment of the FERC Order 745 in 2011, methodology research for a compliant and constructive implementation has been scarce in the academic literature. As the primary significance of this work, we have modeled the demand response decision-making process in a way that conforms to the Order requirements.

A bi-level structure is used in the model to capture the interdependency between the LMP and the dispatch of demand response, which is emphasized via the net benefit test specifications in the Order. We remark that without the bi-level

structure, the interdependency can only be dealt with in a heuristic and inefficient way. To obtain a global solution, we have transformed a nonconvex constraint into a linear form and investigated various methods to reformulate the complementarity relations. We have carried out extensive computational experiments and concluded that:

1. Local NLP solutions are always quick to compute and are useful to generate starting points.

2. Realistic instances, despite their large scale, are not necessarily prohibitive to solve if data characteristics are sufficiently understood and exploited.

3. The bi-level model is able to produce valid DR solutions in compliance with the Order and is readily extensible for other DR compensation rules.

For cases where line limits are not binding or can be ignored, we have developed a graphical method to carry out the net benefit test. In a practical situation, this method can be handy to estimate the monthly threshold price as suggested in the Order.

Future work could include customizing and fine-tuning the model to suit the operational requirements of individual ISOs, extending the modeling idea to facilitate the procurement and dispatch of other resources such as the ancillary services, capacity reserve resources and transmission resources, etc., and applying hierarchical models with sophisticated domain enhancements to inform long-term strategic planning decisions, such as transmission expansion and market restructuring.

3 PAYMENT RULES FOR UNIT COMMITMENT DISPATCH

3.1 Introduction

A substantial portion of the electricity trades in the U.S. wholesale energy markets go through the spot markets organized by Independent System Operators (ISO) or Regional Transmission Organizations (RTO). For example, approximately 45% of New York electricity is transacted in the NYISO day-ahead market, 5% is transacted in the NYISO real time market and half through bilateral contracts[1]. Essentially, ISOs need to organize two things: the electricity flow and the cash flow. For the former, ISOs schedule and dispatch the supply and demand in an economic and reliable way, and for the latter, they make and enforce payment rules to fairly allocate the costs and revenues among market participants. These tasks turn out to be problematic, given the complex nature of the commodity. Electricity differs from an ordinary commodity in two ways: (1) what is produced now cannot be practically stored for use later, as the production rate and capacity exceed the storage rate and capacity by orders of magnitude; (2) the demand for electricity exhibits huge, although to some extent predictable, variations throughout a day, the amplitude of which is way greater than the capacity of any individual generating unit. As a result, many generating units are forced to run intermittently, with frequent startups and shutdowns, as well as prolonged periods of no-load "idling". As in other industrial processes, startup and idling of a unit incur extra costs, but unlike other industries, these costs have significant effect on the overall cost of electricity production due to their frequency and must be compensated for in the settlement. Such characteristics of the electricity supply are recognized in the ISOs' market operations. The discrete decisions, i.e., whether to start and shut down a generator, have been an integral part in ISOs' generator dispatch algorithms, and a compensation rule for the discrete activities is also included in most tariffs. However, the ways these discrete decisions or activities are treated

[1]Data source: NYISO Market Training Material – NY Market Orientation Course 2005.

in the dispatch algorithm and in the payment rule are disjoint and incompatible, which undermines the overall fairness and efficiency. This chapter reviews the issues with the current payment rule, justifies an alternative pay-as-bid scheme and presents some preliminary results regarding bidders' response under pay-as-bid.

3.2 The Problematic Payment Rule

In the day-ahead markets, ISOs take bids from the generators and make the unit commitment and dispatch decisions by solving the security-constrained unit commitment and economic dispatch model (UCED) based on the bid data. The UCED model minimizes the total generation cost, subject to the supply-demand balance constraints, transmission network constraints and the generators' operational constraints as specified in the bids. The operational constraints include minimum up-time and down-time constraints, ramp-up and ramp-down constraints, and lower and upper bounds on the output level once committed. This model is a mixed integer program (MIP) that can be, and is actually being, efficiently solved using modern optimization technology in the ISOs' market operations.

Given a particular commitment decision, e.g., the one arising from the solution of UCED, the economic dispatch (ED) model, which is a linear program (LP), finds the optimal dispatch that minimizes the total energy cost. The optimal multiplier value, or the shadow price, of the power balance constraint in the ED model represents the cost of satisfying the next increment of demand and is set as the market clearing price (MCP), or locational marginal price (LMP) to emphasize its "locational" dimension. In fact, MCP is both locational and temporal, indicating the price at a particular location during a particular time period. This interpretation is implied when we use the term "MCP" in the remainder of this chapter.

As for the payment, all committed generators are paid for the MWh energy output at the uniform MCP. However, as pointed out by Johnson et al. (1997), "the commitment which predicates the optimal dispatch phase strongly affects the market clearing prices... these prices are suboptimal since they do not reflect the inter-temporal costs and constraints that drive the unit commitment." It has been

acknowledged that in the context of unit commitment, there may not be a uniform price (i.e., MCP) that supports the efficient market equilibrium . In other words, the socially optimal dispatch solution cannot be achieved by the market participants' profit-maximizing response to a uniform price. To enforce the central dispatch solution, ISOs have to grant generators "make-whole" payments to compensate their opportunity loss, as well as the unit commitment costs, incurred by complying with the central dispatch and giving up their profit-maximization in response to the MCP. Such payments contribute to the ISOs' overall procurement cost and eventually show up on consumers' electricity bills as "uplift" costs.

"Make-whole" payments are addressed by different names within different ISOs. For example, ISO New England uses the term "Net Commitment Period Compensation", Midwest ISO uses "Offer Revenue Sufficiency Guarantee Payment", PJM includes them in the "Operating Reserves Credit", CAISO uses "Bid Cost Recovery (BCR) uplift payment" and New York ISO uses "Bid Production Cost Guarantee Payments". Despite the differing names and formula, generally speaking, a "make-whole" payment to a generator is calculated as the positive difference between the generator's as-bid cost (including energy, no-load and start-up) and its energy revenue paid at the MCP, evaluated at the actual commitment and dispatch solution. If the energy revenue at MCP exceeds the as-bid cost (as in most cases), the generator simply pockets the surplus and no "make-whole" payment is needed. The outcome is simple and clear: relative to its bid, a generator can be over-paid but will never be under-paid. This is obviously problematic. The problem is two-fold. First, the guarantee of revenue adequacy via "make-whole" payments ostensibly favors the supply side over the demand side, thus violates the "equitable and two-sided" market principle that predicates the uniform price auction format in the first place. Second, if the recovery of the as-bid costs needs a guarantee, then why not directly pay the generators according to their bids, i.e., pay-as-bid? Paying at MCP and then making "make-whole" payments to match up with the as-bid cost seems artificial. In fact, the mere existence of, or reliance on, the "make-whole" payments is evidence that the uniform price payment design is flawed - the uniform MCP neither clears the market, nor reflects the true cost of electricity. Ramifications

are not limited to the unfairness and complexity (of having to use “make-whole” payments), but also open loopholes for market exploitation. For example, JPMorgan was found to game the California power market by exploiting the market rules. The essence of their lucrative strategy was to request a sky-high commitment fee while offering an extremely low price (-$30/MWh) for energy, enough to make the overall cost profile appear competitive so that their units get selected for dispatch. In the end, they are paid for energy at the MCP which was higher than their bid and at the same time reap the high commitment fee via "make-whole" payments. It is reported that JPMorgan amassed $57 million in improper payments over six months in 2010 and 2011 .

3.3 The Imperfect Two-sided Electricity Market

The classic pictorial economic analysis of the price-quantity relationship involves an upward-sloping marginal supply curve and a downward-sloping marginal demand curve. The two curves intersect at the market equilibrium point which identifies the transacted quantity and the market price and also maximizes the social welfare. Although the electricity market is designed to look like a two-sided market in which both suppliers and demanders bid, the suppliers and demanders are not at equitable places: On the supply side, it is relatively easy to estimate the marginal cost of generating a megawatt-hour of electricity, since in most cases, the cost is determined by the fuel cost and the unit’s efficiency. However, it is difficult for the demand-side, especially for the residential and commercial consumers which constitute a significant portion of the total demand, to identify the true marginal value of electricity. Many consumers regard electricity as an essential product and simply consume at whatever price that is passed to them (Kirschen, 2003). In turn, there is no way for their representative wholesale buyers, e.g., load-serving entities (LSE), to come up with bids that reflect the true marginal values of electricity with any accuracy. Therefore, the design of the electricity market as two-sided in the hope of maximizing social welfare is flawed - it does not correspond to reality. Telling evidence of inequitable market participation is that ISOs primarily use the

forecasted demand, in lieu of demand bids, as the input to the dispatch algorithms, whereas they use supply bids for the generation side.

Furthermore, even for those bidders who are able to quantify the marginal values, they are not treated comparably to the generators in the market clearing algorithm and the settlement rules. Specifically, the "consuming unit commitment" issue is practically ignored. Although it is not usually perceived this way, a consumer could legitimately have a commitment requirement. For example, a manufacturing plant might need to run its energy intensive process or unit for a continuous five hours, i.e., minimum run time requirement, to start the process it might take some preparation costs, i.e., start-up cost, and during each hour of the process, extra staff or other fixed costs to facilitate the electricity consumption might be required, analogous to the no-load cost on the generation side. Such a scenario, although hypothetical, raises an equity question: if the consumers are obliged to pay the generating unit commitment fees to generators, shouldn't the generators reciprocate by paying the consuming unit commitment fees to consumers? No ISO/RTO is currently accepting such kind of demand bids, although academic discussions have been around, for example, Su and Kirschen (2009) and Borghetti et al. (2002).

Finally, the energy market over the grid is nothing like the marketplace of an ordinary commodity. A central dispatcher or auctioneer is indispensable, and generation bids come in complicated forms due to the equipment's operating and cost characteristics. As discussed earlier, the classic economic principle simply does not work: a market equilibrium point may not exist.

3.4 Justification of Pay-as-bid in the UCED Context

Previous discussions about market design rarely give adequate consideration to the operational technicalities of the market settlement process. We believe that understanding the settlement algorithm is critical in the design of a sound payment rule.

The unit commitment economic dispatch model is usually solved as a MIP:

MIP formulation:

$$\min_{y\text{binary},x,z,\delta} \quad s^T y + c^T x \tag{3.1}$$

$$\text{s.t.} \quad My \geqslant b \tag{3.2}$$

$$Ax + Bz = d \tag{3.3}$$

$$Ex \geqslant Fy \tag{3.4}$$

$$Qx \geqslant q \tag{3.5}$$

$$Gz + H\delta \geqslant 0 \tag{3.6}$$

At the same time, the problem can also be viewed as a two-stage problem: the unit commitment decision is made in the first stage, whereas the dispatch decision is made in the second stage given the commitment decision. The mathematical model is as follows.
Two-stage formulation:

$$\min_{y \text{ binary}} \quad s^T y + \mathcal{Q}(y) \tag{3.7}$$

$$\text{s.t.} \quad My \geqslant b \tag{3.8}$$

where constraint (3.8) contains the minimum up- and down-time requirements and other commitment related constraints, and $\mathcal{Q}(y)$ is the optimal value of the second-stage economic dispatch (ED) problem:

$$\mathcal{Q}(y) = \min_{x,z,\delta} \quad c^T x \tag{3.9}$$

$$\text{s.t.} \quad Ax + Bz = d \ (\perp p) \tag{3.10}$$

$$Ex \geqslant Fy \ (\perp \eta \geqslant 0) \tag{3.11}$$

$$Qx \geqslant q \ (\perp \mu \geqslant 0) \tag{3.12}$$

$$Gz + H\delta \geqslant 0 \ (\perp u \geqslant 0) \tag{3.13}$$

The model minimizes the total as-bid production cost. Variable y is the com-

mitment decision, e.g., $y_{k,t} = 1$ if unit at node k is committed in time period t (for the remainder, subscripts are omitted as they are easily inferred and the model is presented in vector format). Variable x represents the dispatch, z the line flow and δ the voltage angle. Constraint (3.10) is the power balance equation with the demand d being a random parameter, (3.13) encapsulates all the network constraints, such as line flow equations and line thermal limits, (3.11) represents the lower and upper bound constraints on the power output x, which take the commitment y as a parameter, and (3.12) represents the ramping requirements. Independent of the model, the multiplier of each constraint is listed between the parentheses behind the corresponding constraint. At an optimal solution of the second stage problem, the value of the multiplier p of the constraint (3.10) is the MCP.

Although the make-whole payment rule translates to a total payment of $\max\{s^T y + c^T x,\ d^T p\}$, in stylized analyses of the existing payment structure, the total payment to generators includes the unit commitment cost $s^T y$ and the energy cost charged at the MCP, i.e., $d^T p$. It has been recognized that the solution to the above UCED model does not necessarily minimize the total payment $s^T y + d^T p$, which is the true cost of electricity from the consumers' point of view, see Jacobs (1997). Efforts have been made to minimize the consumer payment by solving the bi-level problem, e.g., Fernández-Blanco et al. (2012):

$$
\begin{aligned}
&\underset{y \text{ binary}, p}{\text{Min}} && s^T y + d^T p \\
&\text{s.t.} && My \geqslant b \\
& && p \text{ comes from (ED) model.}
\end{aligned}
$$

This model could be reformulated as a MIP with big-M constraints, which are needed to express the optimality conditions of the lower level problem. However, no efficient method has been reported to solve this MIP for large-scale instances, due to the inefficiency of the large number of big-M constraints. The bi-level model could also be reformulated as a nonlinear program which is nonetheless hard

to solve due to the non-convex constraints introduced to express the lower level optimality conditions. Even for toy problems, experiments have shown that the minimum payment solution obtained by this model is usually not a practically desirable solution.

We believe that the root cause of the mismatch between the minimum cost solution and the minimum payment solution lies in the pricing rule. In particular, only pricing the power balance constraints (3.10) and neglecting the marginal prices of the other constraints lacks justification from many perspectives. A theoretically correct way is to price all the ED constraints (3.10) to (3.13) with the corresponding multipliers p, η, μ and u, respectively. This will yield a total energy payment of

$$d^T p + (Fy)^T \eta + q^T \mu + 0^T u$$

and the payment minimization problem becomes

$$\begin{aligned} \min_{y \text{ binary}, p} \quad & s^T y + d^T p + (Fy)^T \eta + q^T \mu \\ \text{s.t.} \quad & My \geqslant b \\ & p, \eta, \mu \text{ come from (ED) model.} \end{aligned}$$

Note that the energy payment is the dual objective of the second stage LP. By LP duality, the above expression is equal to $c^T x$ at the optimal solution, hence the total payment becomes

$$c^T x + s^T y$$

which is exactly the total as-bid cost that is being minimized in UCED. On the individual level, it prompts a pay-as-bid scheme: pay a generator according to its bid, for both the unit commitment part and the energy part. This Pay-as-bid scheme not only eliminates the inconsistency between the minimum cost solution and the minimum payment solution, hence enables the ISO to achieve the minimum payment objective by simply solving the UCED model, but also induces accurate

valuation of generating assets (to be illustrated in the example given below), therefore enhances the basis for a truly competitive supply market. This approach is computationally convenient since UCED (minimum cost) is much easier to solve than the bi-level program (minimum payment).

The pay-as-bid scheme has been discussed extensively in the literature (Federico and Rahman, 2001; Anderson et al., 2013), but not adopted in the U.S. market. In fact, most electricity markets use the uniform-price auction format and only a few adopt the pay-as-bid format. For example, the electricity market in Britain and Iran switched to a pay-as-bid format in 2001 and 2003, respectively, and Italy has recently decided to follow suit. A similar move was considered in California in 2001 but was not implemented. We list the common reasons against the pay-as-bid scheme (between quotation marks), and our arguments against them.

1. "It distorts the competitive nature of the market by giving no incentive for technological innovation to suppliers, since the pay-as-bid resembles the cost-based pricing of the rate-of-return regulation (Cramton and Stoft, 2007)." The competitiveness of the market does not depend on the pricing scheme (as long as it is a fair one), but primarily comes from its openness, i.e., the transmission facility is no longer the property of a single generation firm or utility as in the past, instead, any firm willing to and capable of participating the market can have non-discriminatory grid access. As more and more suppliers enter the market, competitiveness is a natural result. To the contrary of the claim, since the pay-as-bid rule explicitly prices every component of the generators' operating and cost characteristics, the incentives for technological innovation and efficiency improvement will be more explicit.

2. "It forces the suppliers to depart from bidding their true marginal costs in order to make a profit, whereas it is believed that under the pay-at-MCP scheme, suppliers have every motivation to bid their marginal costs (Kahn et al., 2001)." First, a sound market design should allow suppliers to behave however they deem best for their interests, within the parameters of the market rules. It is acceptable and perfectly natural to include a profit margin in the

bid under the pay-as-bid scheme. A fuel-efficient generator could afford a higher profit margin while being competitive in its bid profile; likewise, an inefficient generator may have little or no profit. This is a sensible and healthy market outcome, and it poses no systematic discrimination toward suppliers in any tier (such as baseload and peakload). Second, the claim that "suppliers have every motivation to bid their marginal costs" is a textbook scenario and does not apply to the electricity market on the grid (as we demonstrate below). Numerous studies have been conducted on the supplier bidding strategies, and real world examples of manipulative market behavior abound and are happening, e.g., the JPMorgan manipulative bidding story.

3. "Under the pay-as-bid rule, a supplier's best offer is a price equal to its best predicted MCP. If all the suppliers were able to predict the MCP with 100% accuracy, pay-as-bid would result in the same market outcome as pay-at-MCP; otherwise, prediction inaccuracies would lead to dispatches departing from the 'merit order', and consumers would end up bearing the costs of such inefficiency ." If a single price were all that comprises a generator's offer, then all the unit commitment related issues would be gone and the above claim would be correct. However, the reality is that the generators submit multiple blocks of energy offerings each with a price and an incremental quantity, as well as unit commitment requirements that will affect the dispatch (as we also demonstrate below). Therefore, it is premature to assume the criticality of predicting the MCP (under the pay-as-bid rule, MCP means the highest accepted offer price). In fact, even knowing the MCP with certainty does not enable a generator to find an optimal bid in the sense that its profit will be maximized. For example, if a generator bids aggressively in the unit commitment parameters such as requiring a long minimum-up time, it might lose (the opportunity of being selected), irrespective of its price bid, to a competitor who bids modestly in the same parameter. It is also unrealistic to impose an extreme risk-seeking attitude on all suppliers, i.e., always striving to bid (and get paid) at the highest possible price and disregarding the risk

of not being dispatched at all. Furthermore, there is not a clear "merit order" without knowing the commitment, which by itself is a decision variable in the ISO's market clearing algorithm.

The following stylized example demonstrates the advantage of pay-as-bid in the unit commitment context. For simplicity, let us consider one time period and a generator (GEN1) with one block of price-quantity offer, in particular, its marginal cost is \$10/MWh for up to 100 MW, and a start-up cost of \$200. Suppose that the MCP is going to be \$15/MWh and will not be affected by GEN1's offer, be it accepted or not. (Note: This happens if the accepted quantity of the marginal offer is greater than GEN1's capacity of 100MW. In this case, if GEN1's offer is later accepted, it amounts to a deduction of 100MW from the marginal quantity and the marginal price remains unchanged.) Further assume that GEN1 knows all this information. How should GEN1 bid? For the system operator, the net benefit of accepting GEN1 is

$$(15 - p) \times 100 - 200$$

where p is the offer price of GEN1. Clearly, GEN1 will be accepted only if $p \leqslant 13$ and \$13/MWh is the market value of GEN1 (although the MCP is 15). Under pay-as-bid, GEN1 will optimally offer 13 and realize a profit of $(13 - 10) \times 100 = 300$. However, under pay-at-MCP, GEN1 could offer any price lower than 13 and get paid at the MCP of 15, obtaining a profit of $(15 - 10) \times 100 = 500$. This amounts to a \$200 over compensation to GEN1 and consumers will bear the cost.

In the same setting, now suppose that GEN1's marginal cost is 14 instead of 10. Since 14 > 13, GEN1 is not cost-effective and should not be accepted. Under pay-as-bid, GEN1 can do nothing about it; but under pay-at-MCP, it could still offer any price lower than 13 to get accepted and finally realize a profit of $(15-14) \times 100 = 100$. This would be a loss of efficiency at the consumers' cost.

For another scenario, if the dispatched quantity of the marginal offer was less than 100MW and the marginal unit had a higher start-up cost, e.g., \$400, then GEN1 could potentially replace the marginal unit. In this case, the net benefit of accepting

GEN1 will be

$$(15 - p) \times 100 - 200 + 400$$

which indicates that GEN1 could bid up to \$17/MWh and still get dispatched. This represents a case where a lower marginal-cost bid (i.e., 15) is not accepted while a higher marginal-cost bid (i.e., 17) is accepted. Clearly, the "merit order", if one exists, is not simply a ranking of the marginal costs in the bids.

It can be seen from the above examples that in the unit commitment context:

1. The MCP does not represent the market value of all accepted units.
2. Under pay-as-bid, a generator's best bid is not necessarily its predicted MCP.
3. The market outcome will not be the same under pay-as-bid and pay-at-MCP, even if participants know the MCP with certainty.
4. Under pay-as-bid, it is difficult to fool the system operator, as a generator's actual payment price is consistent with its bid price.

3.5 Suppliers' Response under Pay-as-bid

While there is abundant literature on bidding strategy and market equilibrium under various auction designs, the study of bidder's behavior under pay-as-bid in the UCED context is still limited. We leave a comprehensive discourse on this subject to future work and outline a general model accompanied by a simulation experiment, simply to make our point.

Let $O_i := (S_i, A_i, b_i, c_i, F_i, q_i)$ represent the offer (bidding) choice of unit i and let $\theta(O_i)$ be the return secured if offer O_i is accepted, which is easy to calculate. However, whether or not an offer O_i will be accepted depends on the circumstances (e.g., the system demand and other suppliers' offers), and from unit i's perspective, is a Bernoulli random variable (1 if accepted and 0 otherwise). The expected value of this random variable is the probability of O_i being selected, denoted by $p(O_i)$, which is to be approximated via repeated experiments (market participation). In

Table 3.1: Summary of the Generators by Fuel Types

Type	Count	Total MW	Energy Source Description
SUB	14	6059	Subbituminous Coal
NUC	5	4335	Nuclear (Uranium, Plutonium, Thorium)
BIT	10	3055	Anthracite Coal and Bituminous Coal
DFO	20	1123	Distillate Fuel Oil (Diesel, #1, #2 and #4 F. Oils)
RFO	1	450	Residual Fuel Oil (#5, #6 and Bunker C F. Oils)

the long run, as the market is at the state of (dynamic) equilibrium, $p(O_i)$ can be well approximated. Therefore, for a risk-neutral unit i, its profit maximization problem is

$$\max_{O_i \in \mathcal{O}} \theta(O_i) p(O_i) \tag{3.14}$$

where $\mathcal{O}$ is the feasible set of unit i's offer choices.

Realistically, $\theta(O_i)$ and $p(O_i)$ are inversely proportional, and their product is thus concave-shaped and has a maximum. For expositional purpose, we now consider a simple case where the energy bid price c_i is parameterized by a markup factor and other components of O_i are fixed. Based on this case, we show that (3.14) has a solution and that pay-as-bid recognizes and rewards bidders' technological advantage appropriately.

We use the PJM bidding data with masked generator names (publicized by FERC for research purposes). The original data set contains 1011 generators, from which we selected the first 50 generators for the simulation experiments. Table 3.1 lists the composition of the selected portfolio of generators.

We first demonstrate the relations between a generator's profit markup level in the bids and its realized profits. We pick a BIT generator "GEN6" as the subject, the characteristics and original bids of which are listed in Table 3.2 and 3.3. Keeping everything else constant, we multiply GEN6's energy bid prices by different markup factors, i.e., from 1 to 11 with increments of 0.5, and for each factor setting observe the realized profits in 100 randomized demand scenarios. Assuming that the

original bid prices (with markup factor equal to 1) represent the true marginal costs, we can compute the cost once we obtain the dispatch solution. Under the pay-as-bid rule, a generator's revenue is just the corresponding term in the objective function of the UCED model, and the profit is calculated by the formulae

$$\text{Profit} = \text{Revenue} - \text{Cost}$$

Table 3.2: Characteristics of GEN6

Min/Max MW	Ramp Up/Down	Min Up/Down	NoLoad	Start
16.5/86	0.495/0.551 (MW/min)	15/9 (hr)	1000	2000

Table 3.3: Marginal Costs of GEN6

	Block1	Block2	Block3	Block4
Quantity	37.5	11.25	7.5	29.75
Price	21.03	21.04	21.04	22.20

The demand scenarios are generated in two steps. First, we arbitrarily create a base demand curve for a 24-hour period using the following formula,

$$d_t = 10000 + 3000 \sin(2\pi t/24 - 1)$$

where d_t is the base demand in hour t. The sine function is used to mimic the demand variations throughout a day. Next, the actual demand in hour t is treated as a uniform random variable distributed in $[d_t - 100, d_t + 100]$ and 100 samples are drawn for each t to form 100 daily demand scenarios.

To illustrate the effects of a unit's commitment costs on its market competitiveness, we do the same experiments on "GEN5", which we make to have the same characteristics and cost parameters as GEN6, except for a much lower no-load and startup cost, one tenth of that of GEN6. Low commitment cost increases the cost-efficiency, hence competitiveness, of a unit, so we expect to see the advantage of

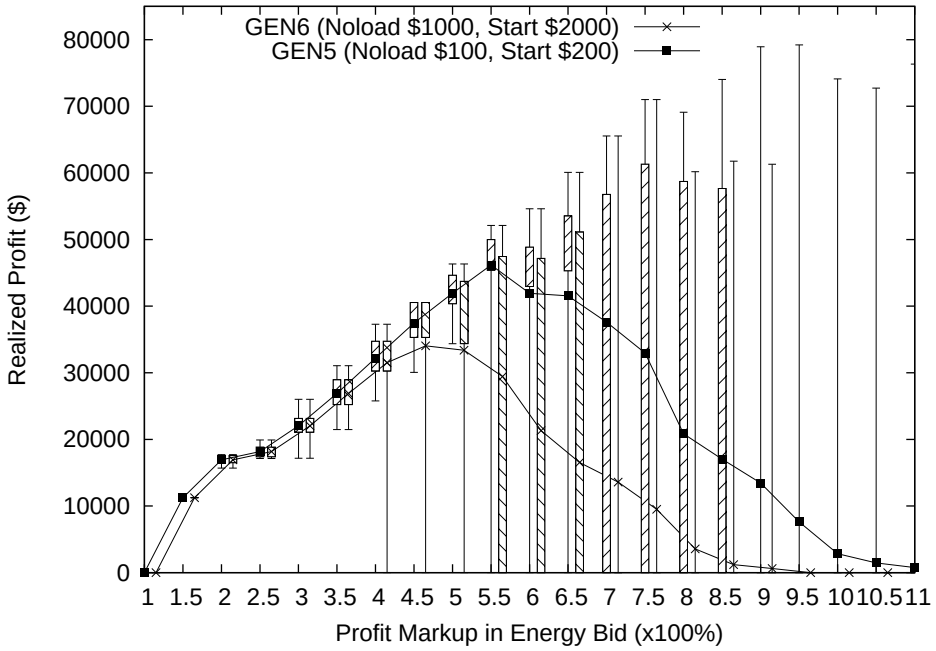


Figure 3.1: Profit curves of two generators with different commitment costs

such features of GEN5 over GEN6 in the experiments. The results are summarized in Figure 3.1.

In the figure, for each profit markup level on the horizontal axis, we plot the mean (the dots, which are connected by the curves), the 25% and 75% percentile (the lower and upper borders of the boxes, respectively), and the minimum and maximum (bottom and top points of the sticks, respectively) of the realized profit over the 100 demand scenarios. For the boxes that are invisible in the plot, they are actually concentrated (both upper and lower borders) at zero profit level. The mean values represent the expected profits, whereas the candlesticks to some extent indicate the risks. Rich insight can be drawn from the figures:

1. The expected profit is indeed a concave-shaped function of the profit markup in the bid and possesses a maximum. This validates the bidder's profit maximization model as discussed earlier. It can also be seen from the candlesticks

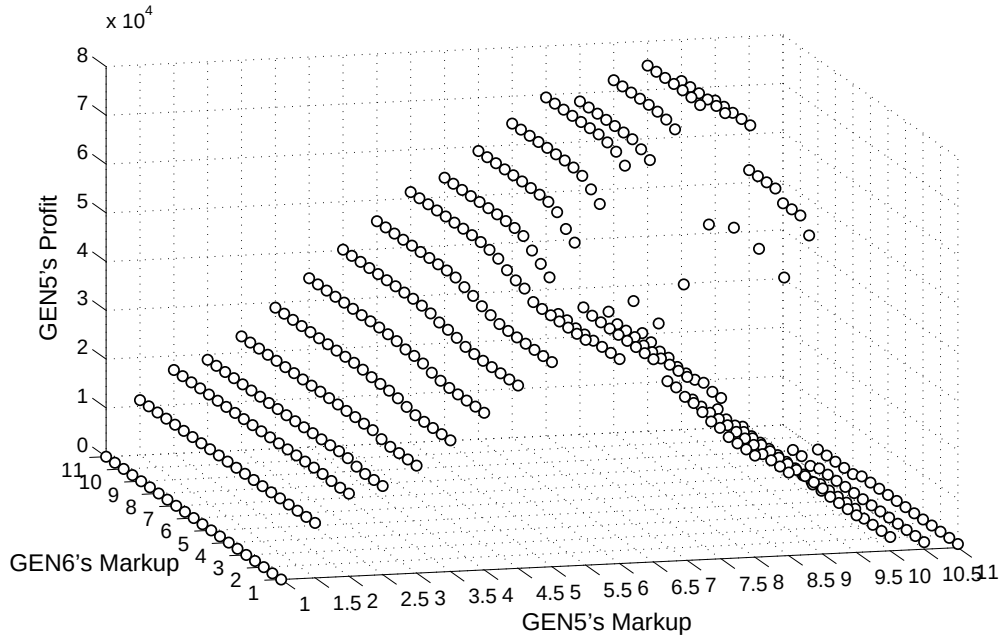


Figure 3.2: Payoff of GEN5 under pay-as-bid

that the higher the markup level, the riskier the bid will be, in the sense of a higher probability for the bid to be rejected.

2. GEN5 exhibits a more competitive profit curve as expected. GEN5 receives a higher maximum expected profit (at optimal markup 5.5) than that received by GEN6 (at optimal markup 4.5). The difference is $8206, or 11.6%. Besides, GEN5's profit curve has a wider and flatter top, which implies a broader risk tolerance for over-bidding. For example, GEN5 could tentatively bid at a markup level of 7 and not bear a big opportunity loss compared to the optimal bid, while bidding at 7 would be an immediate disaster to GEN6 as its 75% highest profit would be 0.

In Figure 3.2 and 3.3, we plot the expected profit (or payoff) of GEN5 and GEN6, respectively, as a function of the markup levels of both GEN5 and GEN6. Such payoff matrices and plots are useful tools to analyze the market equilibrium

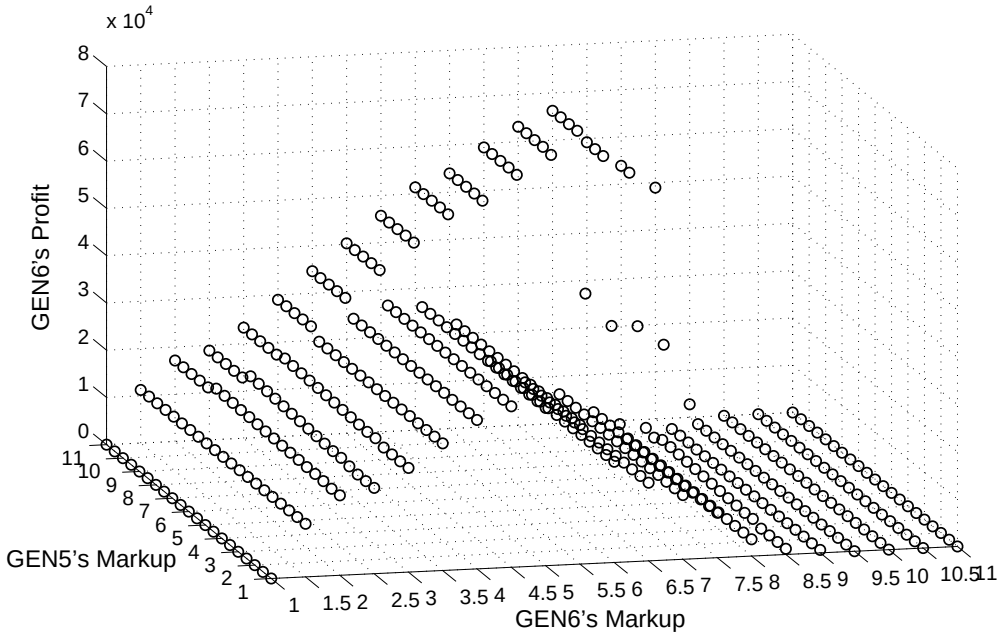


Figure 3.3: Payoff of GEN6 under pay-as-bid

in both the game theoretical and optimization framework. Although a detailed equilibrium study under the proposed pay-as-bid will be deferred to future work, some qualitative and sensible insight is readily available in the plots.

1. A generator's expected profit is consistently concave-shaped as a function of its own profit markup level. This can be seen by fixing the competitor's markup to any level and observing the resulting slice of the 3D plot. This once again validates the profit maximization model postulated earlier.

2. A generator's expected profit is also influenced by its competitor's bid. Take GEN5 (Figure 3.2) for example, when its competitor GEN6's markup is in the low range (which means higher competitiveness), GEN5's maximum profit is relatively low (around), this is because the market is shared with GEN6. As GEN6's raises its markup level, its competitiveness gradually diminishes and so does its market share, so GEN5 gains more market share and therefore

achieves a higher maximum profit (of around 7.5×10^4). Such effects are completely intuitive and the same pattern is present in Figure 3, too.

3. The differences between GEN5 and GEN6, including their maximum profit and the risk tolerance for over-bidding, are clearly exhibited in the two figures and are consistent with the observations from Figure 3.1.

3.6 Conclusion

The existing payment rules, i.e., paying at a uniform MCP for energy and relying on "make-whole" payments to recoup the unit commitment costs, has been shown problematic both in theory and in practice. The root cause of the problem is two-fold: (1) the unavoidable unit commitment constraints make a uniform MCP nonexistent; (2) the supply-side and demand-side do not behave, and are not treated, equitably in the market. In this context, we proposed pay-as-bid as a viable payment rule to be considered by policy makers.

In particular, we showed on the typical two-stage unit commitment model that pay-as-bid prices every component of a unit's energy production characteristics, fairly and accurately evaluates a unit's efficiency, and therefore provides clearer signals for innovations and improvements. Under the pay-as-bid rule, the consumer payment minimization problem, which is usually hard to solve, coincides with the production cost minimization problem that is easy to solve. Using a simple example, we demonstrated that suppliers' bidding behavior is rational under pay-as-bid, and a market equilibrium with a lower social cost is likely to exist and be achieved in the long run.

4 EXTENDED BIDDING STRUCTURE FOR DEMAND RESPONSE

4.1 Introduction

Sufficient demand-side participation is critical to the success of deregulated market design, since the marginal pricing and social welfare maximizing principles underlying this design are predicated on bid-based, competitive participation of both suppliers and demanders (Wellinghoff and Morenoff, 2007). However, reality has shown that the demand side lacks the ability to participate in the market comparably to the supply side and exhibits significant unexpressed elasticity, resulting in inefficient market outcomes, exacerbating oligopoly power and distorting long term investment incentives. There are two main causes. First, not all demanders are able to independently value the electricity ex ante, i.e., before the market clearing price is known, so as to place meaningful price-quantity bids on the market (Kirschen, 2003). This is inherent to the nature of electric energy, as most people regard electricity as an essential and non-substitutable commodity. Second, the bidding system does not provide other mechanisms as an alternative to the price-quantity bid format for demanders to express their willingness to consume, particularly their response to price signals. In fact, demanders can be quite responsive to the price and price variations by modifying and rescheduling usage. For instance, when the price is high, a demander could curtail some usage. Furthermore, if the demander knows a priori that the price is high in some hours of the day and low in other hours of the day, she could reschedule usage to minimize the total cost (Schweppe et al., 1988). Such behaviors are instances of demand response (DR). Incorporating ways in the market rules to induce demand response and encourage demand-side participation has drawn much attention recently from policy makers, practitioners and researchers.

FERC's Ruling on Demand Response

In its recent Order No. 745 (FERC, 2011), Federal Energy Regulatory Commission (FERC) requires that "when a demand response resource participating in an organized wholesale energy market administered by an RTO or ISO has the capability to balance supply and demand as an alternative to a generation resource and when dispatch of that demand response resource is cost-effective as determined by a net benefits test, that demand response resource must be compensated for the service it provides to the energy market at the market price for energy, referred to as the locational marginal price (LMP)". There are two prevalent interpretations of this DR compensation policy, but none is unanimously satisfactory.

The first interpretation allows DR resources to bid in the day-ahead energy market, i.e., the DR resources bid the quantity they are willing to curtail from their (presumably verifiable) expected consumption or baseline and the price for the curtailment. The DR bid is treated the same way as a supply offer in the market clearing economic dispatch algorithm. Cleared DR bids must follow the dispatch and will be compensated at the LMP. PJM RTO implements such a mechanism. In particular, PJM publishes a monthly updated threshold price calculated from certain net benefit criteria, and DR bids are included in the dispatch algorithm only when the LMP resulted otherwise exceeds the threshold.

This interpretation has been argued against by many economists: the DR resources are not entitled to sell energy in the market without physically or contractually owning the energy, see FERC (2011); Ruff (2002); Hogan (2012). A proposed solution is to require the DR resources to buy the baseline amount in an earlier settlement, e.g., futures market and forward contracts, refer to FERC (2011); Chao (2010, 2011); Hogan (2009, 2010a). However, in this case DR becomes no more than energy arbitrage between different markets, similar to the virtual bids between day-ahead and real-time markets. This does not serve the purpose DR is promoted for. The promotion of DR is aimed at eliciting better demand side participation in the market, achieving better social welfare and as a desirable side effect, relieving the strain on the transmission system caused by huge demand variations over time,

as well as damping the price fluctuations (Wellinghoff and Morenoff, 2007; FERC, 2011, 2008; Boisvert et al., 2002). In contrast, arbitrage could make the real-time price converge to the day-ahead price, but could not help reduce the variation of the day-ahead price.

The second interpretation does not treat DR as a sale of energy on the energy market. Instead, DR is treated as a sale of the "consuming right" from certain consumers, i.e., DR providers, to other consumers, i.e., the remaining load. In particular, the remaining consumers pay the DR provider to reduce consumption. When the supply curve is steep, such trades among the demand-side can be beneficial to all consumers, including DR providers who get compensation from the remaining load and the remaining load who enjoys lower LMP. This is done outside the energy market so there is no entitlement issue as in the first interpretation. ISO New England implements such a mechanism. In that market, demand reduction offers are cleared subject to a net benefit test after the day-ahead energy market results are determined, and the compensation level for the cleared DR is set to the LMP, see ISO New England (2012). The work in Chapter 2 implements exactly this interpretation of demand response.

We acknowledge some merits of the second interpretation: compared to the supply-side, electricity buyers are large in number and small in size, hence without a central organization it is impossible for them to have significant leverage on the market. In this context, ISO/RTO serves as an organizer to help the demand-side to form some market power to countervail the suppliers' market power. However, this amounts to a violation of the ISO/RTO's statutory role as an "independent" system operator and in the meantime, the efficiency of countervailing power is up for much debate, see Galbraith (1980), Stigler (1954) and their citing documents.

Other Related Work

A simple monetary compensation rule has not been, at least in theory, successful to elicit a satisfactory solution for the demand response problem. Another alternative is to design a bidding structure that accommodates distinct characteristics and

behaviors of the demand-side participants. Arroyo and Conejo (2002) presented a foundational work on the unit commitment based market clearing mechanism that has been widely adopted in today's markets. Importantly, the mechanism encouraged demanders to submit price-quantity bids to the market operator, instead of being treated as fixed and rigid. Strbac and Kirschen (1999) demonstrated the importance of a realistic demand-side bidding structure. They stressed that the cost of load recovery after, or occasionally before, the load reduction period should be accounted for in an optimal schedule. Su and Kirschen (2009) proposed a complex form of demand bids that allowed for flexible time of consumption. In particular, demanders could submit multiple price-quantity bids for each consumption period and specify the total amount of consumption to be satisfied over the scheduling horizon. However, those demand bids were modeled by integer variables and constraints, thus the dispatch mechanism fell short of good economic properties. Papadaskalopoulos et al. (2011) presented a decentralized market clearing mechanism in which each market participant computes her own optimal generation or consumption schedule and bids given the market prices and the central planner in turn updates the prices based on the bids from market participants. This is an iterative process and the iteration proceeds until an equilibrium is reached. We recognize a merit of this mechanism to be the great freedom available to market participants to interpret and respond to the price signals. However, if such freedom is uncontrolled, it may render the equilibrium nonexistent and the iterative process never converging. We believe that a certain degree of conformity is no less important than flexibility in the design of a bidding structure and adding new bidding formats can be a less drastic and easier to implement change than going to an iterative process.

In this chapter, we propose an extension to the existing price-quantity bid format for the ISO/RTO's economic dispatch model. The extended format enriches the forms of demand-side participation, promotes a broader frontier for load dispatchability and yet preserves the nice properties of the current market design philosophy, such as economic efficiency and incentive compatibility, see Stoft (2002) for a detailed discourse on market design. Following a brief note on the nomen-

clature, Section 4.2 proposes our characterization of different demand types and their respective cost-minimizing or surplus-maximizing problems. Based on this, Section 4.3 develops the new bidding structure and the corresponding central dispatch model, accompanied by the proof of its incentive compatibility. Section 4.4 implements the model for an experiment and presents the experiment results and Section 4.5 draws some conclusions and summarizes the points.

Notes on the Nomenclature

Symbols will be defined where they first appear in the chapter. In general, g and d denote generation and demand in megawatt hour (MWh), respectively, and p denotes the price in dollars/MWh. The superscript on a symbol annotates the specific meaning and the subscript(s) indexes its applicable object. Subscripts k and t index the participant and time period (i.e., hour), respectively. Depending on the context of its occurrence, a symbol may represent a scalar or a vector, with the specific meaning implied by the presence or absence of the subscripts. A symbol topped with a bar or bottomed with a underline is always a parameter instead of a variable, representing the upper or lower bound of a quantity.

4.2 Demand Types and Behavioral Models

In many ISO/RTOs' DR programs, demand response resources are treated comparably to a generation resource. For example, DR providers can specify operating requirements such as minimum curtailment period and DR initialization cost, etc. Energy bids are taken on a similar basis. Almost all ISO/RTOs in north America take demand-side energy bids exclusively in two forms[1]: (1) Fixed, specified by a quantity in MWh, and (2) Price-sensitive (or elastic), specified by a number of price-quantity pairs. These bids impose the demander either to be a price-taker, or to provide an explicit demand curve, which a normal demander and subsequently

[1]ISO/RTOs surveyed include: ISO New England, Midwest ISO, PJM RTO, New York ISO, California ISO and ERCOT. Note that fixed demand bids include the load estimates made by forecast procedures, such as ERCOT's load profiling process.

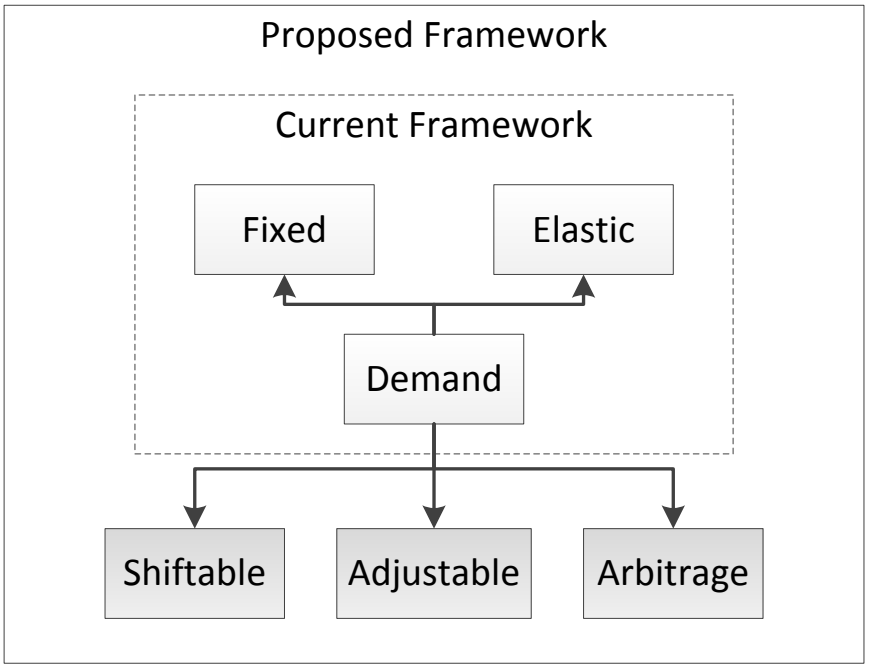


Figure 4.1: Framework for demand-side participation

her wholesale market representative, e.g., a load serving entity (LSE), are unable to estimate accurately, see, e.g., Kirschen (2003). Without the accuracy of this input, social welfare maximization is merely an illusion.

We identify three additional types of demand, in particular, shiftable, adjustable and arbitrage. We will formulate the basic characteristics and model the behaviors for each type of demand, while Figure 4.1 illustrates a structural overview of our work.

Fixed Demand

Fixed demand constitutes a dominant portion of the total demand on the spot market. For example in MISO's day-ahead market in 2008, fixed demand bids accounted for about 98% of total cleared demand (Newell and Hajos, 2010). By submitting a quantity without putting a maximum acceptable price, the bidder effectively tells the market that she places an infinite value on the whole, and each

and every bit, of the specified amount of electric energy. This is unlikely to be true and accurate in such an overwhelming scale, but it is what is happening on the market every day. Using fixed demand bids in cases where additional flexibility is present is contrary to efficiency and should be discouraged. Fixed demand bidders have nothing to optimize because they are unconcerned about the price.

Elastic Demand

Elastic demand exhibits a sloped demand curve. The value (or utility or benefit) is a concave function (decreasing marginal value) of the consumption d, denoted by $V(d)$. Note that the value function can be different for different time periods, but it is separable with respect to the time of consumption. The surplus maximization problem of an elastic demander k is (ELA)(p):

$$\max_{d_k} \quad \sum_t [V_{k,t}(d_{k,t}) - p_t d_{k,t}] \tag{4.1}$$

$$\text{s.t.} \quad \underline{d}_{k,t} \leqslant d_{k,t} \leqslant \bar{d}_{k,t}, \ \forall t \tag{4.2}$$

Typical forms of $V(d)$, like those of the generator cost function $C(g)$, are quadratic or piecewise linear.

Adjustable Demand

Similar to the fixed demand, adjustable demand has a preferred consumption profile, but is willing to make an adjustment at a cost. Let $r^+_{k,t}$ and $r^-_{k,t}$ denote the amount of over- (adjust up) and under- (adjust down) consumption from the target level $d^{ta}_{k,t}$, respectively, and let $D_{k,t}(r^+_{k,t}, r^-_{k,t})$ denote the deviation cost. Over-consumption does not normally incur extra costs, if not making extra benefits, on the demander's side, and we include its cost here simply for the generality of the formulation. Compared to the value function of an elastic demand, the deviation cost function is an alternative valuation of electric energy, also termed the Value

of Lost Load (VOLL) see the term definition in Kirschen (2003); Stoft (2002). An adjustable demander minimizes the cost of consumption by solving (ADJ)(p):

$$\min_{r_k^+, r_k^-} \quad \sum_t [p_t(d_{k,t}^{ta} + r_{k,t}^+ - r_{k,t}^-) + D_{k,t}(r_{k,t}^+, r_{k,t}^-)] \tag{4.3}$$

$$\text{s.t.} \quad 0 \leqslant r_{k,t}^+ \leqslant \bar{r}_{k,t}^+, \ \forall t \tag{4.4}$$

$$0 \leqslant r_{k,t}^- \leqslant \bar{r}_{k,t}^-, \ \forall t \tag{4.5}$$

Realistically, the parameter $\bar{r}_{k,t}^-$ is upper bounded by $d_{k,t}^{ta}$. Note that $D_{k,t}(r_{k,t}^+, r_{k,t}^-)$ is assumed to be a convex function and takes value zero when $r_{k,t}^+$ and $r_{k,t}^-$ are both zero. We envision a typical form of $D_{k,t}(r_{k,t}^+, r_{k,t}^-)$ to be:

$$D_{k,t}(r_{k,t}^+, r_{k,t}^-) = \alpha_{k,t}^+(r_{k,t}^+)^2 + \beta_{k,t}^+|r_{k,t}^+| + \alpha_{k,t}^-(r_{k,t}^-)^2 + \beta_{k,t}^-|r_{k,t}^-| \tag{4.6}$$

where α and β are parameters.

Shiftable Demand

Shiftable demand requires a total amount of electricity to be delivered within a given time range, but is flexible with regard to the time of delivery within that range. For instance, demander k partitions the planning horizon T into time ranges indexed by m and requires $d_{k,m}^{tr}$ amount to be delivered within the time range $T_{k,m} \subset T$. A shiftable demander minimizes her consumption cost by solving (SHI)(p):

$$\min_{d_k^{sh}} \quad \sum_t p_t d_{k,t}^{sh} \tag{4.7}$$

$$\text{s.t.} \quad \sum_{t \in T_{k,m}} d_{k,t}^{sh} = d_{k,m}^{tr}, \ \forall m, T_{k,m} \tag{4.8}$$

$$\underline{d}_{k,t}^{sh} \leqslant d_{k,t}^{sh} \leqslant \bar{d}_{k,t}^{sh}, \ \forall t \tag{4.9}$$

The shiftable demand bid requires no explicit valuation of the electricity and opens a door for demanders to respond to the market prices. It can be expected to substitute for an appreciable portion of the fixed demand and hence increase the general dispatchability of the demand. Typical shiftable loads include plug-in electric vehicles (PEV) and their aggregators, industrial laundry facilities and sewage treatment plants, etc.

Arbitrage

Arbitrage here means physical (instead of financial) arbitrage over time in a given market (instead of between different markets). A storage facility is a typical arbitrage type of demand (Walawalkar et al., 2007). An arbitrageur seeks to profit from the price discrepancies over time – buy energy when the price is low, store it, and sell when the price is high. There are no target levels of storage and no deviation penalties, but there is efficiency loss in the charge-discharge cycles. Let $s_{k,t}$ and $b_{k,t}$ denote sell (discharge) and buy (charge), respectively, and $h_{k,t}$ denote the storage level. An arbitrageur maximizes its profit by solving (ARB)(p):

$$\max_{b_k, s_k, h_k} \quad \sum_t p_t(s_{k,t} - b_{k,t}) \tag{4.10}$$

$$\text{s.t.} \quad h_{k,t} = h_{k,t-1} + b_{k,t}e_k - s_{k,t}, \quad \forall t \tag{4.11}$$

$$h_{k,1} = h_{k,|T|} \tag{4.12}$$

$$0 \leqslant b_{k,t} \leqslant \bar{b}_k, \quad \forall t \tag{4.13}$$

$$0 \leqslant s_{k,t} \leqslant \bar{s}_k, \quad \forall t \tag{4.14}$$

$$0 \leqslant h_{k,t} \leqslant \bar{h}_k, \quad \forall t \tag{4.15}$$

In the defining equation (4.11) for $h_{k,t}$, e_k is the efficiency factor with $e_k \in [0, 1]$, indicating that each unit of energy input will convert to e_k unit of output. Realistically, e_k may be a function of h_k, e.g., the efficiency of a Sodium Sulfur (NaS) battery depends on the depth of discharge (J. Himelic, 2011), which needs more

constraints to express. For expositional purpose, we make e_k a constant bidding parameter. Constraint (4.12) nails the net change of h_k in the planning horizon to zero, for sustainable operations, although in practice it can appear in different forms.

Note that we do not aim to enumerate all possible demand characteristics and the above nominated types are not strictly exclusive to one another. For example, the elastic demand and adjustable demand share a similar basis for valuation (i.e., both have no intertemporal component) and are mathematically generalizable to one form. The important point is that when demanders, despite their formal differences, all naturally behave as if they are solving a convex minimization problem, we can open up the existing bidding structure to explicitly account for these natural behaviors, without sacrificing its nice properties. This will be addressed in the next section.

4.3 Bidding and Central Dispatch Model

While market participants have their own optimal response to the prices, the actual dispatch and the market clearing prices are determined by the central auctioneer (ISO/RTO), whose objective is maximizing the social welfare. If a dispatch and pricing model is designed such that the central dispatch solution with the accompanying prices coincides with the market participants' optimal response to these prices, then competitive participants have every reason to bid their true parameters, thus the model is incentive compatible. We will develop such a model incorporating the above mentioned demand types.

Central Model and its Properties

Table 4.1 lists the parameters and variables in the model, with subscripts omitted for clarity. The parameters represent the bids submitted to the system operator.

In a distributed decision-making paradigm, given the market clearing prices p_t, demanders solve their respective behaviorial models presented in the last section.

Table 4.1: Bidding Parameters and Decision Variables

Type	Bidding Parameters	Variables
Generator	$C(\cdot)$, $\underline{g}$, $\bar{g}$, R_k^U, R_k^D	g
Fixed	d^{fx}	
Elastic	$V(\cdot)$, $\underline{d}$, $\bar{d}$	d
Shiftable	T_m, d^{tr}, $\underline{d}^{sh}$, $\bar{d}^{sh}$	d^{sh}
Adjustable	d^{ta}, $D(\cdot)$, $\bar{r}^+$, $\bar{r}^-$	r^+, r^-
Arbitrage	e, $\bar{b}$, $\bar{s}$, $\bar{h}$	b, s, h

On a similar basis, generator k responds to the price p by solving (GEN)(p):

$$\max_{g_k} \quad \sum_t [p_t g_{k,t} - C_{k,t}(g_{k,t})] \tag{4.16}$$

$$\text{s.t.} \quad \underline{g}_{k,t} \leqslant g_{k,t} \leqslant \bar{g}_{k,t}, \ \forall t \tag{4.17}$$

$$g_{k,t} - g_{k,t-1} \leqslant R_k^U, \ \forall t \tag{4.18}$$

$$g_{k,t-1} - g_{k,t} \leqslant R_k^D, \ \forall t \tag{4.19}$$

where R_k^U and R_k^D are the ramp-up and ramp-down rates (in MW/hour), respectively. The system operator maintains the supply-demand balance

$$\sum_k (g_{k,t} - d_{k,t} - d_{k,t}^{sh} - r_{k,t}^+ + r_{k,t}^- + s_{k,t} - b_{k,t}) = \sum_k (d_{k,t}^{fx} + d_{k,t}^{ta}), \quad \forall t \tag{4.20}$$

by adjusting the prices p_t.

We postulate a central dispatch model, as follows.
(Central Model):

$$\min_{\substack{g,d,d^{sh},r^+ \\ r^-,b,s,h}} \quad \sum_{k,t} [C_{k,t}(g_{k,t}) - V_{k,t}(d_{k,t}) + D_{k,t}(r_{k,t})]$$

$$\text{s.t.} \quad \text{(4.2), (4.8), (4.9), (4.4), (4.5)}$$

$$\text{(4.11)-(4.15), (4.17)-(4.20)}$$

The price p_t is set as the optimal Lagrangian multiplier (or dual variable) of the corresponding constraint in (4.20). Note that the model minimizes the total social cost (negative of the social welfare), hence it is economically efficient.

Theorem 4.1. *Given a set of bidding parameters, suppose that* $\hat{x} := (\hat{g}, \hat{d}, \hat{d}^{sh}, \hat{r}^+, \hat{r}^-, \hat{b}, \hat{s}, \hat{h})$ *solves the Central Model and* $\hat{p}$ *is the optimal Lagrangian multiplier of the constraint (4.20). Then* $\hat{g}$ *solves (GEN)*$(\hat{p})$*,* $\hat{d}$ *solves (ELA)*$(\hat{p})$*,* $\hat{d}^{sh}$ *solves (SHI)*$(\hat{p})$*,* $(\hat{r}^+, \hat{r}^-)$ *solves (ADJ)*$(\hat{p})$*, and* $(\hat{b}, \hat{s}, \hat{h})$ *solves (ARB)*$(\hat{p})$*.*

Proof. By duality theory, we know that $(\hat{x}, \hat{p})$ solves the Wolfe dual, formulated by dualizing constraint (4.20), of the Central Model:

$$\begin{aligned}
\max_{p}\min_{x} \quad & \sum_{k,t}[C_{k,t}(g_{k,t}) - V_{k,t}(d_{k,t}) + D_{k,t}(r_{k,t})] \\
& + \sum_{t} p_t[\sum_{k}(g_{k,t} - d_{k,t} - d^{sh}_{k,t} - r^+_{k,t} + r^-_{k,t} \\
& \qquad + s_{k,t} - b_{k,t} - d^{fx}_{k,t} - d^{ta}_{k,t})] \\
\text{s.t.} \quad & \text{(4.2), (4.8), (4.9), (4.4), (4.5), (4.11)-(4.15), (4.17)-(4.19)}
\end{aligned}$$

Consequently, $\hat{x}$ solves

$$\begin{aligned}
\min_{x} \quad & \sum_{k,t}[C_{k,t}(g_{k,t}) - V_{k,t}(d_{k,t}) + D_{k,t}(r_{k,t})] \\
& + \sum_{t} \hat{p}_t[\sum_{k}(g_{k,t} - d_{k,t} - d^{sh}_{k,t} - r^+_{k,t} + r^-_{k,t} \\
& \qquad + s_{k,t} - b_{k,t} - d^{fx}_{k,t} - d^{ta}_{k,t})] \\
\text{s.t.} \quad & \text{(4.2), (4.8), (4.9), (4.4), (4.5), (4.11)-(4.15), (4.17)-(4.19)}
\end{aligned}$$

which is a separable model by participant types, i.e., can be decomposed into (GEN)$(\hat{p})$, (ELA)$(\hat{p})$, (SHI)$(\hat{p})$, (ADJ)$(\hat{p})$ and (ARB)$(\hat{p})$, thus the conclusion follows. □

It is widely believed that this property of the economic dispatch model, coupled with the reality that nonconvex cost (e.g., unit commitment cost) is relatively minor,

makes the existing bidding structure incentive compatible, see Stoft (2002). This leads to the conclusion that the extended Central Model is incentive compatible.

Abstraction

While the specific formats proposed above focus on the demand side, the structure can be applied to both sides of the market. For example, a hydro generator may have time-shiftable supply needs. In the abstract form, each market participant k has a benefit function $f_k(x_k)$ and operating constraint $x_k \in X_k$, where x_k is the energy consumption/supply. The participant's optimal response to the market price p is

$$\max_{x_k \in X_k} \quad f_k(x_k) - x_k^\top p \tag{4.21}$$

Note that time dimension is embedded in the vectors x_k and p, so all kinds of intertemporal relations can be expressed in the objective function as well as in the constraint X_k. In the bid-based central dispatch mechanism, each participant k simply informs (via bidding) the dispatcher its $f_k(\cdot)$ and X_k, and the dispatcher maximizes the social welfare by solving

$$\max_{x} \quad \sum_k f_k(x_k) \tag{4.22}$$

$$\text{s.t} \quad \sum_k x_k = 0 \ (\perp p) \tag{4.23}$$

$$x_k \in X_k, \ \forall k \tag{4.24}$$

The existing market model (where only fixed and elastic bids are allowed) is a special case of this formulation, having two specialties: (1) the value function f is separable across time, thus $f_k(x_k)$ is restricted to the form $\sum_t f_{k,t}(x_{k,t})$; (2) the constraint set X_k of a demander k is also separable across time, i.e., $X_k = \prod_t X_{k,t}$. These restrictions hinder efficient market participation. For example, a shiftable demander with no way to express the shiftability in bids may have to predict the

price path so as to approximate this feature using the time-separable price-quantity bids. The prediction and approximation are error-prone and most likely to lead to suboptimal outcomes.

In contrast, the general model avoids such barriers and retains nice properties. It is straightforward to generalize that as long as each $f_k(\cdot)$ is a convex function and each X_k is a convex set, the economic properties will hold and the model will remain easy to solve.

Two Additional Merits

There are two related points that we need to clarify:

Network Integration

The above framework is developed only on an economic basis, devoid of the transmission network variables and constraints. This is purely for the clarity of the main point. In fact, the framework can be easily adapted to a DC-based (linearly constrained) network model, and the nice properties will hold as well. Suppose the network is represented by a set of nodes $\mathcal{N}$ and a set of arcs $\mathcal{A}$ (each physical transmission line is modeled by two arcs, one for each direction). Let variable z denote the power flow on arcs, bounded within the thermal limits $[-\bar{z}, \bar{z}]$, variable δ denote the voltage angle at nodes and parameter B denote the susceptance of arcs. Then the system operator maintains the arc flow equation and the nodal power balance, as follows:

$$z_{k,l,t} - B_{k,l}(\delta_{l,t} - \delta_{k,t}) = 0,\ \forall (k,l) \in \mathcal{A}, t \tag{4.25}$$

$$g_{k,t} - d_{k,t} - d^{sh}_{k,t} - r^{+}_{k,t} + r^{-}_{k,t} + s_{k,t} - b_{k,t} - \sum_{l:(k,l)\in\mathcal{A}} z_{k,l,t} = d^{fx}_{k,t} + d^{ta}_{k,t},\ \forall k, t \tag{4.26}$$

It is easy to see that these additional variables and linear equations can be readily incorporated in the central model.

Unit Commitment

In practice, the economic dispatch is usually preceded by the unit commitment (UC) process (i.e., to decide which generators are to be used in the dispatch, based on costs and operating characteristics), which shapes the feasible set of the economic dispatch problem. In the proposed bidding context, the unit commitment process can be performed by taking all the bidding demand, i.e., d^{fx}, $\bar{d}$, $\bar{d}^{sh}$, d^{ta}, $\bar{b}$, as fixed demand, and we claim that the UC decision thus obtained is guaranteed to be feasible for the subsequent central dispatch model. To see this, simply note that our central dispatch model boasts a relaxed feasible region compared to the conventional one where all demands are taken as fixed, and that the fixed demand is a feasible solution to the Central Model.

The UC decision obtained in the above way may not be the optimal one to the unit commitment model formulated directly based on the Central Model, although one can solve such a UC model if an "optimal" UC solution is desired. However, we offer an important caveat: the unit commitment model, which is usually a mixed integer program, lacks economic justification for the market clearing function, see Johnson et al. (1997); O'Neill et al. (2005), which is part of the reason why unit commitment and economic dispatch are usually practiced as two decision processes rather than one.

4.4 Implementation and Experiments

While the proposed model opens up new ways for demand bidding, the actual penetration rate of the new demand forms is yet to see and the exact bidding parameters are still unknown. These parameters are set fictitiously in the experiments. Therefore, the experimental results of this section should be assimilated as a qual-

itative, rather than rigorously quantitative, projection of the current and future states of the market.

Data and Setting

The generator bids and the fixed demands are obtained from the FERC eLibrary Docket Number AD10-12, ACCNNUM 20120222-4012. The data set represents a typical summer operating day of the PJM day-ahead market (Krall et al., 2012). For the demand data, we sum up the fixed demand bids from all the 13760 buses for each hour to create an aggregate hourly demand profile, for use as the base case in the experiments[2]. The base case is illustrated in Figure 4.2 as the "Fixed" demand. For the generator data, there are altogether 1011 generators, each offering up to 10 pairs of price-quantity bids for energy along with various unit commitment requirements and costs. A unit commitment process similar to the one documented in (Krall et al., 2012) was executed on the base-case demand, which selected 365 generators for commitment. We fix the unit commitment status according to this result in the subsequent experiments.

We make up four aggregate demanders, one for each demand type. The omission of subscript k in the following should cause no confusion.

Elastic Demand

We assume that 1% of each hour's base-case demand becomes elastic, which is then bid into the market in ten equally sized MWh blocks, coupled respectively with 10 decreasing prices ranging from \$99/MWh to \$0/MWh with even decrements, see Figure 4.3 for an illustration. This piece-wise linear demand curve for hour t is represented by a linear cost function $V_t(d_t)$ and two linear constraints in the

[2]There are also price-responsive demand bids, demand response bids and incremental and decremental virtual bids in the data file. We disregard them because (1) they are negligible in quantity, (2) the on-going demand response rule is unclear and controversial, and (3) virtual bids are irrelevant to our topic. We also disregard the network data because it is inaccessible to the public.

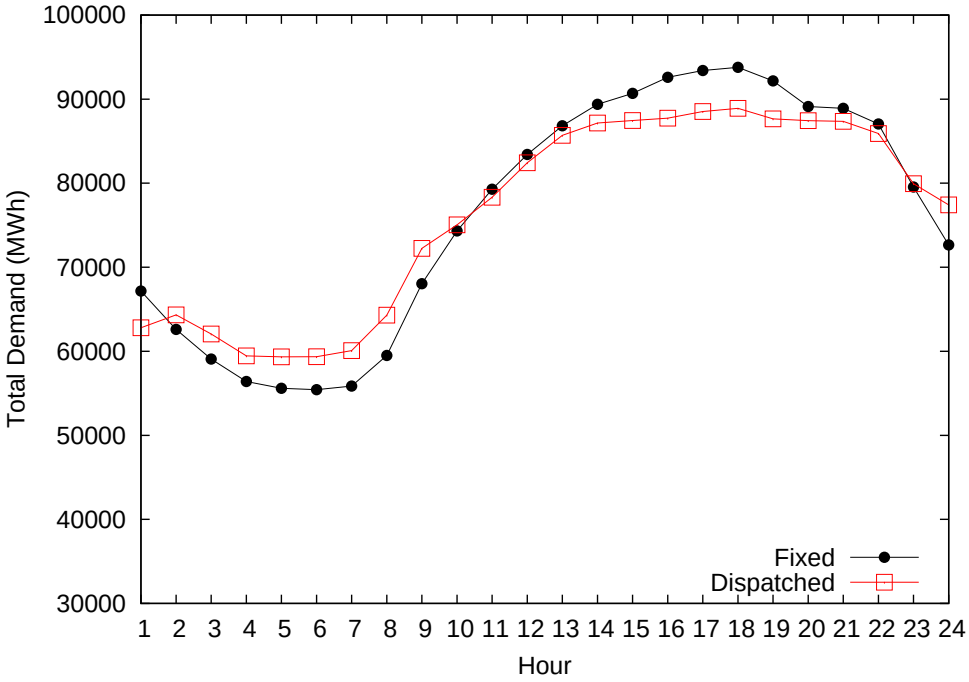


Figure 4.2: Day-ahead demand profile of FERC dataset 4012

minimization problem, as follows.

$$V_t(d_t) = \sum_{o \in \mathcal{O}} p^{db}_{t,o} d^{db}_{t,o} \tag{4.27}$$

$$d_t = \sum_{o \in \mathcal{O}} d^{db}_{t,o} \tag{4.28}$$

$$d^{db}_{t,o} \leqslant \bar{d}^{db}_{t,o}, \ \forall o \in \mathcal{O} \tag{4.29}$$

where $\mathcal{O}$ is the set of bid blocks, the bidding pair $(p^{db}_{t,o}, \bar{d}^{db}_{t,o})$ indicates that an increment of $\bar{d}^{db}_{t,o}$ MWh is worth $p^{db}_{t,o}$ dollars/MWh to the demander, and the variable $d^{db}_{t,o}$ represents the dispatched quantity in bid block o.

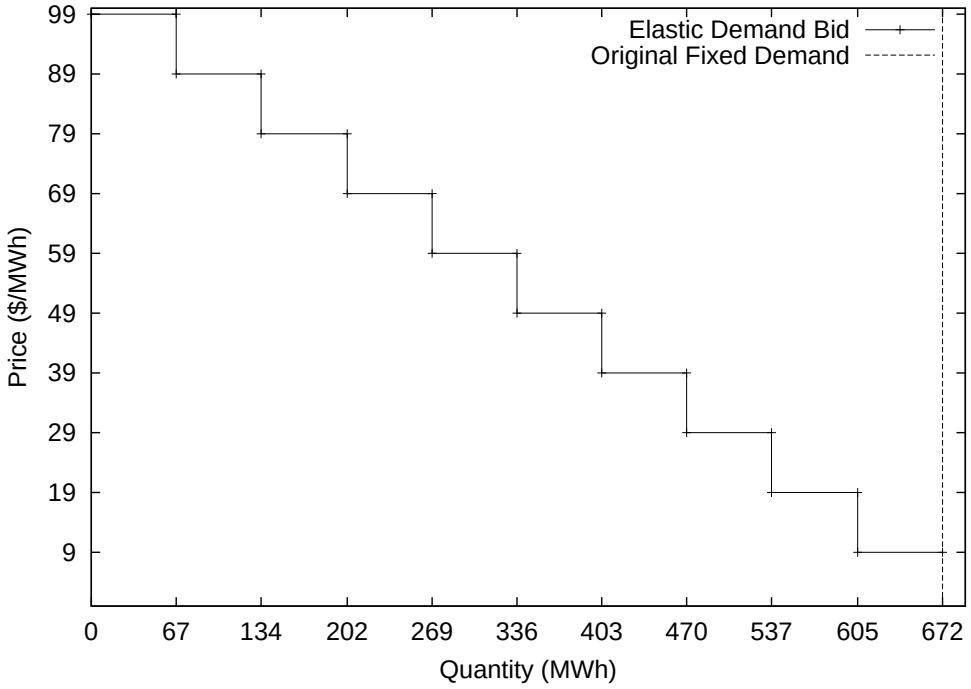


Figure 4.3: Elastic demand bid for hour 1

Adjustable Demand

We assume 1% of each hour's base-case demand becomes the target level p_t^{ta} of the adjustable demand. The deviation function $D_t(r_t^+, r_t^-)$ is taken in the form of (4.6), with the linear penalty β_t^+ and β_t^- arbitrarily set to the minimum LMP 0 and the average LMP 30.1 of the base-case, respectively, and the quadratic penalty α_t^+ and α_t^- arbitrarily set to 0.05 and 0.1, respectively. The bound $\bar{r}_t^-$ is set equal to p_t^{ta} while $\bar{r}_t^-$ is set to $\sum_t p_t^{ta}$.

Shiftable Demand

We partition the 24-hour period into three 8-hour ranges, i.e., T_m, $m = 1, 2, 3$, and assume 1% penetration of shiftable demand by setting the total demand requirement d_m^{tr} for range m to be 1% of the sum of the hourly base-case demand in the range.

Arbitrage

We assume an arbitrageur (storage) the size of 1% of the base-case demand is present besides the base-case demand and set $\bar{h}$ accordingly. We set the hourly buy (charging) rate $\bar{b}$ and sell (discharging) rate $\bar{s}$ to be $0.2\bar{h}$, to mimic the characteristics of a 5-hour storage facility. The efficiency factor e is set to 0.75.

Comparative Effect of Different Demand Types

We tested the effect on LMP and social welfare of 1% penetration of the outlined forms of demand-side bids, separately and aggregatively. The elastic, shiftable and adjustable demands are substitutes for the fixed demand, so the fixed demand will reduce to 99% of the original level in these individual cases. The arbitrage is an additional form of participation on top of the base-case demand, so the base-case demand remains at the 100% level. We examined two aggregative cases, both consisting of 97% fixed demand and 1% each of the elastic, shiftable and adjustable demand, one with 1% arbitrage and the other without arbitrage. The actual dispatched demand of the "97% Fixed + 1% (E+S+A+AR)" case is plotted in Figure 4.2 as the "Dispatched" curve.

Effect on the LMP

Figure 4.4 plots the LMP resulted from each case. As expected, the base case exhibits the roughest (with the biggest dip and spike) price path while the aggregative case exhibits the mildest. The penetration of each individual demand type smoothens the LMP to a certain extent. Among them, arbitrage is the most effective, followed by shiftable demand, whereas elastic demand is the least effective, in terms of dampening the price fluctuation.

Effect on the Social Welfare

Table 4.2 lists the cost (negative of the social welfare) results. The first column indicates the hypothesized market composition, the second column is the cost from

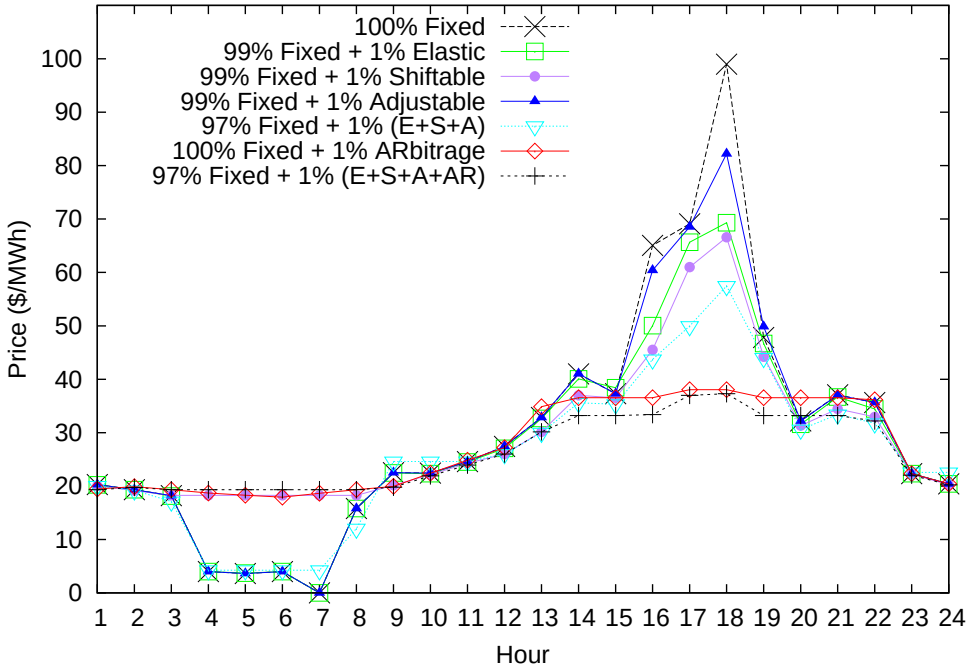


Figure 4.4: Effect of extended bidding on LMP

Table 4.2: Cost Results

	Current	Optimal	Saving	%Saving
1% **E**lastic	23317039	23215798	101242	0.43%
1% **S**hiftable	24315018	24069303	245715	1.01%
1% **A**djustable	24315018	24299083	15935	0.07%
1% (**E**+**S**+**A**)	23317039	22991408	325632	1.40%
1% **AR**bitrage	24315018	23748933	566085	2.33%
1% (**E**+**S**+**A**+**AR**)	23317040	22566391	750649	3.22%

the current bidding design, i.e., treating all demand as fixed, the third column is the optimal cost from our proposed bidding design, and the fourth and the fifth columns compare the costs and list the savings and percent savings, respectively. The benefit of the proposed bidding design is apparent and significant.

Arbitrage Effect on the LMP and Profit

As demonstrated above, arbitrage is the most impactive on the LMP among other participant types of the same penetration level of 1%. This is fathomable, as an arbitrageur's buy/sell schedule is driven solely by the temporal price differences and is unfettered by any target level of consumption or private valuation of the electric energy (because practically there are none). However, unlike the other types of demand bids which are direct alternatives or substitutes for the fixed demand bid, the arbitrage bid must be backed by physical storage capability that takes time to construct and deploy, so the penetration level is likely to be small in the foreseeable future.

In Figure 4.5, we plotted the effect of arbitrage on the LMP for different penetration levels, ranging from 0.2% to 1%. As expected, the increase of the arbitrage level will gradually damp the LMP variation. It is also interesting to note that the effect does not grow linearly with the penetration level, as the first 0.2% increment of the arbitrage level has contributed about half of the peak price reduction. This observation prompts a question: what is the "optimal" percentage of storage on the market? Figure 4.6 below provides some useful information to address this question.

In Figure 4.6, we plotted the profits of arbitrage for penetration levels ranging from 0% to 2% with an increment of 0.1%, and for three different efficiency factors, 0.65, 0.75 and 0.85. Seen from the figure, high marginal value of storage expansion can be expected when the level is below $0.4 \sim 0.6\%$ for all three efficiency options. From a level higher than 0.6%, the marginal benefit of expanding storage capacity starts to decrease, plateau or even reverse sign, depending on the technology type (efficiency factor). Of course, in making the storage expansion decision, construction and operation costs and a myriad of other factors need to be considered, but the above observation at least shed some light on such a decision-making process.

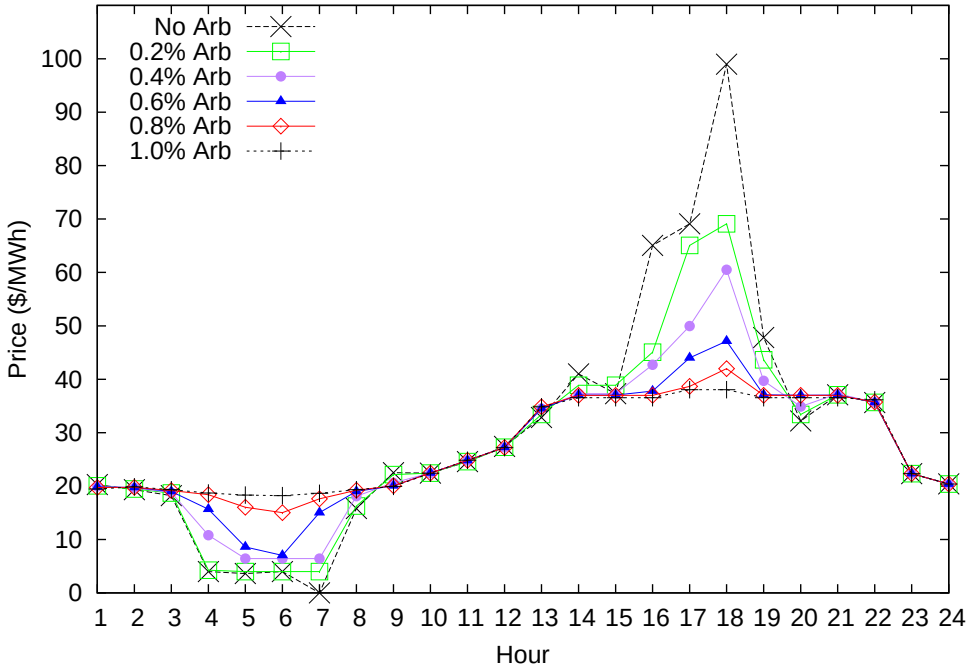


Figure 4.5: LMP for different arbitrage levels

4.5 Conclusion

The existing demand response compensation policy has been widely and fiercely questioned for its economic efficiency, equality and fairness. Recognizing that a simple monetary compensation rule is unlikely to settle the issue, we proposed an alternative route to reach the end – opening up the bidding structure to allow for more forms of bids that reflect realistic demand characteristics and behaviors. Specifically, existing bid formats are all separable over time. But a significant and growing segment of demand can be shifted across time and therefore has no way to bid its true valuation of consumption. We proposed additional bid types that allow time-shiftable demand to better express its value, thus elicit demand response in the most natural way – direct participation in the market. The additional bid types are easily incorporated into the existing market and that they preserve its efficiency and incentive-compatibility properties, both of which are critical design principles that must be instantiated, but are commonly seen violated, in ISO/RTO's

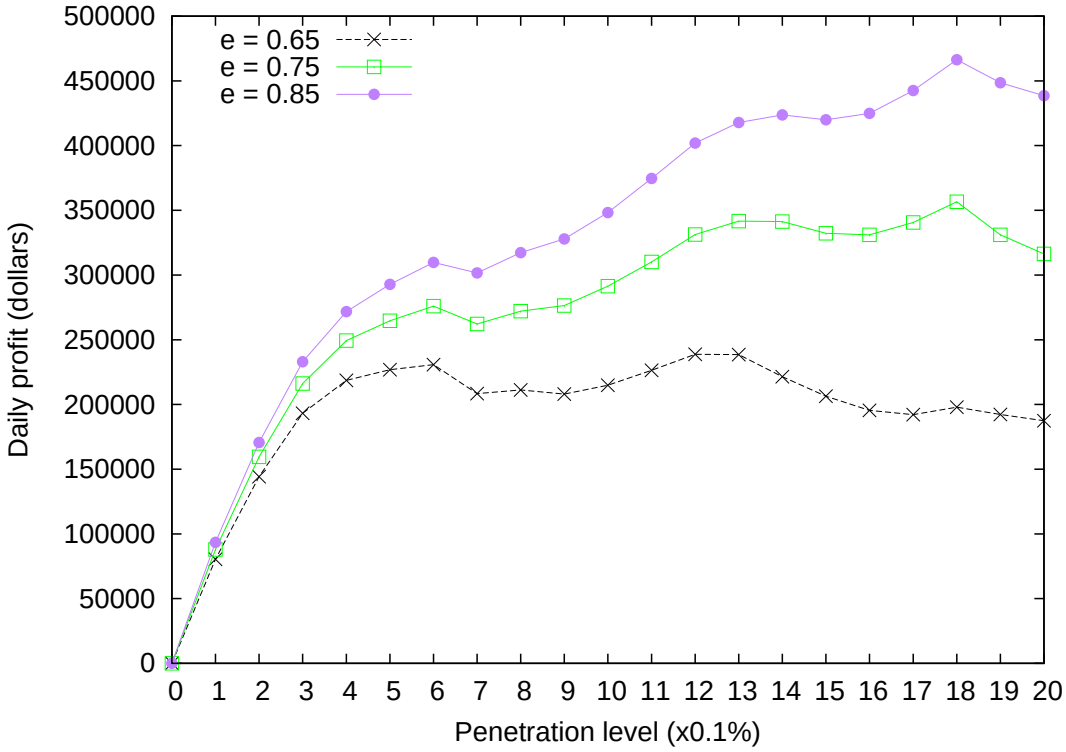


Figure 4.6: Profit of arbitrage for different penetration and efficiency

demand response programs. Experiment has shown that significant savings could be realized even from a small market presence of those demand types, if this mechanism were put to use. Some useful insight on storage expansion has also been drawn from the experiments. We have also abstracted the design philosophy in a general mathematical form, which serves as a blueprint for further extension and implementation.

5 STOCHASTIC UNIT COMMITMENT WITH DERAND SAMPLING METHOD

5.1 Introduction

The day-ahead unit commitment (DUC) schedule is computed on the basis of the day-ahead market bids which cannot be relied upon to infer the actual load profile during the operating day. It is likely that the day-ahead UC schedule is insufficient or uneconomical to support the real-time dispatch of generation resources to meet the real-time load. Reserve adequacy assessment (RAA) aims to identify the extra unit-hours for commitment[1] in preparation for the actual load and reserve requirements in the operating day.

The terminology residual unit commitment (RUC) is often used instead of RAA in industry. The problem essentially embodies a unit commitment model similar in structure to the day-ahead UC model. The differences between the RAA UC and day-ahead UC include:

- While the day-ahead UC supports (non-physical) financial transactions, RAA plans for the anticipated (forecast) physical load. In the input data to RAA, all demand bids, virtual bids and bilateral transactions are removed. Prior to the RAA run, uncommitted resources that are backed by physical injection/withdrawal capabilities are allowed to revise or resubmit their bids and the zonal load values are to be instantiated with the load forecast information.

- While the day-ahead UC model includes the DC power flow constraints, the RAA UC model substitutes the proxy constraints (i.e., transmission proxy and SFT proxy) from the day-ahead security constrained economic dispatch (SCED) process for the DC power flow constraints.

[1] Unit-hours already committed in the day-ahead UC solution are locked in and will not be decommitted in RAA.

- In the RAA model, the day-ahead UC decision is respected to the extent that the unit-hours committed by the day-ahead model will be held fixed in the committed state. RAA only assesses the "residual" unit-hours for commitment, in particular for slow-start units which are not committed by the DUC run. Therefore, RAA is smaller and easier than the full-size day-ahead UC problem.

Load forecast software provides very accurate predictions. For example, the mean relative error in 2011 is less than 1.3% and its 95-th percentile is about 3.4%. The existing deterministic model using these values is generally effective to serve the RAA goal.

A stochastic unit commitment model aims to utilize richer information from the load forecast to provide a better solution than its deterministic counterpart, see, e.g., Yu et al. (2013), Constantinescu et al. (2009) and Wang et al. (2013). It is well-known that a stochastic programming model harnessing the true probability distribution of the random data is in theory superior in solution quality compared to its deterministic counterpart[2]. However, including the probabilistic information significantly increases the size of the model and the solution difficulty. For instance, the size of the extensive form of a stochastic model, which is a mixed integer program (MIP), grows linearly as the number of scenarios increases. For the RAA problem at hand, each extra scenario adds about 190,000 variables and 166,000 constraints to the model. Table 5.1 lists the model (instance) sizes resulted from increasing number of scenarios[3]. In the worst case, the computational complexity of a MIP may grow quickly as the number of scenarios increases and exponentially as the number of discrete variables increases.

The tradeoff between the solution quality and computational efficiency is present in any practical stochastic programming design, see, for example Feng and Ryan (2014) and Papavasiliou and Oren (2013). In this chapter, we develop and showcase

[2]The solution quality is measured by evaluating the mean value of the objective function over a large number of independent samples drawn from the distribution.

[3]The numbers in the table come from the instance of June 3, 2011, and may differ by instance, due to varying number of active resource bids and proxy constraints.

Table 5.1: Stochastic RAA UC Instance Size v.s. Number of Scenarios

Scen.	Row	Col.	Non-zero	Disc. Col.	Model
1	217,390	220,709	819,309	15,073	168 MB
2	384,012	410,717	1,470,303	15,073	258 MB
3	550,634	600,725	2,121,297	15,073	347 MB
10	1,716,988	1,930,781	6,678,255	15,073	977 MB
20	3,383,208	3,830,861	13,188,195	15,073	1876 MB
30	5,049,428	5,730,941	19,698,135	15,073	2775 MB

a derandomization (Derand) sampling method that caters for a tight computational budget, i.e., a situation where only a few (3 or 5) scenarios can be computationally afforded. Such a stringent condition is rarely studied in the stochastic programming literature, but the reality in ISO New England's RAA project does necessitate it. The idea of Derand is to partition the sample space and use the conditional expectation as an informed and unbiased guess (sample) of the underlying probability distribution. This method is most applicable and advantageous for low-dimensional space with a small sample size budget. In the stochastic UC model for RAA, the load forecast error is a random parameter and only a few samples of it can be afforded in the computation, which constitutes an ideal application of the method. We acknowledge that Derand shares the same idea as the Gaussian quadrature technique in numerical integration (Brandimarte, 2014) and vector quantization in signal processing (Lloyd, 1982; Gersho and Gray, 1992), whereas the latter has been applied to scenario generation in stochastic models, see Goodwin et al. (2009) and Cooper et al. (2012) for example. Compared to the existing work, our development is based on probability theory and is uniquely tied to the general form of stochastic programming. Furthermore, we prove its unbiasedness property and emphasize its systematic (non-random, hence called derandomization) approach, which warrants a greater degree of stability and repeatability particularly desirable in an ISO's operation.

5.2 Problem Formulation

The problem is formulated as a two-stage stochastic program. The first stage decision involves the binary unit commitment variables and the second stage includes continuous variables for net energy injection, power flow, reserve level and constraint violation, etc. The objective is minimizing the total cost which is the sum of unit commitment cost, expected (over all sampled scenarios) energy and reserve costs and expected penalty cost for constraint violations. While being a mathematically valid solution status, infeasibility is inconvenient for solution quality assessment and problem diagnosis and thus should be circumvented (Liu et al., 2014). In order to make each candidate UC solution feasible for all scenarios (i.e., to have a relatively complete recourse), artificial variables are introduced to the second stage to allow for constraint violations. These variables are heavily penalized in the objective function and the penalty cost measures the extent of infeasibility. For detailed formulation of general stochastic UC problem, we refer readers to Section 10 of Chapter 1.

5.3 Derandomization Sampling Method

We base our discussion on a stochastic program of the form

$$\min_{x \in X} \{f(x) := \mathbb{E}[F(x, \xi)]\} \tag{5.1}$$

where X is a nonempty compact subset of $\mathbf{R}^n$, ξ is a random vector defined on the probability space $(\Omega, \mathcal{F}, P)$, and F is a real-valued function of x and ξ. The expectation operator is taken with respect to the probability distribution of ξ. We denote by $\Xi \subset \mathbf{R}^d$ the support of the probability distribution of ξ.

Given a natural number N, let the subsets $B_i \subset \mathbf{R}^d, i = 1, \ldots, N$ be a disjoint partition of Ξ, i.e.,

$$B_i \cap B_j = \emptyset \text{ for } i \neq j \text{ and } \cup_{i=1}^{N} B_i = \Xi$$

In addition we assume that

$$P(\{\omega : \xi(\omega) \in B_i\}) > 0 \text{ for } i = 1, \ldots, N \tag{5.2}$$

Let $A_i = \{\omega : \xi(\omega) \in B_i\}$ for $i = 1, \ldots, N$, and thus $A_1, \ldots, A_N$ form a disjoint partition of Ω. Let $\mathcal{A}_N$ be the σ-field generated by the collection $\{A_1, \ldots, A_N\}$ of subsets of Ω, i.e., $\mathcal{A}_N = \sigma(\{A_1, \ldots, A_N\})$. Clearly, we have $\mathcal{A}_N \subset \mathcal{F}$. Now define a random variable ξ_N by

$$\xi_N = \sum_{i=1}^{N} \frac{\mathbb{E}[\xi 1_{A_i}]}{P(A_i)} 1_{A_i}, \text{ where } 1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

Note that ξ_N is a discrete random variable which takes the value $\mathbb{E}[\xi 1_{\{\xi \in B_i\}}]/P(\xi \in B_i)$ with probability $P(\xi \in B_i)$, for $i = 1, \ldots, N$. In fact, ξ_N is an unbiased estimator of ξ, as shown in the following lemma.

Lemma 5.1. *ξ_N is the conditional expectation of ξ given $\mathcal{A}_N$, i.e.,*

$$\xi_N = \mathbb{E}[\xi | \mathcal{A}_N]$$

Proof. From the definition of $\mathcal{A}_N$ and the fact that $A_1, \ldots, A_N$ partition Ω, we have that every element A of $\mathcal{A}_N$ is of the form

$$A = \cup_{i \in I} A_i, \ I \subset \mathbf{N} \tag{5.3}$$

For A given by (5.3),

$$\mathbb{E}[\xi 1_A] = \mathbb{E}[\xi \sum_{i \in I} 1_{A_i}] = \sum_{i \in I} \mathbb{E}[\xi 1_{A_i}]$$

On the other hand, we have

$$\mathbb{E}[\xi_N 1_A] = \mathbb{E}[\sum_{i \in I} \frac{\mathbb{E}[\xi 1_{A_i}]}{P(A_i)} 1_{A_i}] = \sum_{i \in I} (\frac{\mathbb{E}[\xi 1_{A_i}]}{P(A_i)} \mathbb{E}[1_{A_i}]) = \sum_{i \in I} \mathbb{E}[\xi 1_{A_i}]$$

We have shown that $\mathbb{E}[\xi_N 1_A] = \mathbb{E}[\xi 1_A]$ for all $A \in \mathcal{A}_N$, therefore proven the lemma. □

Let $F_N(x, \xi) := F(x, \xi_N)$ and by noting that $F_N(x, \xi)$ can be equivalently expressed as

$$F_N(x, \xi) = \sum_{i=1}^{N} F(x, \frac{\mathbb{E}[\xi 1_{A_i}]}{P(A_i)}) 1_{A_i},$$

the following problem

$$\min_{x \in X} \{f_N(x) := \mathbb{E}[F_N(x, \xi)] = \sum_{i=1}^{N} P(A_i) F(x, \frac{\mathbb{E}[\xi 1_{A_i}]}{P(A_i)})\} \tag{5.4}$$

is well-defined and solvable.

For a fixed N, $\mathcal{A}_N$ encapsulates the information that we could exploit to tie down the randomness of ξ in the optimization problem. The increase of N enriches the information encapsulated in $\mathcal{A}_N$. In particular, if the sequence of σ-fields $\mathcal{A}_N$ forms a filtration, i.e., $\mathcal{A}_N \subset \mathcal{A}_{N+1}$ for each N, then the process $(\xi_N, N = 1, 2, \ldots)$ is a martingale with respect to the filtration, because we have

$$\mathbb{E}[\xi_{N+1} | \mathcal{A}_N] = \mathbb{E}[\mathbb{E}[\xi | \mathcal{A}_{N+1}] | \mathcal{A}_N] = \mathbb{E}[\xi | \mathcal{A}_N] = \xi_N$$

where the second equality comes from the tower property of conditional expectation. This to some extent indicates that N represents the richness of the information we could possibly have on hand while estimating ξ.

If Ξ is a finite set, the maximum number of nonempty subsets needed to partition Ξ is the size of Ξ, denoted by $|\Xi|$, in which case each subset becomes a singleton representing a possible value of ξ. The resulting $\mathcal{A}_{|\Xi|}$ is the last member in the filtration, which is clearly equal to $\sigma(\xi)$, the σ-field generated by ξ. Therefore, we have

$$\xi_{|\Xi|} = \mathbb{E}[\xi | \mathcal{A}_{|\Xi|}] = \mathbb{E}[\xi | \sigma(\xi)] = \xi$$

and consequently,

$$\min_{x\in X} f_{|\Xi|}(x) \text{ is equivalent to (5.1).} \tag{5.5}$$

To discuss the case where Ξ is infinite, we first note that it is always possible, via progressive partitioning of Ξ, to obtain an increasing sequence of σ-fields $\mathcal{A}_N$ that converges to $\sigma(\xi)$, i.e., $\mathcal{A}_N \subset \mathcal{A}_{N+1}$ for any N, and $\mathcal{A}_\infty := \sigma(\cup_{N\in\mathbf{N}}\mathcal{A}_N) = \sigma(\xi)$. Then we have the following proposition.

Proposition 1. *Suppose $\{\mathcal{A}_N\}$ is an increasing sequence of σ-fields and $\mathcal{A}_\infty = \sigma(\xi)$. As $N \to \infty$,*

$$\xi_N \to \xi \text{ almost surely.} \tag{5.6}$$

Proof. Since we have $\mathbb{E}[\xi|\sigma(\xi)] = \xi$, and $\xi_N = \mathbb{E}[\xi|\mathcal{A}_N]$ by Lemma 5.1, Theorem 5.5.7 in Durrett (2010) completes the proof. □

Results in (5.5) and Proposition 1 indicate that ξ_N is a consistent estimator of ξ. We can regard (5.4) as a knowledge-guided approximation to the original problem (5.1) in the sense that, given a computing budget N and the associated σ-field $\mathcal{A}_N$, the solution of (5.4) is based on an informed and unbiased guess of the underlying random parameter. Note that for the same N, different partitioning schemes will result in different $\mathcal{A}_N$'s, which leaves plenty of freedom for algorithm design.

Theorem 5.2. *Let X be a nonempty compact subset of $\boldsymbol{R}^n$ and suppose, in addition to a given $\{\mathcal{A}_N\}$ satisfying the assumptions in Proposition 1, that for any $x \in X$,* (i) *$F(x,\cdot)$ is a bounded continuous function, and* (ii) *$F(\cdot,\xi)$ is continuous at x almost surely. Then $f(x)$ is finite valued and continuous on X, and $f_N(x)$ converges to $f(x)$ uniformly on X.*

Proof. Proposition 1 implies that ξ_N converges to ξ in distribution, which then implies that

$$\mathbb{E}[g(\xi_N)] \to \mathbb{E}[g(\xi)]$$

for all bounded continuous function g, so by the assumption (i) we have $f_N(x) \to f(x)$ for each $x \in X$. It also follows from (i) that for all $x \in X$ there is a number $M(x)$ with $|F(x,\xi)| \leqslant M(x) < \infty$, and consequently $|f(x)| = |\mathbb{E}[F(x,\xi)]| \leqslant |\mathbb{E}[M(x)]| =$

$|M(x)| < \infty$. Consider a point $x \in X$ and let x_k be a sequence of points in X converging to x. By (i) coupled with the Bounded Convergence Theorem, we have

$$\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} \mathbb{E}[F(x_k, \xi)] = \mathbb{E}[\lim_{k\to\infty} F(x_k, \xi)].$$

By (ii), we have $\lim_{k\to\infty} F(x_k, \xi) = F(x, \xi)$ almost surely, then it follows that $f(x_k) \to f(x)$, hence $f(x)$ is continuous.

Now choose a point $\bar{x} \in X$, a sequence γ_k of positive numbers converging to zero, and define $V_k := \{x \in X : \|x - \bar{x}\| \leqslant \gamma_k\}$ and

$$\delta_k(\xi) := \sup_{x\in V_k} |F(x, \xi) - F(\bar{x}, \xi)|.$$

By (ii) we have $\delta_k(\xi) \to 0$ almost surely as $k \to \infty$ and by (i) we have that $\delta_k(\xi), k \in \mathbf{N}$ are bounded, hence by the Bounded Convergence Theorem we have

$$\lim_{k\to\infty} \mathbb{E}[\delta_k(\xi)] = \mathbb{E}[\lim_{k\to\infty} \delta_k(\xi)] = 0. \tag{5.7}$$

In the meantime, for any $x \in V_k$ we have

$$\begin{aligned} |f_N(x) - f_N(\bar{x})| &= |\mathbb{E}[F(x, \xi_N)] - \mathbb{E}[F(\bar{x}, \xi_N)]| \\ &= |\mathbb{E}[F(x, \xi_N) - F(\bar{x}, \xi_N)]| \\ &\leqslant \mathbb{E}|F(x, \xi_N) - F(\bar{x}, \xi_N)| \end{aligned}$$

where the second line is by the linearity of the expectation operator and the third line is by Jensen's Inequality. Consequently,

$$\begin{aligned} \sup_{x\in V_k} |f_N(x) - f_N(\bar{x})| &\leqslant \sup_{x\in V_k} \mathbb{E}|F(x, \xi_N) - F(\bar{x}, \xi_N)| \\ &\leqslant \mathbb{E}[\sup_{x\in V_k} |F(x, \xi_N) - F(\bar{x}, \xi_N)|] \\ &= \mathbb{E}[\delta_k(\xi_N)] \end{aligned}$$

Because $F(x, \cdot)$ is continuous and V_k is compact, $\delta_k(\cdot), k \in \mathbf{N}$ are continuous, and

since ξ_N converges to ξ almost surely, we then have $\mathbb{E}[\delta_k(\xi_N)] \to \mathbb{E}[\delta_k(\xi)]$ as $N \to \infty$. Together with (5.7) this implies that for any $\epsilon > 0$ there exists a neighborhood W of $\bar{x}$ such that for sufficient large N,

$$\sup_{x \in W \cap X} |f_N(x) - f_N(\bar{x})| < \epsilon.$$

Since X is compact, there exists a finite number of points $x_1, \ldots, x_m \in X$ and corresponding neighborhoods $W_1, \ldots, W_m$ covering X such that for N large enough the following holds

$$\sup_{x \in W_j \cap X} |f_N(x) - f_N(x_j)| < \epsilon, j = 1, \ldots, m. \tag{5.8}$$

Furthermore, since $f(x)$ is continuous on X, these neighborhoods can be chosen in such a way that

$$\sup_{x \in W_j \cap X} |f(x) - f(x_j)| < \epsilon, j = 1, \ldots, m. \tag{5.9}$$

Since we have shown $f_N(x) \to f(x)$ for each $x \in X$, this means that

$$|f_N(x_j) - f(x_j)| < \epsilon, j = 1, \ldots, m \tag{5.10}$$

holds for N large enough. It follows from (5.8)-(5.10) that for N large enough

$$\sup_{x \in X} |f_N(x) - f(x)| < 3\epsilon. \tag{5.11}$$

Since $\epsilon > 0$ was arbitrary, (5.11) indicates that $f_N(x)$ converges to $f(x)$ uniformly on X and hence the proof is complete. □

As $\mathcal{A}_N$ is constructed by dividing Ξ into N parts and given such a division there is complete freedom as for where to place the next cut point to form $\mathcal{A}_{N+1}$, we can see that there are numerous ways to form the filtration $\{\mathcal{A}_N\}$ that satisfies the assumptions in Proposition 1. Suppose $\{\mathcal{A}_N\}$ is formed randomly so that the formation is modeled by a measurable map from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$, where S is the set of all sequences $\{\mathcal{A}_N\}$ satisfying the assumptions in Proposition 1 and $\mathcal{S}$ denotes

the set of all subsets of S, and since the above theorem works for each and every value of this map, i.e., "a given $\{\mathcal{A}_N\}$" as stated in the theorem, then it is possible to establish that $f_N(x)$ is a random variable which converges to $f(x)$ uniformly on X and the convergence is in the almost sure sense.

Theorem 5.2 resembles the Theorem 7.48 in Shapiro et al. (2009) in many aspects, from the assumptions to the conclusions. In fact, a significant portion of the proofs overlap. Since the establishment of the consistency properties of the SAA estimators is primarily based on the uniform convergence of the sample average function $\hat{f}_N(x)$, see Section 5.1.1 of Shapiro et al. (2009), we claim that those properties also hold for the Derand method on similar minor assumptions.

We defer more detailed theoretical discussion to future work. In what follows, let us focus on analyzing the practical problem and making a concrete case for Derand method's industrial application.

Forecast Error Analysis

Let us examine the forecast errors based on the historical data of 2011. There are 8759 entries in the data set[4], each entry i consisting of the system-wide forecast d_i^f and the actual load d_i^a for the hour, $i = 1, 2, \ldots, 8759$. We calculate the relative error δ_i by the formula $\delta_i = (d_i^f - d_i^a)/d_i^f$. Our plan is to analyze the error distribution and make informed guesses of the error in future forecasts. When a new forecast d^f arrives and a stochastic UC with n load scenarios needs to be solved, we can draw n error samples, $\hat{\delta}_1, \ldots, \hat{\delta}_n$, and construct the load scenarios $\hat{d}_1, \ldots, \hat{d}_n$ using the formula

$$\hat{d}_s = d^f(1 - \hat{\delta}_s), \; s = 1, \ldots, n \tag{5.12}$$

The error distribution is visualized in Figure 5.1. In the histogram, the relative error is densely concentrated around zero, suggesting that the forecast is very accurate. The errors also exhibit a notable bias toward the positive quadrant (mean=0.002599), indicating a trace of over-forecasting.

[4]In the data set, the hour 02:00 of March 13, 2011 is missing due to daylight saving, but the missing hour is not added back on November 6; otherwise, there would be 8760 entries for the year.

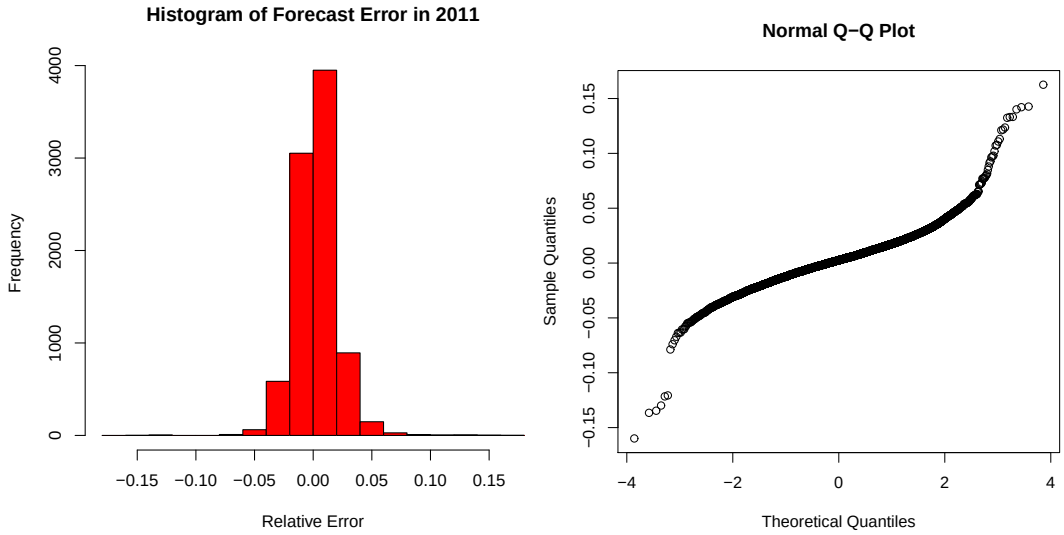


Figure 5.1: Relative forecast error distribution in 2011.

The Normal Q-Q plot suggests that Normal distribution is a poor fit for the error, which is also indicated by the small p-value (<0.005) in a goodness of fit test. We have tried a number of off-the-shelf distributions and none of them could provide a good fit that passes the Chi-square test[5].

In summary, *we need a few high-quality samples from an empirical distribution.*

Scenario Generation

We adopt the Derand method to generate samples for the error distribution. The empirical distribution, formed by a finite number of historical observations, is the best available approximation of the true distribution of δ. We will use those historical samples as building blocks to form a smaller set of samples to approximate the error distribution.

Suppose the error distribution has a support of the real line. Given a target sample size N, we partition the support into N pieces, i.e., $(-\infty, c_1] \cup (c_1, c_2] \cup \ldots \cup (c_{N-1}, \infty)$. The $N-1$ cut points can be obtained by inversely transforming

[5]The fit is performed by Arena® Input Analyzer.

Uniform(0,1) samples by the empirical CDF of δ. For simplicity, we use a direct and deterministic approach to set the cut points. Observing that δ rarely falls outside the interval $[-0.1, 0.1]$, we use the $N-1$ points that uniformly partition the interval $[-0.1, 0.1]$ as the cut points. For instance, when $N = 3$, we have $v_1 = -0.0333$ and $v_2 = 0.0333$, and the partitioning outcome consists of three intervals, $[-\infty, -0.0333]$, $(-0.0333, 0.0333]$ and $(0.0333, \infty)$.

For each interval (partition) k, we then generate a sample point (v_k, p_k), where v_k is the value and p_k is its probability. In the Derand method, v_k is the expectation of δ conditioning on the interval k and p_k is the probability of interval k, i.e., $H(c_k) - H(c_{k-1})$, where $H(\cdot)$ is the CDF of δ.

This is straightforward in the present case since we work directly with the historical data. In particular, v_k is the average value of all observations that fall in the interval k, and p_k is the relative frequency of observations falling in this interval.

Once the load scenarios have been determined, the system load is apportioned to the eight zones as zonal loads, according to a certain percentage mix, i.e., a weight matrix $w_{t,z}$ with $\sum_z w_{t,z} = 1, \forall t$. The weight matrix is given in Figure 5.2, in which the entries were calculated by averaging the historical data *2011_smd_hourly.xls* accessible at *www.iso-ne.com*

5.4 Performance Evaluation

Experiment Design

We have the full market and network data for a single day, i.e., June 3, 2011, as well as hourly system load and forecast data for the year 2011 and 2012. We will use these data in the experiments.

We compare the proposed Derand method against two alternative approaches: (1) Sample average approximation (SAA) with Monte Carlo sampling (Kleywegt and Shapiro, 2001), and (2) the Scenario reduction algorithm, SCENRED2, within GAMS (Heitsch and Römisch, 2003; GAMS). We fix the sample (scenario) size to

	ME	NH	VT	CT	RI	SEMASS	WCMASS	NEMASSBOST
Hour1	0.0884	0.0884	0.0465	0.2449	0.0649	0.1172	0.1412	0.2086
Hour2	0.0897	0.0884	0.0470	0.2436	0.0647	0.1164	0.1414	0.2088
Hour3	0.0906	0.0885	0.0472	0.2428	0.0644	0.1160	0.1416	0.2090
Hour4	0.0912	0.0888	0.0474	0.2422	0.0642	0.1158	0.1413	0.2090
Hour5	0.0922	0.0896	0.0475	0.2418	0.0640	0.1157	0.1412	0.2081
Hour6	0.0931	0.0908	0.0477	0.2418	0.0634	0.1160	0.1410	0.2063
Hour7	0.0931	0.0919	0.0481	0.2426	0.0629	0.1164	0.1404	0.2045
Hour8	0.0922	0.0922	0.0480	0.2436	0.0633	0.1170	0.1397	0.2040
Hour9	0.0913	0.0923	0.0473	0.2446	0.0637	0.1177	0.1392	0.2038
Hour10	0.0904	0.0923	0.0468	0.2451	0.0643	0.1184	0.1389	0.2038
Hour11	0.0894	0.0923	0.0462	0.2458	0.0649	0.1186	0.1390	0.2038
Hour12	0.0884	0.0921	0.0460	0.2460	0.0652	0.1190	0.1391	0.2042
Hour13	0.0878	0.0917	0.0456	0.2462	0.0654	0.1191	0.1391	0.2051
Hour14	0.0873	0.0916	0.0455	0.2467	0.0655	0.1192	0.1392	0.2051
Hour15	0.0870	0.0913	0.0453	0.2471	0.0655	0.1191	0.1391	0.2056
Hour16	0.0867	0.0914	0.0453	0.2471	0.0653	0.1196	0.1389	0.2058
Hour17	0.0873	0.0916	0.0451	0.2469	0.0652	0.1204	0.1387	0.2047
Hour18	0.0876	0.0921	0.0451	0.2473	0.0651	0.1216	0.1385	0.2027
Hour19	0.0876	0.0925	0.0453	0.2475	0.0652	0.1226	0.1384	0.2008
Hour20	0.0876	0.0924	0.0452	0.2473	0.0654	0.1231	0.1384	0.2006
Hour21	0.0872	0.0918	0.0451	0.2477	0.0655	0.1231	0.1387	0.2009
Hour22	0.0862	0.0905	0.0455	0.2481	0.0657	0.1221	0.1391	0.2027
Hour23	0.0861	0.0893	0.0453	0.2481	0.0656	0.1203	0.1399	0.2055
Hour24	0.0869	0.0886	0.0459	0.2468	0.0653	0.1186	0.1405	0.2074

Figure 5.2: 2011 average zonal share of the system load.

3 and experiment different sampling methods. For a 3-scenario stochastic RAA UC problem, it may take CPLEX, with all options at default values, more than two hours to find the global solution (zero optimality gap). For each run, we set the computation time limit to one hour (reslim = 3600). It is observed that within one hour, the 3-scenario stochastic model is able to find the global solution in most cases, see column 7 in Table 5.2 and 5.3.

All scenario generation methods (Derand, SAA and SCENRED) use the knowledge gained from the historical data of 2011 and their performances are tested on the data of 2012. The performance test of an RAA solution against the 365 days in 2012 takes about 5 minutes.

The SAA method is implemented as follows: for each hour, draw three Uniform(0,1) random numbers (e.g., 0.618, 0.824 and 0.264), and then find the corresponding quantiles through the empirical CDF of the historical error distribution (e.g., 0.007, 0.016 and -0.006). These quantiles serve as the error scenarios of the forecast for the given hour, each with an equal probability of 0.333. The load sce-

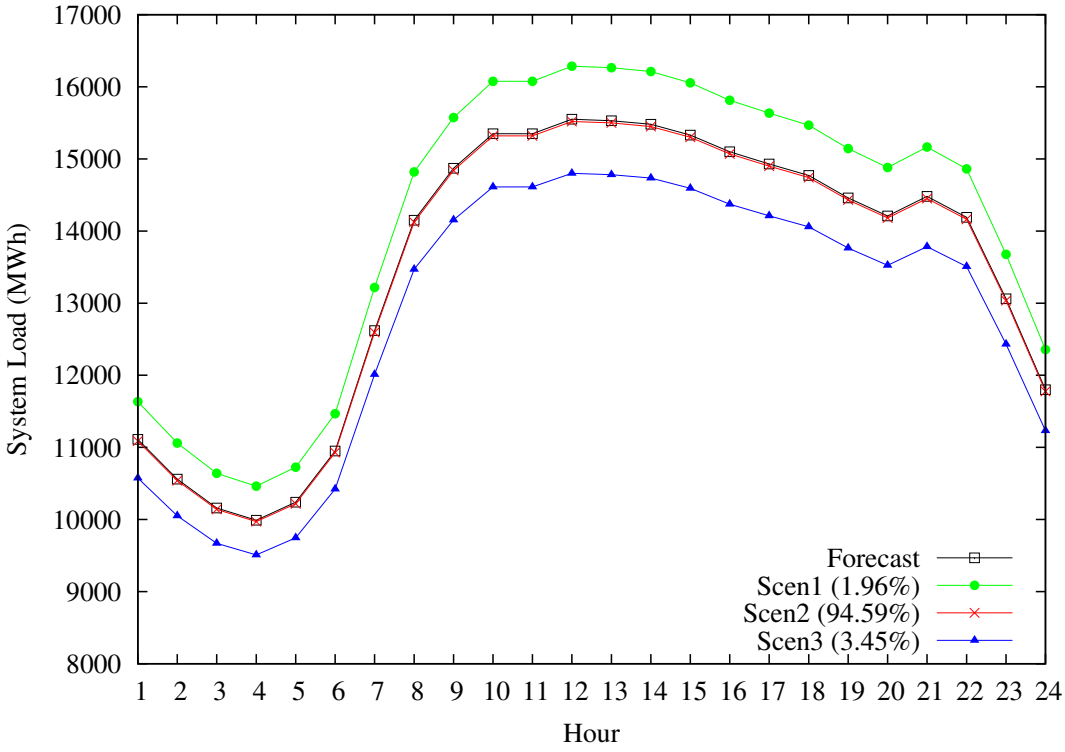


Figure 5.3: Load scenarios generated by Derand.

narios are then constructed by formula (5.12), where d^f is the load forecast of the hour in the operating day June 3, 2011.

The Derand method assigns the same error and probability to all hours in a scenario, so the load profiles of the three scenarios are "parallel" to each other. The Derand scenarios and the SAA scenarios are plotted in Figure 5.3 and 5.4.

Given an initial probability distribution P with a finite support Ξ (i.e., an initial population), SCENRED finds a "reduced" probability distribution that is supported by a subset of Ξ of prescribed cardinality (in this case, 3) and is close to P in terms of a Fortet-Mourier probability metric (Heitsch and Römisch, 2003). In the experiments two parameters, cost and forecast error, are used each to form an initial population of 364 scenarios for SCENRED. The cost for each scenario is pre-computed as the optimal objective value of a deterministic RAA UC problem taking as input the actual forecast errors for the corresponding day in 2011. SCENRED selects 3 out of

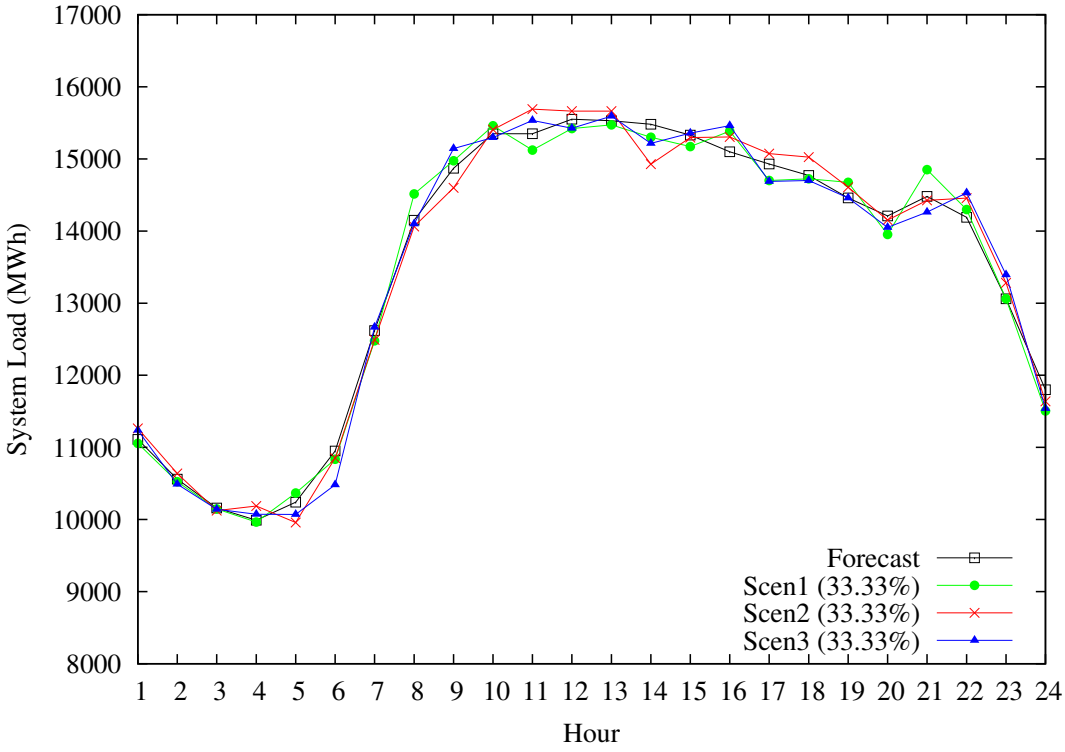


Figure 5.4: Load scenarios generated by SAA.

the 364 scenarios and assigns a probability to each selected scenario. The scenarios selected by-cost and by-error are plotted in Figure 5.5 and 5.6, respectively.

Result Analysis

The main results arising from an evaluation over 365 simulation cases from 2012 are listed in Table 5.2. In the table, the first column indicates the method. In order, they are the deterministic (1-scenario) model run on the load forecast, Derand method, two scenario reduction (SR) methods and five trials of the SAA method. Columns 2 to 5 report the expected (daily average) costs: Total = Commit + (Energy + Reserve) + Penalty. Column 6 shows the number of cases where penalty costs are incurred (some constraint is violated) and Column 7 notes the execution time in minutes (or percentage gap, for a one hour time limit).

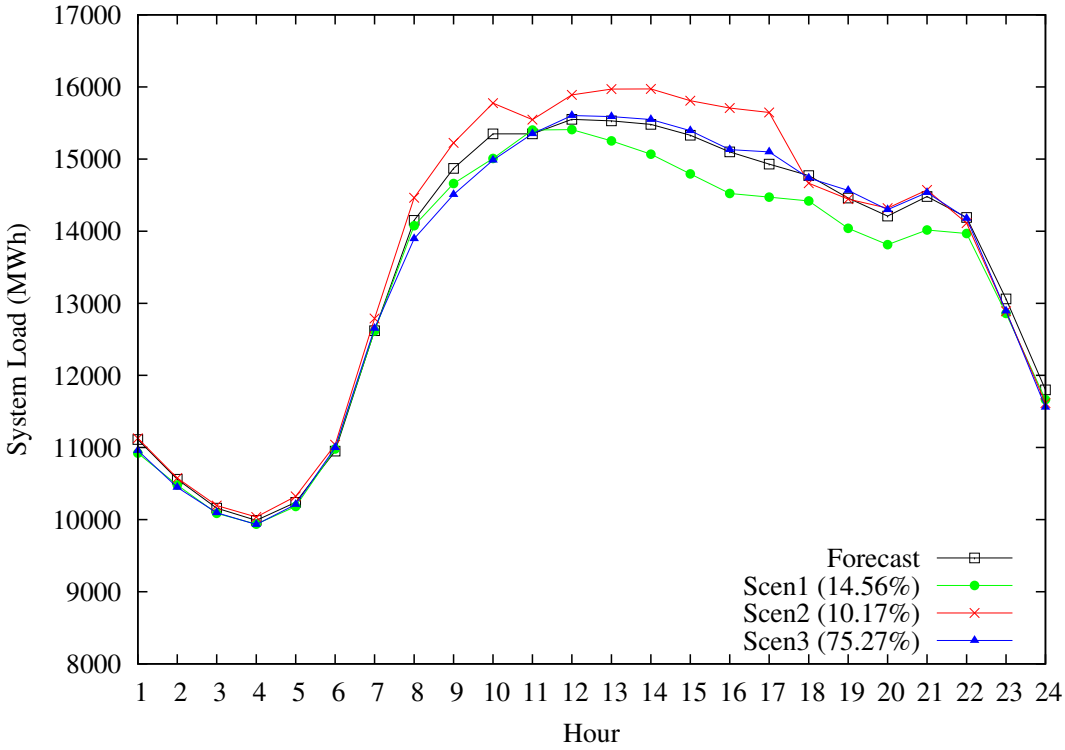


Figure 5.5: Load scenarios of SCENRED by cost.

Table 5.2: Stochastic UC Performance with 3 Scenarios

Model	Total	Commit	En + Re	Penalty	Viol.	Time
Dtmnstc	23,009,667	1,335,604	18,061,648	3,612,415	215	2.5′
Derand	19,567,526	1,325,029	18,081,097	161,399	11	0.013%
SR (cost)	20,404,708	1,286,661	18,104,208	1,013,839	143	8.33′
SR (error)	26,090,031	1,328,095	18,098,266	6,663,670	298	5.91′
SAA trial 1	21,332,370	1,350,344	18,032,758	1,949,268	183	17.33′
SAA trial 2	22,290,783	1,344,929	18,055,468	2,890,386	250	20.28′
SAA trial 3	21,498,931	1,349,540	18,052,558	2,096,833	187	24.10′
SAA trial 4	23,284,301	1,336,093	18,068,259	3,879,950	252	4.19′
SAA trial 5	21,712,592	1,350,305	18,046,846	2,315,441	220	12.94′

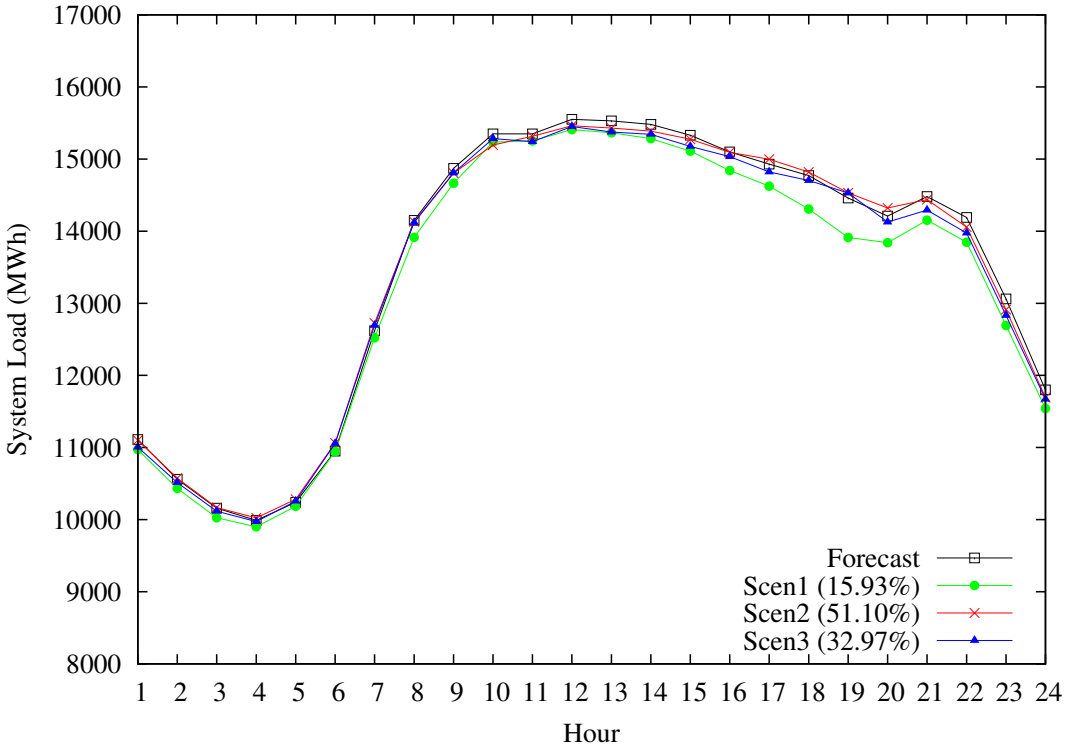


Figure 5.6: Load scenarios of SCENRED by error.

Table 5.3: Stochastic UC Performance with 5 Scenarios

Model	Total	Commit	En + Re	Penalty	Viol.	Time
Dtmnstc	23,009,667	1,335,604	18,061,648	3,612,415	215	2.5′
Derand	19,811,759	1,466,692	18,128,757	216,310	1	6.80′
SR (cost)	19,750,688	1,399,637	18,102,261	248,790	28	26.48′
SR (error)	21,074,742	1,343,998	18,039,367	1,691,377	132	40.69′
SAA trial 1	21,432,672	1,358,369	18,044,793	2,029,510	225	0.004%
SAA trial 2	21,680,361	1,335,077	18,065,749	2,279,536	196	33.85′
SAA trial 3	20,807,084	1,358,466	18,033,642	1,414,977	143	51.48′
SAA trial 4	21,550,597	1,351,373	18,041,205	2,158,019	204	0.004%
SAA trial 5	20,603,777	1,352,102	18,037,938	1,213,737	127	45.46′

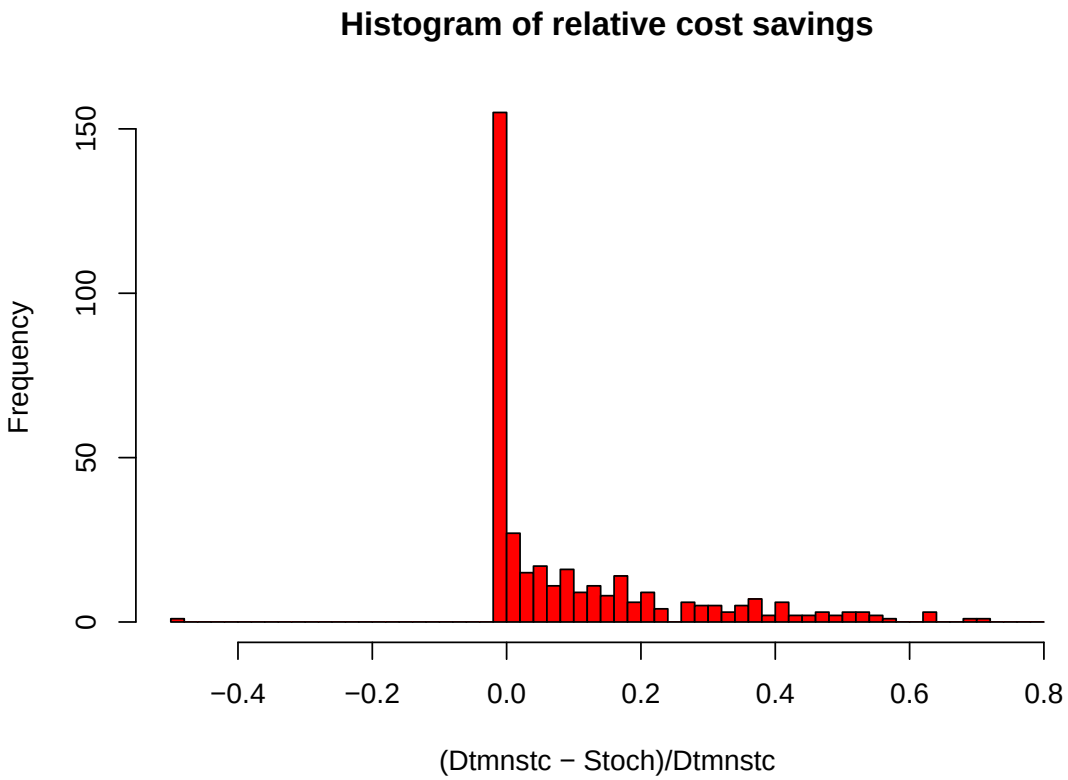


Figure 5.7: Cost savings of 3-scenario stochastic model with Derand sampling.

It is apparent that in expectation the stochastic model equipped with the Derand sampling significantly outperforms the deterministic model. For detail, Figure 5.7 shows the histogram of the relative differences in total costs between the stochastic solution and the deterministic solution over the 365 simulation cases. In about half of the cases, the stochastic solution results in a slightly ($\leqslant 2\%$) higher cost than its deterministic counterpart and in almost all remaining cases, the stochastic solution yields significant savings, up to 70%.

Compared to the other scenario generation methods tested here, Derand yields the best results, with the lowest expected cost and the fewest constraint violations. On the other hand, the model populated by Derand samples also appears harder to solve – there remains a 0.013% optimality gap after one-hour solution time.

The stochastic models populated by SR-by-error samples gave worse results than

the deterministic model and the model with SAA samples did not perform well. This suggests that the quality of samples is crucial to the performance under such a stringent sample size budget and that the performance gain over the deterministic model observed here is largely due to the superior samples provided by Derand (and SR-by-cost), rather than the theoretical advantage of stochastic programming, which relies on the law of large numbers.

Experiment results on 5-scenario programs are listed in Table 5.3. All the above conclusions remain valid.

5.5 Conclusion

RAA determines the supplemental unit commitments in addition to the Day-ahead unit commitment schedule, according to which generating asset operators arrange the fuel delivery and other preparation work for the next day. It is important from the reliability standpoint that the unit commitment schedule is robust to the load uncertainty for the operating day, as re-scheduling units in a short notice in the case of real-time resource inadequacy can be costly. The potential economic and reliability benefits of a better RAA solution warrant this investigation of stochastic unit commitment.

In face of uncertainty, a stochastic program is guaranteed to provide a better solution in the long run, provided that the uncertainty pattern or distribution is known with accuracy and the pattern will repeat in the future. In particular, the more samples involved, the better the solution becomes. However, due to the large problem size and computational constraints, a stochastic RAA model cannot take as many scenarios as we wish. Therefore, in reality only a small number of scenarios will be incorporated into the stochastic model.

In this circumstance, selecting good samples is paramount. We have tested a few sampling methods and found that the Derand method, which makes informed guesses based on partitioning and properties of conditional expectation, could substantially boost the RAA performance. Unlike other sampling methods such as Monte Carlo or Latin hypercube which draws “random” samples, Derand employs

a systematic approach to generate samples with a greater degree of stability and repeatability. Stability and repeatability are desirable features in the ISO's operation. Scenario reduction techniques (in particular SR-by-cost) also appear to be a competitive alternative, although the original large number of scenarios typically come from the solution of optimization models which take extra time and resources to solve.

6 MULTI-STAGE SECURITY-CONSTRAINED ECONOMIC DISPATCH

6.1 Introduction

Motivation

Economic efficiency and system reliability are top concerns in the day-to-day operation of the restructured energy market over the grid. However, the two goals are intrinsically competing with each other. Efficiency pushes for the maximum use of available transmission capacity to facilitate the merit-order dispatch of generation resources, while reliability requires a certain degree of conservatism in the use of transmission capacity to prepare for unexpected events such as line and generator outages. To balance the two goals, system operators typically solve a security-constrained economic dispatch (SCED) model. Economic dispatch (ED) seeks a nodal injection/withdrawal arrangement (i.e., dispatch solution) to minimize the total generation cost in a base-case network setting. Security constraints (SC) require that the economic dispatch solution must simultaneously support a feasible power flow under a list of counterfactual scenarios of component failure, called contingencies. A *feasible* power flow is one that does not cause overloads in lines, as power flow automatically redistributes across lines following physical laws in case of a contingency. When the contingency list spans all elements in the system, the corresponding SCED solution, if one exists, is said to meet the N-1 security criterion. In practice, the solution is oftentimes obtained or approximated via an iterative process: obtain an ED solution and test if it is feasible for all contingency cases, if not, refine the solution and test again. This process is termed as a simultaneous feasibility test (SFT).

Practical reliability standards usually allow for some flexibility in the security constraints. In particular, the post-contingency power flow may temporarily exceed the normal line rating as long as system operators are able to correct it in a limited amount of time via rescheduling actions. For example, ISO New England uses four levels of thermal capacity ratings for transmission facilities: Normal, Long Time

Emergency (LTE), Short Time Emergency (STE) and Drastic Action Limit (DAL), with increasing rating numbers. While the transmission line and equipment loadings should not exceed the Normal rating for pre-contingency system conditions, the operating procedure (OP), see ISO New England, approves the use of other less restrictive ratings under contingency conditions. Specifically, the OP requires that the post-contingency line loadings should not rise beyond the DAL and must be reduced below the STE rating in 5 minutes, reduced below the LTE rating in 15 minutes and reduced below the Normal rating in 30 minutes, see Figure 6.1 for an illustration. Although imposing the Normal rating at all times is sufficient for reliability, the relaxed standards should be properly implemented in the SCED software to preserve economic efficiency. However, there is no evidence that the post-contingency rescheduling procedures are actually considered in prevalent dispatch software. We postulate that the main obstacle comes from the computational difficulty due to the increased model size.

In this chapter, we present a SCED model that takes the multi-stage contingency response actions into account. In order to solve large instances of the model, we develop a series of algorithmic enhancements based on the Benders' decomposition method. We also analyze the causes of infeasibility and propose an approach to diagnose and correct infeasible situations in the solution process. Thus our solution approach provides not only an optimal dispatch solution, but also a list of contingencies that need to be treated separately.

Related Work

SCED with corrective rescheduling (SCED-C) has been studied for over two decades. The pioneering work by Monticelli et al. (1987a) described the mathematical framework with great clarity and illustrated the economic gain of taking into account system rescheduling capabilities. The authors also pointed out many extensions that motivated our work, including multiple dispatch stages each considering different line ratings for different time frames of emergency control and the prospect of processing the subproblems in parallel. Recent advances have been made along

two major avenues: (1) contingency filtering (CF) techniques to effectively reduce the problem size, e.g., Capitanescu and Wehenkel (2008), Capitanescu et al. (2007a) and Fliscounakis et al. (2013); (2) decomposition and parallel algorithms to obtain/approximate global solutions efficiently, e.g., Phan and Kalagnanam (2014) and Lubin et al. (2011b).

Capitanescu and Wehenkel (2008) studied the (single-stage) corrective security-constrained optimal power flow (CSCOPF) problem. The authors exploited the fact that in practice most contingencies are not binding at the optimum by iteratively solving CSCOPF (using an interior-point method as described in Capitanescu et al. (2007b)) with increasing size. In each iteration, the CSCOPF model only incorporates those post-contingency constraints that have been identified to be "potentially binding" by a contingency filtering procedure. In another paper, Capitanescu et al. (2007a) proposed two CF techniques to efficiently identify a minimal "dominating" subset of contingencies, the complement set of which is redundant for the solution of SCOPF and can be removed, thus reducing the size of the problem. A recent work by Fliscounakis et al. (2013) incorporates uncertain demand in the SCED-C context and uses a mixed integer bi-level optimization model to ensure a worst-case coverage of the dispatch solution. The authors ranked contingencies into four clusters based on severity and carefully chose the solver options for computational performance. In the present work, we embed a contingency filtering idea in the Benders' algorithm. Compared to existing work in the literature, our method has three desirable features. First, the CF step does not incur extra computation load. Second, the filtering is not a once-and-for-all procedure, but is dynamically integrated in the iterative algorithm. Third, the filtering requires minimal domain-knowledge-based judgement about the network or contingency but is entirely based upon numerical results of the subproblem. Admittedly, domain knowledge might also be useful to augment our method.

To solve the nonlinear nonconvex SCOPF problem, Phan and Kalagnanam (2014) investigated a global optimization algorithm based on Lagrangian duality, as well as two decomposition schemes, namely, Benders' decomposition and the alternating direction method of multipliers (ADMM). Since a Benders' cut is not valid (i.e.,

may cut off feasible regions and the global solution) in the nonconvex AC context, as a computational alleviation the authors proposed to shift the cutting plane by an adaptively chosen distance so as to cut off less of the feasible region. The authors also briefly remarked that a contingency, once feasible for the base-case solution, can be "switched off" from future iterations. We derive a concrete and rigorous algorithmic enhancement that includes adaptive switching off of contingencies and demonstrate its effectiveness quantitatively. Pinto and Stott briefly reviewed the Benders' algorithmic framework applied to SCED-C and stressed that a computational study on a full-scale prototype was needed, which we aim to provide in this chapter.

Parallel computing is becoming a standard technique for "large-scale" computation in decomposable systems. Recent work from Argonne National Laboratory, in particular Lubin et al. (2011b) and Lubin et al. (2011a), provided efficient parallel algorithms for solving huge LPs. The algorithms were based on interior-point methods and a Schur complement technique, which the authors demonstrated to achieve a high scaling efficiency on supercomputers. In this chapter, we implemented the well-established scheme of parallel computing for Benders' decomposition, with the aim of showcasing its practical effectiveness on an affordable computing server, hence providing a realistic estimate of deployment potential of our model.

In the aspect of modeling, Capitanescu and Wehenkel (2007) warned that if the immediate post-contingency state (power flow) violates limits too much, the system may collapse before corrective actions take effect. They postulated a constraint to be added to the corrective SCED model in order to prevent this from happening. Our multi-stage corrective model naturally contains such a constraint, i.e., the one for period $T = 0$ with the DAL line rating. While part of the SCED-C literature is based on the AC power flow equations (Monticelli et al., 1987a; Capitanescu and Wehenkel, 2008), in the present work a linear "DC" model is more aligned with our objective. First, the decomposition theory for convex optimization is well-established and proven to guarantee global solutions, so we can focus on algorithmic enhancements for faster solution. Second, most ISO/RTOs use a linear model in the dispatch software, thus our algorithm as well as the computational

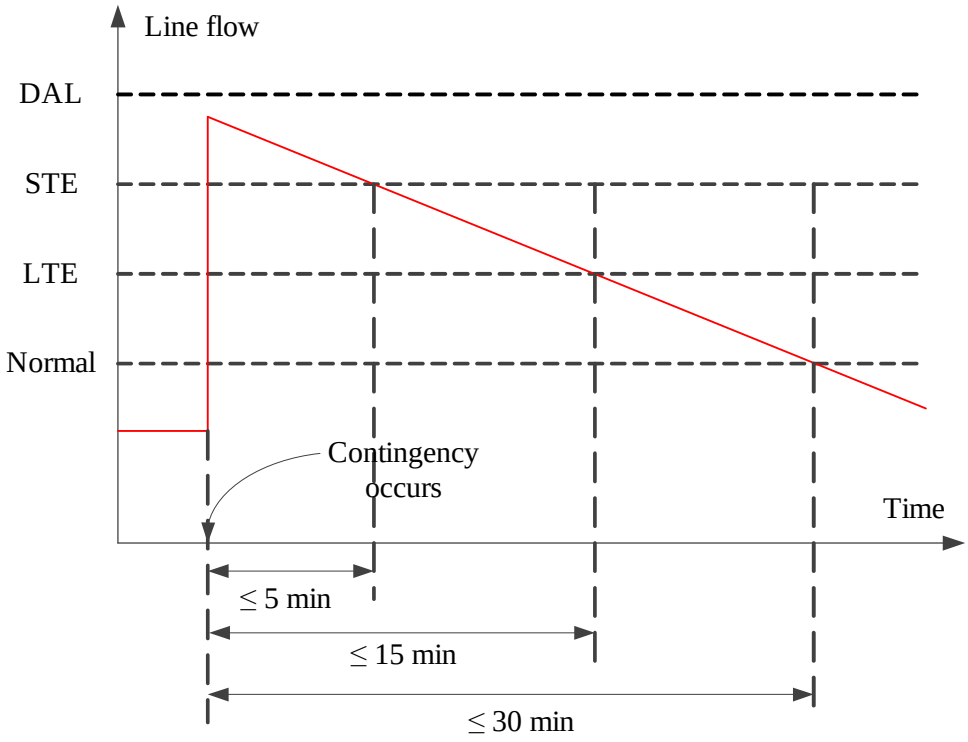


Figure 6.1: Post-contingency line flow requirements

results can directly compare with those of the existing software, enabling more credible evaluation of its industrial potential. Readers can consult Capitanescu et al. (2011) for an insightful review of the SCOPF problem and methodology.

6.2 The Model and its Structure

SCED with post-contingency corrective rescheduling (SCED-C) with K contingencies is written in the general form as follows (Monticelli et al., 1987a; Capitanescu and Wehenkel, 2008; Pinto and Stott):

$$\begin{aligned} \min_{x_0,\ldots,x_K,u_0,\ldots,u_K} \quad & f_0(x_0,u_0) \\ \text{s.t.} \quad & g_k(x_k,u_k)=0 \quad k=0,\ldots,K \\ & h_k(x_k,u_k)\leqslant 0 \quad k=0,\ldots,K \\ & |u_k-u_0|\leqslant \Delta_k \quad k=1,\ldots,K \end{aligned} \tag{6.1}$$

where f_0 is the base-case objective function and h_k and g_k are constraint functions. For the k-th system configuration, x_k is the vector of state variables and u_k is the vector of control variables. Δ_k is the vector of maximal allowed variation of control variables between the base case ($k=0$) and the k-th post-contingency configuration.

In a simplified linear "DC" network setting which we work with in this chapter, the control variables are the generation level P and the state variables are the voltage angle δ and the line flow F. The equality $g(x,u)=0$ corresponds to

$$\sum_{g(i)} P_g - \sum_{\substack{(j,c):\\(i,j,c)\in BR}} F_{i,j,c} + \sum_{\substack{(j,c):\\(j,i,c)\in BR}} F_{j,i,c} = D_i \qquad \forall i\in BUS$$

$$F_{i,j,c} - b_{i,j,c}(\delta_j-\delta_i)=0 \qquad \forall (i,j,c)\in BR$$

and the inequality $h(x,u)\leqslant 0$ corresponds to

$$F_{i,j,c}-\bar{F}_{i,j,c}\leqslant 0 \qquad \forall (i,j,c)\in BR \tag{6.2}$$

$$-F_{i,j,c}-\bar{F}_{i,j,c}\leqslant 0 \qquad \forall (i,j,c)\in BR \tag{6.3}$$

$$P_g^{min}-P_g\leqslant 0 \qquad \forall g\in GEN \tag{6.4}$$

$$P_g-P_g^{max}\leqslant 0 \qquad \forall g\in GEN \tag{6.5}$$

For simplicity we consider a linear cost function, i.e., $f_0(x_0,u_0):=c_0^T u_0$, where c_0 can be regarded as the marginal cost of generation. There may be multiple lines (or circuits) between two buses, hence the branch (i,j,c) indicates the c-th circuit connecting bus i and j. Functions g and h are identified by the following sets and parameters:

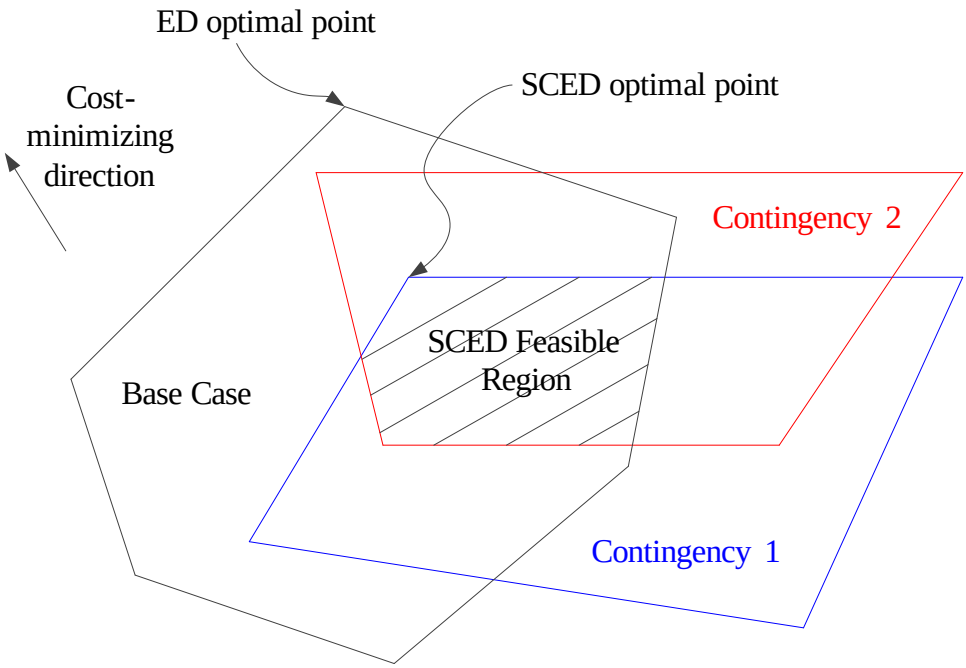


Figure 6.2: On the u_0 plane, the feasible region of an N-1 SCED is the intersection of N polyhedra. Since it would involve huge numbers of variables and constraints to represent all these polyhedra, we leave most of them out from the master model and use Benders' cuts to approximate the relevant pieces.

BR, BUS, GEN	Set of branches, buses and generators
D_i	Fixed load at bus i
$g(i)$	Set of generators connected at bus i
$b_{i,j,c}$	Susceptance of branch (i,j,c)
$\bar{F}_{i,j,c}$	Thermal rating of line (i,j,c)
P^{min}, P^{max}	Bounds of generator output

In this chapter, contingencies are the single line outages and there are several levels of post-contingency line ratings, so the functions g_k and h_k of post-contingency configuration k is actually represented by the set BR_k and parameters $(\bar{F}_{i,j,c})_k$. The feasible region of a DC-based SCED is the intersection of many polyhedra, as illustrated in Figure 6.2.

It is well-known that the linear DCOPF model is a coarse representation of the

physics of AC circuits in power systems. In particular, the power flow equations outlined above ignore real power losses as well as reactive power constraints. In fact, it is acceptable practice not to consider the reactive power in the economic dispatch solution which is primarily used to settle the market for real power. Detailed AC power flow studies usually follow at a later stage where various ancillary services come into play. It is possible to incorporate system losses in the linearized DC formulation, see Li and Bo (2007) for detail. However, because line loss is a quadratic function of the real power flow, it takes multiple iterations of the DCOPF run to achieve an accurate approximation of the total system loss as well as the loss factors (LF), which determine how the total losses are distributed/compensated across buses. Due to computational constraints, the ISOs often solve the real-time dispatch problem with estimated system loss and LFs and without the iterative process. This can be easily be considered into our model. Therefore, our choice of a DC model is practical. We have tested an extension of the model to account for line losses in the base case, adopting the fictitious nodal demand (FND) idea from Li and Bo (2007) (basically, evenly dividing/allocating the estimated loss on line (i, j, c) to the buses i and j as FNDs). The computational performance is indistinguishable to that of the model without considering the losses. For simplicity and limited by data availability, we use the lossless formulation for subsequent discussion and experiments.

SCED with Multi-stage Rescheduling

We now extend the general formulation (6.1) to the multi-stage rescheduling situation, where the time dimension plays an explicit role. The time index t will appear as superscripts on applicable symbols and the subscript k now indexes contingency cases with $k = 0$ being the pre-contingency case (base-case). Suppose the post-contingency operating procedure involves T checkpoints in time and there are K contingencies to prepare for in the SCED. When the system is operating at normal state (x_0, u_0), base-case feasibility requires that $g_0(x_0, u_0) = 0$ and $h_0(x_0, u_0) \leqslant 0$. When contingency k occurs, the state variable x will instantaneously change to

x_k^0 following physical laws, i.e., $g_k(x_k^0, u_k^0) = 0$, where $u_k^0 = u_0$ since the control variable cannot change abruptly. In general, security constraints require that the following conditions hold

$$\begin{aligned} & g_k(x_k^t, u_k^t) = 0 \\ & h_k(x_k^t, u_k^t) \leqslant \epsilon_t \\ & |u_k^t - u_k^0| \leqslant \Delta_t \end{aligned} \tag{6.6}$$

for a discrete set of time checkpoints t. For instance, ISO New England's operating procedure imposes the following checkpoints:

- $t = 0$ corresponds to the immediate checkpoint to ensure that the line flow is within the DAL rating. The components of ϵ_t for (6.4) and (6.5) are 0 (same for cases below) and the components of ϵ_t for (6.2) and (6.3) is $\bar{F}^{DAL} - \bar{F}^{Normal}$ (for notation convenience, we write $\epsilon_t = \bar{F}^{DAL} - \bar{F}^{Normal}$), $\Delta_t = 0$.

- $t = 1$ corresponds to the 5-minute checkpoint to ensure that the line flow is reduced within the STE rating, $\epsilon_t = \bar{F}^{STE} - \bar{F}^{Normal}$, $\Delta_t = 5R$, where R is the vector of per minute ramp rate of injection at the buses.

- $t = 2$ corresponds to the 15-minute checkpoint to ensure that the line flow is reduced within the LTE rating, $\epsilon_t = \bar{F}^{LTE} - \bar{F}^{Normal}$, $\Delta_t = 15R$.

- $t = 3$ corresponds to the 30-minute checkpoint to ensure that the line flow is reduced within the Normal rating, $\epsilon_t = 0$, $\Delta_t = 30R$.

For a given contingency k, condition (6.6) requires that there exists a recourse solution u_k^t for the time period between when the contingency occurs and the time of the checkpoint t. However, satisfying (6.6) for each t does not guarantee that the recourses at different time points are compatible with each other. For example, there may be a (injection) solution for the 5-minute checkpoint and a solution for the 15-minute checkpoint, respectively, but it may not be feasible to ramp from the 5-minute solution to the 15-minute solution. Figure 6.3 provides a numerical example that concretely demonstrates this issue. In the pre-contingency state, the 150 MW load at Bus 1 is supplied by Line 1 and Line 2. After the contingency

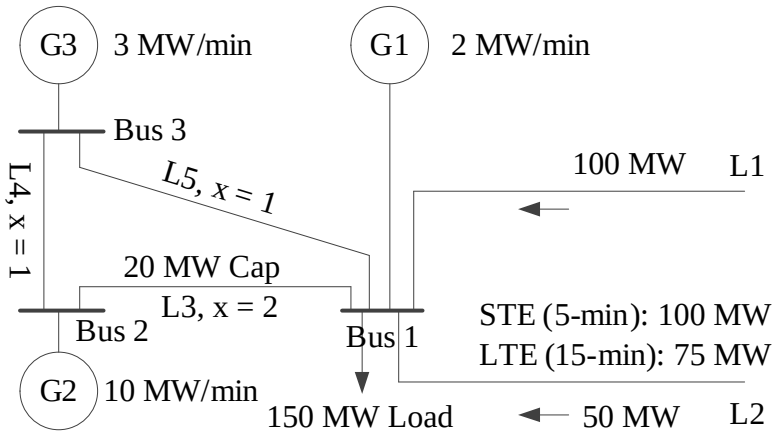


Figure 6.3: When L1 fails, flow on L2 will instantly rise to 150 MW. Ramping up G1 and G2 can meet STE requirement, while ramping up G1 and G3 can meet LTE requirement. However, STE and LTE can not be satisfied simultaneously.

(outage of Line 1) occurs, STE and LTE ratings of Line 2 requires the generators (G1, G2 and G3) to provide a total ramp-up of 50 MW in 5 minutes and 75 MW in 15 minutes, respectively. The STE rating requirement can be satisfied via ramping up G1 by 10 MW and ramping up G2 by 40 MW (of which 20 MW will flow on L3 and another 20 WM will flow on L4 and L5). At this stage, the flow on L3 has reached its capacity of 20 MW, so we cannot continue to ramp up G3 to meet the LTE requirement. Note that the parameter x on a line indicates the line's reactance, which dictates how the power flow distributes along different paths.

To deal with this situation, we postulate an alternative feasibility condition for contingency k that couples the contiguous time points, as follows,

$$\begin{aligned} & g_k(x_k^t, u_k^t) = 0 \\ & h_k(x_k^t, u_k^t) \leqslant \epsilon_t \\ & |u_k^t - u_k^{t-1}| \leqslant \Delta_t - \Delta_{t-1} \end{aligned} \tag{6.7}$$

for $t = 1, \ldots, T$ and $u_k^0 = u_0$. The resulting optimization problem is

$$\begin{aligned}
\min_{x,u} \quad & f_0(x_0, u_0) \\
\text{s.t.} \quad & g_0(x_0, u_0) = 0 \\
& h_0(x_0, u_0) \leqslant 0 \\
& g_k(x_k^t, u_k^t) = 0 \qquad k = 1, \ldots, K,\ t = 0, \ldots, T \\
& h_k(x_k^t, u_k^t) \leqslant \epsilon_t \qquad k = 1, \ldots, K,\ t = 0, \ldots, T \\
& |u_k^t - u_k^{t-1}| \leqslant \tilde{\Delta}_t \qquad k = 1, \ldots, K,\ t = 1, \ldots, T \\
& u_k^0 - u_0 = 0 \qquad k = 1, \ldots, K
\end{aligned} \tag{6.8}$$

where $\tilde{\Delta}_t = \Delta_t - \Delta_{t-1}$. This is a linear program when f, g and h are all linear as defined above. In the remainder of the chapter, we discuss computational techniques for efficient solutions of this LP. For illustration, we plot the sparsity pattern of a small problem instance in Figure 6.4. In the plot, the columns (variables) are arranged in the order $\{u_0, x_0, u_1^0, x_1^0, \ldots, u_1^T, x_1^T, u_2^0, x_2^0, \ldots\}$, and the rows (constraints) are arranged in the corresponding appropriate order. Note the inequalities $h_k(\cdot) \leqslant \epsilon_t$, i.e., (6.2) to (6.5), are handled as variable bounds hence do not appear in the matrix. It is apparent that the Jacobian is almost a band matrix if not for the constraints that link the control variable u_0 of the base-case with those of the contingency cases. Also note that the problem size grows linearly as the number of contingencies increases. These characteristics make the problem suitable for decomposition methods, which we discuss below.

6.3 Benders' Decomposition

The common Benders' decomposition scheme (Benders, 1962; Conejo et al., 2006; Li and McCalley, 2009) reformulates model (6.8) into an equivalent form, as follows.

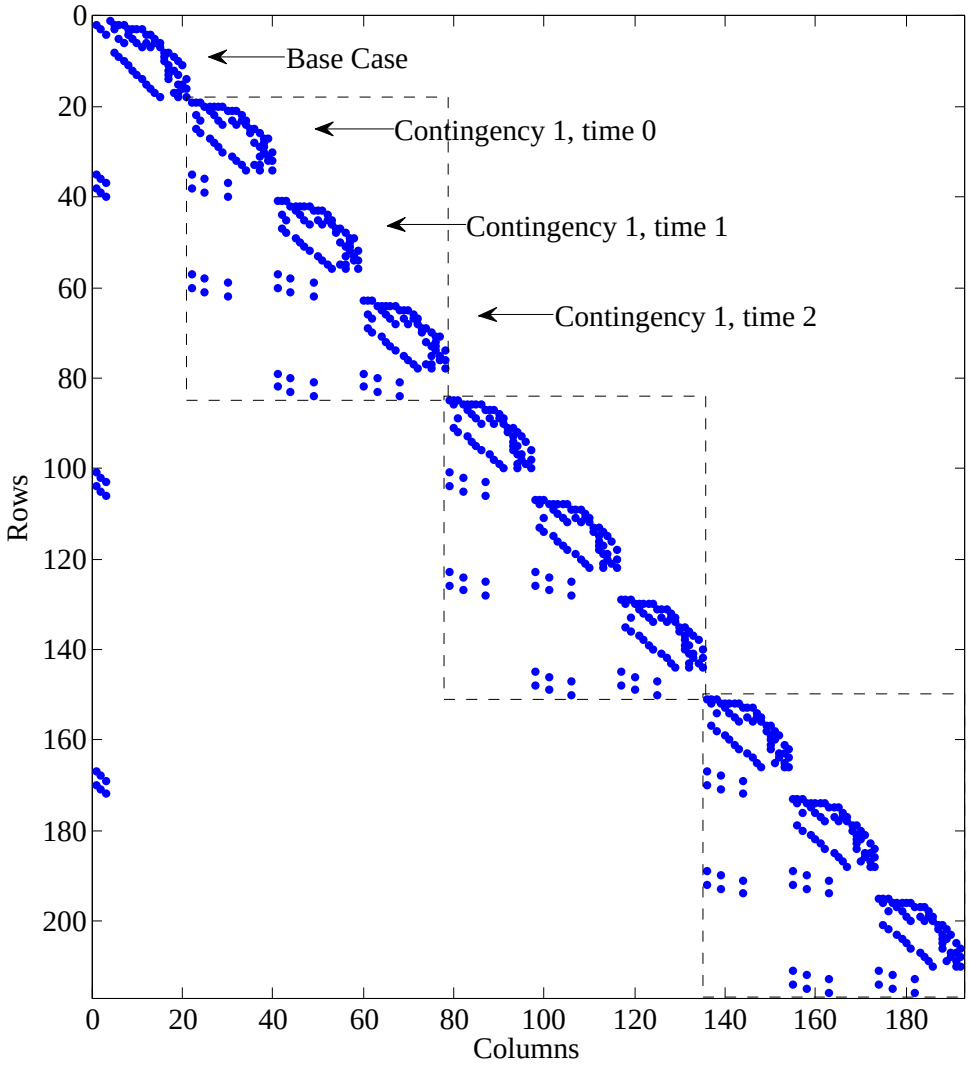


Figure 6.4: Sparsity structure of the Jacobian matrix of a 6-bus case with 3 contingencies and 3 post-contingency checkpoints.

$$\min_{x_0,u_0} \quad f_0(x_0,u_0) \tag{6.9}$$

$$\text{s.t.} \quad g_0(x_0,u_0)=0 \tag{6.10}$$

$$h_0(x_0,u_0)\leqslant 0 \tag{6.11}$$

$$w_k(u_0)\leqslant 0 \qquad k=1,\dots,K \tag{6.12}$$

where $w_k(u_0)$ is the value function of the k-th sub-problem, given by

$$\begin{aligned} w_k(u_0)= \min_{x_k,u_k,s_k} \quad & \|s_k^t\| & \\ \text{s.t.} \quad & g_k(x_k^t,u_k^t)=0 & t=0,\dots,T \\ & h_k(x_k^t,u_k^t)\leqslant \epsilon_t & t=0,\dots,T \\ & |u_k^t-u_0^t|-s_k^t\leqslant \tilde{\Delta}_k & t=1,\dots,T \\ & u_k^0-u_0=0 & \\ & s_k^t\geqslant 0 & t=1,\dots,T \end{aligned} \tag{6.13}$$

Note that the s_k is an artificial variable added to model the constraint violations. Since the sub-problem is a linear program, it follows from LP duality that any given point $\bar{u}_0$ and the associated value $w_k(\bar{u}_0)$ (denoted by $\bar{w}_k$) can provide a linear function of u_0 that underestimates $w_k(u_0)$, as follows,

$$w_k(u_0)\geqslant \bar{w}_k+\bar{\lambda}_k(u_0-\bar{u}_0) \tag{6.14}$$

where $\bar{\lambda}_k$ is the Lagrangian multiplier of constraint (6.13) at the solution. It is easy to see that

$$\bar{w}_k+\bar{\lambda}_k(u_0-\bar{u}_0)\leqslant 0 \tag{6.15}$$

is a necessary condition for (6.12) hence is a valid inequality for the master problem. As a substitute for (6.12) which is hard to impose directly, it will cut off the point $\bar{u}_0$ if $\bar{w}_k$ is positive.

The Benders' decomposition algorithm alternates between solving the master problem and the sub-problems, hence approaches a better and better satisfaction of (6.12) until certain convergence criteria are met. In each iteration, the master problem (6.9) - (6.11) with previously added cuts is solved. Subsequently the subproblems are solved one by one given the master solution. Each subproblem solution having a positive objective value will supply a new cut to the master problem for the next iteration. For a given problem instance, the number of variables in the master problem and size of the subproblem are fixed regardless of how many contingencies there are to consider. In this sense, the algorithm "decomposes" the big LP by approaching its solution via repeatedly solving smaller LPs.

It is worth noting that model (6.8) can be succinctly expressed as a two-stage stochastic program (SP) in the extended mathematical modeling (EMP) framework within the GAMS modeling software. An SP solver, e.g., DE and Lindo, can then be called to solve the problem in both the deterministic equivalent form (big LP) and the Benders' decomposition form. However, the Benders' algorithm implemented in Lindo (which is the only general purpose Benders' code available) is unable to handle any problem-specific structure or solve the subproblems in parallel. For example, it takes Lindo about 31 minutes to solve the 118-bus 183-contingency case using its Benders' algorithm, even worse than the "Vanilla" Benders' algorithm that we implemented directly in GAMS (computation times for this case is listed in the first row of Table 6.2).

6.4 Computational Enhancements

Formulation

The variable s_k^t in the subproblem captures the violation or infeasibility of the ramping constraints evaluated at the candidate solution u_0. In the literature, e.g., Monticelli et al. (1987a), Capitanescu and Wehenkel (2008) and Pinto and Stott, all authors used the L_1 norm in the subproblem objective function, i.e., minimizing the sum (over all buses) of violations. We find that the following modifications to

the subproblem formulation reduce the number of Benders' iterations required for convergence in many cases, as demonstrated in Figure 6.5. We will apply these modifications in all subsequent discussions.

- Use L_∞ norm of s_k^t in the objective, i.e., minimizing the maximum (over all constraints for which the violation variable is added) violation. This is aligned with the normalization idea from Fischetti et al. (2010).
- Allow violation in the inequality constraints, i.e., substitute $h_k(x_k^t, u_k^t) - s_k^t \leqslant 0$ for $h_k(x_k^t, u_k^t) \leqslant 0$ in the subproblem. This also provides convenience in detecting infeasible contingencies, as will be discussed below.

We do not have a proof of the advantage of using L_∞ norm over using the L_1 norm. However, reductions in the number of iterations are consistently observed (although not always as significant as shown in the figure) and we have not encountered any case that takes more iterations using L_∞ rather than L_1. Since all known papers on SCED-C happened to explicitly adopt the L_1 norm in their formulations, we find it useful to report our observations here. As a side note, the 118-bus instance used in Figure 6.5 was made harder to solve (so that it takes Benders' algorithm many iterations to converge) by using nodal loads 1.8 times higher than the original values.

Dealing with Infeasibility

In practical use of the SCED model (with or without the corrective rescheduling, single- or multi-stage), there is an implicit assumption (belief) that a feasible solution exists, i.e., the security constraints are satisfiable. Indeed, if it frequently occurs that no operating point is able to meet the security criteria, it probably indicates that the criteria are too restrictive and need a change. Realistically, not all lines in the network are included in the contingency list of SCED but only those having crucial importance, e.g., high-voltage backbone transmission lines, and those for which the consequence of failure is controllable (by which we mean "no load is lost") via dispatch or rescheduling. Lines whose failure would island a load bus, for

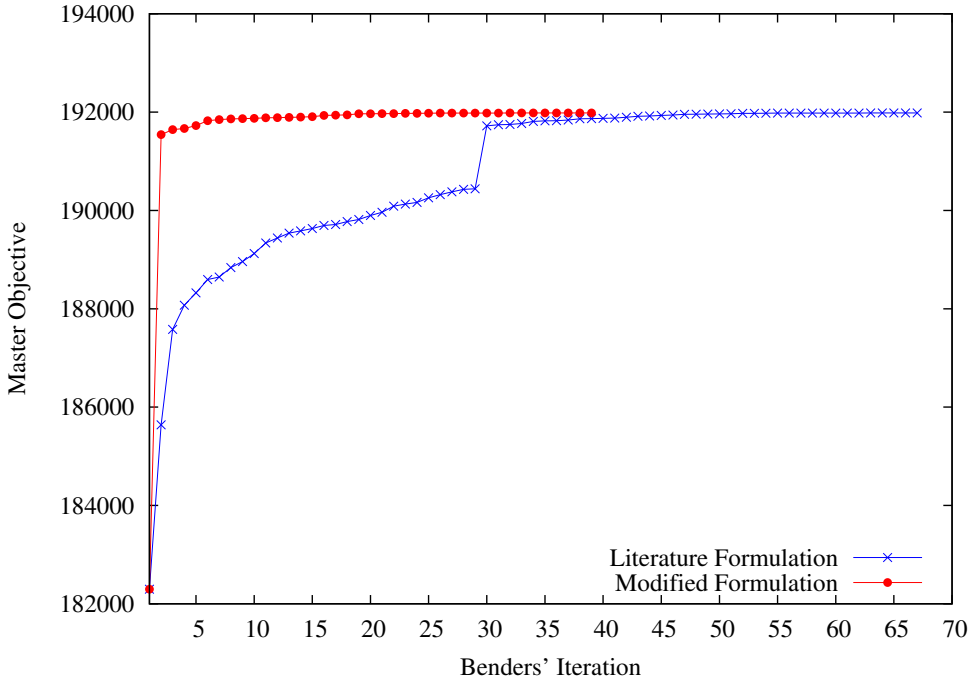


Figure 6.5: Performance of different formulations on an instance of 118-bus case with 20 contingencies (subproblems).

example, should not be included in the contingency list for SCED, because nothing can be done beforehand to avoid shedding load should the failure occur. Despite the sensible selection of contingencies, there can be no a-priori guarantee for the existence of a feasible solution before actually running the SCED. Being notified that the model is infeasible is the last thing system operators want to see from a SCED run – in this case, they at least need to know what is causing the infeasibility, if not how to correct it. One way of avoiding infeasibility is to allow constraint violation (i.e., load shedding) and penalize it in the objective function, see, e.g., Jiang and Xu (2013). There are three drawbacks for this method: (1) Determination of the penalty factor is almost entirely arbitrary; (2) Shedding a load simply because its supply line MAY fail is impractical[1]; (3) Keeping a large number of penalty variables (i.e.,

[1]This would happen if the only line that connects a load bus with the rest of the network were in the contingency list. A practical treatment should be to shed the load when the contingency actually occurs.

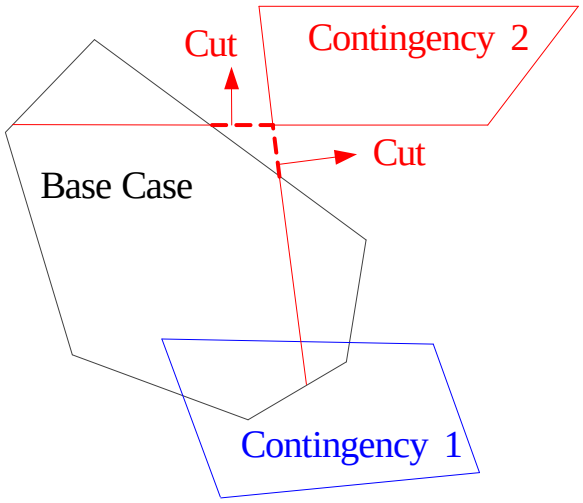


Figure 6.6: Contingency 2 is intrinsically infeasible. Either the corresponding subproblem is infeasible or its Benders' cuts will render the master problem infeasible.

one for each equality constraint) in the (master) model to prepare for infeasible situations which only occur occasionally is not efficient modeling practice.

Alternatively, we integrate in the Benders' algorithm a mechanism to dynamically identify and remove the contingencies that would cause infeasibility. Denote the base-case feasible set by $\mathcal{F} = \{u_0 \in R^n | \exists x_0$ such that $g_0(x_0, u_0) = 0, h_0(x_0, u_0) \leqslant 0\}$ and the feasible set for contingency k by $\mathcal{F}_k = \{u_0 \in R^n | \exists (x_k^t, u_k^t)$ such that (6.7) holds.$\}$. Let us assume base case feasibility, i.e., $\mathcal{F} \neq \emptyset$.

We call a contingency k *intrinsically infeasible* if $\mathcal{F} \cap \mathcal{F}_k = \emptyset$. For such a contingency, the corresponding subproblem is either infeasible for all $u_0 \in \mathcal{F}$, which indicates $\mathcal{F}_k = \emptyset$, or optimal with a positive objective value for all $u_0 \in \mathcal{F}$. In the former case, the subproblem will be infeasible in the first run (in the first iteration) and we can remove the contingency immediately[2]. In the latter case, the subproblem keeps generating cuts for the master problem until the master problem becomes infeasible due to conflicting cuts.

[2]It must be the power balance constraint in the subproblem that makes it infeasible, since violation is allowed everywhere else. This corresponds to the situation where the contingency isolates a load node or sub-network from the rest of the network, a situation that is not considered insecure in the "N-1" security context.

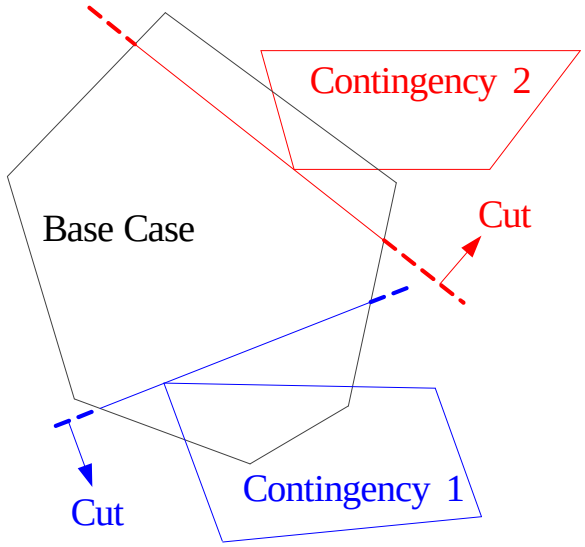


Figure 6.7: Each individual contingency is feasible, but they are not simultaneously feasible. Their Benders' cuts will render the master problem infeasible.

Another source of infeasibility comes from the case where multiple contingencies are not simultaneously feasible, e.g., $(\cap_{k=1}^{K}\mathcal{F}_k) \cap \mathcal{F} = \emptyset$. Such a case manifests itself in the form of an infeasible master problem caused by conflicting cuts. The two cases of conflicting cuts are illustrated in Figure 6.6 and 6.7.

Our order of business is to remove the "problematic" contingencies whenever the master problem becomes infeasible. We do this by solving a modified master model, constructed by adding a nonnegative violation variable v_k^i to each of the previously added cuts as well as adding a linear term in the objective function to penalize the violation (with a penalty factor M). The solution of this model indicates the violated cuts, those for which the violation variable is positive. We then remove any contingency that has contributed one or more violated cuts. The modified master problem is outlined below.

$$
\begin{aligned}
\min_{x_0,u_0} \quad & f_0(x_0,u_0) + \sum_{(k,i)\in \mathrm{CUT}} M v_k^i \\
\text{s.t.} \quad & g_0(x_0,u_0) = 0 \\
& h_0(x_0,u_0) \leqslant 0 \\
& \bar{w}_k^i + \bar{\lambda}_k^i (u_0 - \bar{u}_0^i) - v_k^i \leqslant 0 \qquad \forall (k,i) \in \mathrm{CUT} \\
& v_k^i \geqslant 0 \qquad \forall (k,i) \in \mathrm{CUT}
\end{aligned}
$$

where the superscript i indexes the iteration and $(k,i) \in \mathrm{CUT}$ means that contingency k has generated a cut in iteration i.

Note that by penalizing the "sparsity inducing" L_1 norm of the violation, we intend to approximately identify a minimal number of problematic contingencies whose removal would restore feasibility, which corresponds to a NP-hard problem. This approach is motivated by sparse optimization methods (Wright, 2009) and is shown to be effective in our experiments.

Algorithmic Enhancement

The master problem is a small-sized (relative to the subproblems) linear program and is easy to solve. It is also easy to update the optimal solution from one iteration to the next, since adding cuts does not change the dual feasibility of an LP. Taking a 2383-bus case for example, it takes the CPLEX dual simplex method less than 1 second to solve the master model from scratch and takes less than 0.3 second for solution updates between successive iterations. In contrast, the majority of the solution time is spent on solving the many subproblems, each of which is also three times larger than the master problem (not counting the cuts), see Figure 6.4. We tailor several enhancement schemes to solving the subproblems using GAMS and CPLEX.

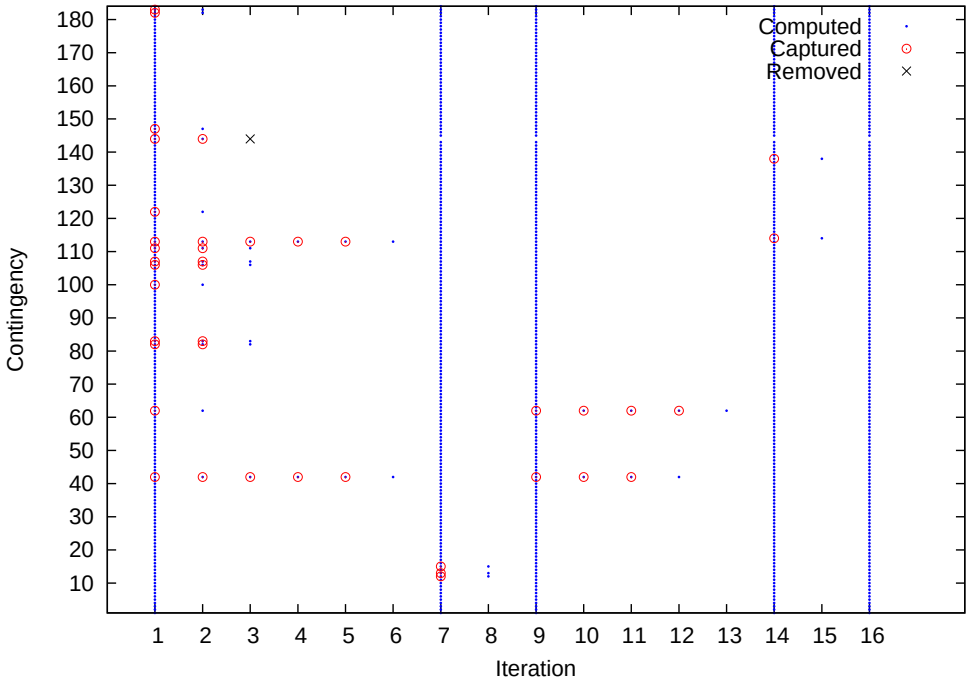


Figure 6.8: Algorithm progress on the 118-bus case.

Reducing the number of LP runs

In practice, although a long list of contingencies need to be considered, most of them turn out not to be binding in the final SCED solution. In the Benders' solution process, we observe via experiments that if a contingency is feasible (i.e., its subproblem has an optimal value of 0) in an iteration, it is also likely to be feasible in subsequent iterations. This observation can be exploited to improve the algorithmic design, as follows. At any iteration if a contingency becomes feasible, we temporarily exempt it from the feasibility test in subsequent iterations, which results in a shrinking list of *active* contingencies as the algorithm progresses. When this list is empty, we import the whole list of contingencies again and test the simultaneous feasibility of the current master solution. If it is feasible for all contingencies, the algorithm terminates; otherwise, the above process continues. This method could greatly reduce the total number of LP runs.

We apply this approach to the 118-bus case and demonstrate the solution process

in Figure 6.8. There are 186 lines in the network and two of them (i.e., bus 12 to bus 117 and bus 68 to bus 116) are determined to be intrinsically infeasible by pre-screening, so we monitor the remaining 184 lines in the experiment. In iteration (iter) 1, 184 subproblems are "computed" among which 14 are "captured" to have a positive objective value. In iter 2, only those 14 subproblems captured in the previous iteration are computed, so in this fashion we are dealing with a shrinking list of active contingencies. In iter 6, two are computed but none is captured, which means the active list becomes empty. Therefore, in iter 7 the active list is reset to the whole list and three were captured. This process repeats until in iter 16 none is captured after computing for the whole list, which means that the current u_0 is feasible for all contingencies and the algorithm terminates. Note that in iter 3, contingency #146 (i.e., bus 85 to bus 86) is removed due to its causing infeasibility. The correctness of its removal has been verified by solving the full LP formulation.

Using barrier method without crossover for subproblems

CPLEX offers several options for LP methods, including primal simplex, dual simplex, network simplex, barrier and concurrent. By default, CPLEX chooses the simplex method to solve the subproblem LPs. We found via experiments that the barrier method actually solves the subproblem faster than the simplex method for big instances (e.g., 2383-bus case). CPLEX barrier optimizer automatically invokes a crossover process when the barrier algorithm terminates, in order to produce a basic solution. However, any optimal solution, not necessarily a basic one, suffices to generate a valid cut (6.15). Therefore, we can turn off the crossover to save time, i.e., setting CPLEX option *barcrossalg=-1*. Table 6.1 shows the potential time saving of this choice of solver options.

Solving batches of subproblems in parallel

Benders' decomposition algorithm is naturally amenable to parallel computing. We harness the multi-core hardware and multi-threading capability of the operating system to solve the subproblems in parallel during each iteration. Specifically,

Table 6.1: Time (seconds) Spent in Sequentially Solving 100 Subproblems using Different LP Methods, on a Dell Laptop[6]

Case	Subproblem Size			Dflt Splx	Barrier	Barrier \Xover
	Row	Col	NZ			
118-bus	1070	2668	8545	14.9	14.4	13.4
2383-bus	16814	37129	115006	453.6	139.0	79.8

we evenly divide the active contingencies into N batches and run the batches on separate processors. In solving the series of subproblem LPs in each batch, we use the GUSS facility (Bussieck et al.) within GAMS to shorten the overall solution time. As the subproblems are similar in structure, GUSS constructs the model rim once and plugs in different parameter data for different LPs. This eliminates the repetitive work of building the model from scratch for each LP, thus saves time.

Invoking feasibility checker and adding difficult contingencies to the master problem

Experiments on different realistic data sets provide the following observation. Between two successive whole-list scans (e.g., between iter 1 and 7 in Figure 6.8), most contingencies will be removed from the active list in a few (less than ten) iterations, but at times a small number of contingencies will remain in the active list for many iterations before becoming feasible or proven to be intrinsically infeasible (by the conflicting cuts of Figure 6.6). Such contingencies incur extended computational costs in two ways: (1) Dealing with only a few subproblems per iteration is an inefficient use of parallel computing considering its overhead and (2) if the persistent contingency was intrinsically infeasible, all the iterations spent in detecting the infeasibility would be "wasted". In other words, if infeasibility was detected earlier, much time could be saved. We mitigate these difficulties as follows. When the size of the active contingency list drops to a certain threshold level L^{fc}, we initiate a "feasibility checker" job to run in parallel with the main Benders' loop. This job checks on an individual basis the feasibility for all the contingencies on the active list, by solving for each contingency a reduced SCED model consisting of only the

base case and the target contingency that is being checked. The feasibility checker (FC) model for contingency k is as follows.

$$\begin{aligned} \min_{u_0, x_0, r^+, r^-} \quad & \|r^+ + r^-\| \\ \text{s.t.} \quad & g_0(x_0, u_0) + r^+ - r^- = 0 \\ & h_0(x_0, u_0) \leqslant 0 \\ & u_0 \in \mathcal{F}_k \\ & r^+,\, r^- \geqslant 0 \end{aligned}$$

FC cannot be infeasible since we assumed that $\mathcal{F} \neq \emptyset$. A positive optimal value of FC indicates that the corresponding contingency is intrinsically infeasible. At the beginning of each subsequent iteration after the checker job is invoked and before it is finished, the main algorithm checks the status of the ongoing checker job. Once the checker job is finished, its results, i.e., whether the contingencies are feasible or not, are passed to the main algorithm. A myriad of heuristics can be designed to utilize these results. The most obvious step is to remove the intrinsically infeasible contingencies, if any, from the contingency list. In addition to this, we also add the feasible contingencies that are still in the active list[3], which are anticipated to be "difficult" ones that would cost many more iterations to deal with, directly to the master problem, forming and solving a larger master problem and starting the next whole-list scan immediately afterwards.

The choice of the L^{fc} level is influenced by different factors. If it is too big, the feasibility checker results may suggest adding excessive number of contingencies to the master problem which would increase its subsequent solution time. On the other hand, a small value of L^{fc} means the algorithm is likely to spend more iterations reducing the size of the active list to L^{fc} before the feasibility checker is initiated. The relative speed of progress between the main Benders' loop and the feasibility checker is also an influencing factor for choosing L^{fc}.

[3]The size of the active list may be smaller than L^{fc} by the time the feasibility checker job is finished.

6.5 Numerical Experiments

We use the IEEE 118-bus test case as well as several sets of the Polish network data for experiments. Although these data sets are available from various sources, e.g., the Matpower package, the original data lack meaningful values for line thermal ratings and generator ramp rates, which are critical for the SCED with multi-stage rescheduling[4]. We adopt the data provided by a FERC project, which made up the missing values based on reasonable engineering assumptions and enriched the existing format, refer to Molzahn (2013). In particular, the rateA, rateB and rateC of a branch are taken as the $\bar{F}^{\text{Normal}}$, $\bar{F}^{\text{LTE}}$ and $\bar{F}^{\text{STE}}$, respectively, associated with three post-contingency stages (checkpoints). The algorithms are implemented in the GAMS modeling software. Throughout the experiments, we use the feasibility tolerance of 10^{-6}, i.e., if the subproblem objective value $||s_k||$ is lower than 10^{-6}, the contingency is considered feasible.

Comparison of Performance on Feasible Instances

Table 6.2 demonstrates the effect of different algorithmic enhancements on the solution speed. The "Big LP" formulation is solved by the simplex algorithm (CPLEX default) and the barrier method without crossover. Note that the crossover step may take significantly longer time than the core barrier algorithm. The subsequent columns of the table lists results (i.e., number of iterations taken, number of subproblem LPs solved and the total solution time in seconds) brought by incrementally added features (as described by the header of the column). The "Vanilla Benders" represents the original Benders' algorithm with the formulation improvement described in Section 6.4. "RedLP+Opt" implements the first two enhancement schemes described in Section 6.4, i.e., implementing algorithmic control to reduce the number of LP runs and using appropriate solver options for subproblem speedup. "Paraguss", in addition, solves the subproblems in (8 or 40, depending on problem size) parallel batches and uses the GUSS facility in each batch. "Fatmaster

[4]The original data may have all otherwise non-trivial contingencies feasible for any base-case dispatch, which would make experiments unilluminating.

(5)" represent the complete set of features presented in this chapter. In addition to "Paraguss", it enables the feasibility checker (with $L^{fc} = 5$) and adds difficult contingencies to the master model. All the implementations that have solved the given case within the time limit (2 hours) have obtained the same optimal objective value, listed in the last column.

Because neither the "Big LP" formulation nor the original Benders' algorithm is capable of handling infeasible contingencies and an infeasible problem is not an interesting subject for comparison, we use a pre-screened list of lines as contingencies which are guaranteed to constitute a feasible SCED case. In Table 6.2, the first five cases in the upper half of the table are small to medium-sized. For the 118-bus case, we monitor (meaning: set as contingency) all the 183 lines in the pre-screened list and for the 2383-bus case[5], we monitor the first 20, 50, 100 and 400 lines, respectively, in the pre-screened list. These cases are run on a Dell laptop[6]. We use 8 parallel processes in the "Paraguss" implementation. It is apparent that each enhancement scheme yields substantial time savings.

The next five cases in Table 6.2 are the largest possible feasible instances on the corresponding network data (wp means winter peak and wop means winter off-peak, etc.), as we monitor the complete pre-screened list of lines. These cases are run on a Dell R710 server with two 3.46G X5690 Xeon Chips, 12 Cores and 288GB Memory. The "Big LP" formulation is unable to solve any of these cases within 2 hours due to the large model size. For example, the 2383-bus 2349-contingency case results in a 18GB LP for the solver. We ignored the uncompetitive "Vanilla Benders" algorithm on these big cases but instead ran the "RedLP+OPT" without the 2-hour time limit. It is worth noting that our final approach "Fatmaster" is able to solve all cases in 10 minutes and solve most cases well within 5 minutes.

[5]In the Polish network case names, the suffix "wp" means winter peak, "sop" means summer off-peak and so on.

[6]Dell precision M4500 with Intel Core i7 CPU Q840 @1.87GHz, 8GB RAM, on Windows 7.

Table 6.2: Solution Statistics of Different Formulations

Case	Ctgcy	Big LP		Vanilla Benders			RedLP+Opt			Paraguss			Fatmaster (5)			Cost
		Splx	Bar	Iter	LPs	Time	Iter	LPs	Time	Iter	LPs	Time	Iter	LPs	Time	
118-bus	183	207.8	13.8	8	1464	123.5	10	764	72.6	14	776	15.1	12	755	13.5	86206.9
2383wp	20	175.0	205.5	52	1040	1281.2	46	115	99.8	48	117	95.4	11	60	41.5	1832425.7
2383wp	50	1403	123.1	49	2450	2799.3	48	193	160.3	48	193	101.7	11	135	46.5	1832425.7
2383wp	100	3621	240.6	32	3200	3688.6	33	289	226.0	32	288	96.3	12	245	79.4	1832813.4
2383wp	400	-	2354.5	-	-	-	35	953	913.3	38	956	218.0	13	879	197.8	1838910.7
2383wp	2349						106	12123	12165	104	9788	769.5	21	9529	515.7	1894241.7
2736sp	2749						45	5543	5836	44	5542	366.2	4	5500	220.9	1289173.9
2737sop	2753						1	2753	2801	1	2753	100.1	1	2753	100.5	764008.6
2746wop	2794						1	2794	3046	1	2794	118.3	1	2794	118.5	1178164.0
2746wp	2719						262	8646	9738	278	8622	1427.7	14	5558	333.5	1608584.3

Performance on Possibly Infeasible Instances

Given a network case and a list of contingencies, it is not known a priori whether the data represents a feasible SCED instance or not. A straightforward first step to "purify" the data is pre-screening the intrinsically infeasible contingencies one by one, which involves solving the model FC for each contingency k in the list. Table 6.3 lists the results of this process run on all lines in the network. We can see that pre-screening is very time-consuming. Even if parallel computing involving 100 processors were utilized, it would still take several hundred seconds to pre-screen a large network case.

Table 6.3: Time (seconds) Spent to Pre-screen for Different Cases. The LPs Are Solved Sequentially.

Case	# Lines	# Feasible	# Removed	Time
2383wp	2896	2353	543	49670.8
2736sp	3269	2749	520	76068.8
2737sop	3269	2753	516	13069.2
2746wop	3307	2794	513	20160.2
2746wp	3279	2719	560	43618.7

Table 6.4: Solution for Big Cases, 80 threads, $L^{fc} = 5$

Case	Ctgcy	Iter	LPs	Time	Added	Tabu
2383-bus	2896	15	7694	522.1	6	547
2736-bus (sp)	3269	4	6020	252.9	1	520
2737-bus (sop)	3269	4	6023	242.2	0	516
2746-bus (wop)	3307	4	6102	280.2	0	513
2746-bus (wp)	3279	8	6053	334.3	4	560
2383-bus	2353	16	7156	460.6	6	4
2736-bus (sp)	2749	4	5498	245.9	1	0
2737-bus (sop)	2753	1	2753	110.8	0	0
2746-bus (wop)	2794	1	2794	131.7	0	0
2746-bus (wp)	2719	14	5558	354.4	4	0

In contrast, our approach of dealing with infeasibility takes little extra time and is able to identify infeasible cases (as in Figure 6.7) to which the pre-screening is blind. This is demonstrated in Table 6.4. The upper half of the table are experiments that include all lines in the initial contingency list, i.e., the "N-1" cases. The column "Added" lists the number of contingencies that have been added to the master problem in the solution process. The numbers are small, indicating that the choice of $L^{fc} = 5$ is appropriate for these cases. The column "Tabu" lists the number of of contingencies that have been removed during the run due to infeasibility. The lower half of the table are experiments that only takes the pre-screened lines (the "Feasible" lines coming from Table 6.3) as contingencies. We can see that the solution times of the two cases do not differ much, which means that our approach of removing infeasibility is much more efficient than pre-screening. Furthermore, the fact that 4 extra lines (other than those identified by the pre-screening) are removed in the 2383-bus case indicates that (1) Pre-screening is indeed unreliable in practice and (2) our method is effective at approximating a minimal set of problematic contingencies when the problem is infeasible.

Table 6.5: Active Contingencies at Optimum

Case	Ctgcy	Active ctgcy number at optimal solution
118-bus	186	**43**, **63**, 124, 184, 185
2383-bus	2896	**344**, **414**, **546**, **1798**, **2164**
2736-bus (sp)	3269	170, 292, **576**
2737-bus (sop)	3269	
2746-bus (wop)	3307	
2746-bus (wp)	3279	**3**, **4**, **440**, **573**

Table 6.5 provides the list of binding contingencies at the optimal solution of the "N-1" cases. A contingency is identified as binding at the solution if any cut it contributes has a nonzero multiplier value, or the ramping constraint (6.13) has a nonzero multiplier if the contingency has been added to the master model. In the table, numbers in bold face are binding contingencies in the master model. We can see that very few contingencies are binding at the optimal solution, although a larger number of contingencies have been active along the algorithm iterations.

As a companion of Table 6.5 which only lists the contingency numbers, Table 6.6 provides the mapping from a contingency number k to line identification (i, j, c), in the form of $k : (i, j)$. The circuit number c is omitted with a note that all the lines listed here has a circuit number $c = 1$. For example, the first entry "43: (26,30)" in the 118-bus case reads "contingency # 43 is the outage of the line (circuit # 1, if there are multiple) connecting bus 26 and bus 30."

Table 6.6: Contingency-to-line Mapping for Active Contingencies at Optimum

118-bus	2383-bus	2736-bus	2746-bus
43: (26,30)	344: (310,6)	170: (131,75)	3: (7,8)
63: (38,65)	414: (367,14)	292: (131,75)	4: (8,14)
124: (71,73)	546: (477,420)	576: (508,361)	440: (395,21)
184: (110,111)	1798: (1514,894)		573: (503,493)
185: (110,112)	2164: (1845,135)		

Economic Gain of Multi-stage Corrective Rescheduling

We compare the ISO's current "0-stage" SCED, i.e., imposing the Normal line rating for both the base case and contingency cases and not considering post-contingency corrective actions, and our proposed model considering 3-stage rescheduling. In the experiments, we use the same algorithm and enhancements for both models. The computational results for "N-1" cases are shown in Table 6.7. Due to its larger sized subproblems, the 3-stage model takes more time to solve. However, the dispatch solution yields noticeable cost savings compared to that of the 0-stage model. Furthermore, in some cases the 3-stage model also gives rise to fewer uncontrollable contingencies (those put in Tabu), which is an advantage in real-world operations.

Table 6.7: SCED Solution Considering Different Post-contingency Stages

Case	0-Stage			3-Stage			% Saving
	Tabu	Time	Cost	Tabu	Time	Cost	
118-bus	3	13.1	93046.4	3	26.6	86206.9	7.35%
2383wp	553	213.9	1903510.4	547	522.1	1894241.7	0.49%
2736sp	520	124.3	1297672.0	520	252.9	1289173.9	0.66%
2737sop	516	80.5	764056.7	516	242.2	764008.6	0.01%
2746wop	513	93.4	1178683.4	513	280.1	1178164.0	0.04%
2746wp	566	207.8	1632181.2	560	334.3	1608584.3	1.45%

6.6 Conclusion

Incorporating the post-contingency rescheduling actions into the SCED model provides better economic efficiency but also increases the computational difficulty which hinders its industrial application. Our work contributes to the advancement of the SCED-C research in both modeling and computation aspects. First, we have proposed a model to correctly address the multiple stages of rescheduling requirement found in realistic operating procedures. Second, we have devised a series of computational enhancements to solve the proposed model. The enhanced

algorithm was shown to be much faster than the original Benders' algorithm and was able to solve large instances within reasonable amount of time. Finally, our computational results could serve as an estimate on how far/close the current technology is to the industrial deployment of the multi-stage SCED-C model. Future work will quantify the economic benefit of multi-stage rescheduling and investigate the application of this solution approach to an online setting.

7 SECURITY-CONSTRAINED ECONOMIC DISPATCH USING SEMIDEFINITE PROGRAMMING

7.1 Introduction

Economic dispatch is an optimization problem that seeks a minimum-cost generation plan to meet system load in an electrical power network. Assuming that system load is constant within the planning horizon, economic dispatch is usually cast as a short-term resource planning problem (rather than a real-time control problem). In restructured electricity markets, it is primarily used to settle energy transactions among market participants, at which stage numerous power engineering details, including reactive power, voltage stability and ancillary services, are beyond the scope of the model. Therefore, a linearized "DC" representation of the network often suffices for dispatch purposes. To guard against disruptions in supply and transmission, various constraints are included in the dispatch model that ensure operability after a set of failure conditions, giving rise to a class of security-constrained economic dispatch (SCED) models. A widely used security criterion requires that the dispatch solution must leave room for an "escape route" to prevent the system from collapsing in the case of a (single) major component failure. Unlike the economic aspect of the problem which is a "planning" problem by nature, security issues are formulated at the operation level and therefore call for a more accurate representation of the system. For one thing, to judge if the system state is truly operational involves not only looking at the real power flows, but also examining the voltage magnitudes and reactive power flows. Therefore, a SCED model is preferably built upon the AC power flow equations. An AC-based SCED is widely studied under the name of security-constrained optimal power flow (SCOPF) in the literature, e.g., Capitanescu and Wehenkel (2008) and Jiang and Xu (2013), as the AC-based economic dispatch is traditionally termed an ACOPF problem.

The *ultimate goal* of SCOPF is to find a dispatch solution that is capable of being

ramped to a feasible AC operating point under all contingency cases. But such a goal is in general difficult to attain, for two reasons. First, the AC power flow equations pose quadratic equality constraints on the OPF formulation, which make the optimization problem nonconvex. Second, the model size grows linearly with the number of contingencies and can easily outgrow a manageable scale. Various methods have been explored in the literature, e.g.,Capitanescu and Wehenkel (2008), Capitanescu et al. (2007a) and Phan and Kalagnanam (2014), to tackle different aspects of the problem. However, a scalable method is still to be developed, let alone deployed in a practical setting.

In this chapter, we propose an approach to obtain high-quality local solutions for large-scale SCOPF instances. The solution can guarantee a feasible AC operating point under all controllable contingencies while maintaining the dispatch cost close to that of the global solution. In essence, we use a Benders' decomposition framework to tackle the scale issue, and use a semidefinite programming (SDP) relaxation of the AC model to serve as convex subproblems in the Benders' algorithm. In the sections to follow, we will develop the approach in detail and evaluate its effectiveness in relation to the ultimate goal.

Brief Review of the SDP Literature

The remarkable accuracy of our local solution is mainly attributed to the use of an SDP relaxation. Semidefinite programming has been intensively studied in power systems applications in recent years; its efficacy for finding global solutions being widely acknowledged. Among the growing literature, we find the following papers especially helpful in developing our work. In a clear step-by-step manner, Bai et al. (2008) presented the reformulation of optimal power flow (OPF) problem as an SDP model and applied a primal-dual interior point method for its solution. Despite the paper's richness in details, the SDP variable could have been formed more parsimoniously (i.e., with a smaller size) to cater for computational efficiency. Lavaei and Low (2012) provided a succinct SDP formulation for OPF and derived a sufficient condition for zero duality gap between the convex SDP solution and

the nonconvex ACOPF solution. The authors demonstrated the effectiveness of the formulation by globally solving several standard power systems test cases. However, a subsequent work by Lesieutre et al. (2011) showed that Lavaei and Low's SDP formulation can fail to give a physically meaningful solution (i.e., it has a non-zero duality gap) in some scenarios of practical interest. The authors went on to investigate an SDP approach utilizing modified objective and constraints to compute all solutions of the nonlinear power flow equations. Molzahn et al. (2013) extended Lavaei and Low's SDP formulation to incorporate cases with multiple generators at the same bus and with multiple lines between two buses, enabling a more general model of the power system. More recent work has extended these results somewhat to different settings.

Problem Description and the Benders' Framework

The general formulation of SCED with post-contingency corrective rescheduling (with K contingencies) is written as follows (Liu et al., 2014; Monticelli et al., 1987a; Capitanescu and Wehenkel, 2008), Pinto and Stott:

$$\begin{aligned} \min_{x_0,\ldots,x_K,y_0,\ldots,y_K} \quad & f_0(x_0, y_0) \\ \text{s.t.} \quad & g_k(x_k, y_k) = 0 \quad k = 0, \ldots, K \\ & h_k(x_k, y_k) \leqslant 0 \quad k = 0, \ldots, K \\ & |y_k - y_0| \leqslant \Delta_k \quad k = 1, \ldots, K \end{aligned} \tag{7.1}$$

where f_0 is the base-case objective function and h_k and g_k are constraint functions. For the k-th system configuration, x_k is the vector of state variables including the (real and imaginary part of) voltage V^{re} and V^{im} and power flow F^{re} and F^{im}, y_k is the vector of control variables including the real and reactive power injection P and Q. Δ_k is the vector of maximal allowed variation of control variables, specifically the ramping radius of generation, between the base case ($k = 0$) and the k-th post-contingency configuration.

Two characteristics of the formulation are important to note. First, contingency

related variables, x_k and y_k for $k \neq 0$, do not play a role in the objective function, indicating that each contingency is essentially a feasibility problem, specifically an AC power flow problem. Second, different contingencies are only linked to the base case but are not directly linked to each other, which makes the overall problem decomposable and thus inviting to parallel computing techniques.

In this work, we develop an SDP model that specifically deals with the feasibility subproblem arising from the Benders' decomposition framework for the SCED problem. Our approach features a high degree of scalability thanks to recent advances in decomposition algorithms and parallel computing techniques. Benders' decomposition method, as well as many of its variants, has served as an effective tool to ameliorate scale-related computational difficulties, see, e.g., Monticelli et al. (1987a) and Pinto and Stott.

The idea of Benders' algorithm is to break a very large-scale model into a master model and many similarly structured sub-models (subproblems), all small enough to be solved efficiently and amenable to be processed in parallel. In SCOPF, where contingency scenarios only pose security constraints but do not alter the objective (which is minimizing the base-case dispatch cost), the subproblem is typically formulated as a feasibility problem, i.e., to test whether a master solution (base-case dispatch) can ensure a feasible operating point in the contingency. If not, the subproblem solution is used to provide a constraint that cuts off the given master solution in future iterations. Our main contribution resides in developing a convex subproblem that copes with the AC feasibility requirement, whereas several computational techniques are directly inherited from Chapter 6, which include (1) maintaining a dynamic list of active contingencies in order to defer low-impact computational tasks to when they are really necessary, and (2) leveraging a parallel computing framework within GAMS to boost computational efficiency.

7.2 Benders' Decomposition with SDP Subproblems

SDP Formulation of the AC Feasibility Subproblem

Let Y be the n-by-n bus admittance matrix. The net injection of apparent power at bus i is

$$\begin{aligned} S_i = \quad & P_i + iQ_i = V_i I_i^* = V_i (YV)_i^* \\ = \quad & (V_i^{re} + iV_i^{im})[\sum_j (Y_{i,j}^{re} + iY_{i,j}^{im})(V_j^{re} + iV_j^{im})]^* \end{aligned}$$

Hence, the real and reactive power injections at bus i are

$$P_i = \sum_j (V_i^{re} V_j^{re} Y_{i,j}^{re} - V_i^{re} V_j^{im} Y_{i,j}^{im} + V_i^{im} V_j^{im} Y_{i,j}^{re} + V_i^{im} V_j^{re} Y_{i,j}^{im}) \tag{7.2}$$

$$Q_i = \sum_j (V_i^{im} V_j^{re} Y_{i,j}^{re} - V_i^{im} V_j^{im} Y_{i,j}^{im} - V_i^{re} V_j^{im} Y_{i,j}^{re} - V_i^{re} V_j^{re} Y_{i,j}^{im}) \tag{7.3}$$

The apparent power flow on line (i, j) measured at bus i is

$$\begin{aligned} F_{i,j} = \quad & F_{i,j}^{re} + iF_{i,j}^{im} = V_i (Y_{i,j} V_j)^* \\ = \quad & (V_i^{re} + iV_i^{im})[(Y_{i,j}^{re} + iY_{i,j}^{im})(V_j^{re} + iV_j^{im})]^* \end{aligned}$$

Hence, the real and reactive power flows along line (i, j) are given as

$$F_{i,j}^{re} = V_i^{re} V_j^{re} Y_{i,j}^{re} - V_i^{re} V_j^{im} Y_{i,j}^{im} + V_i^{im} V_j^{re} Y_{i,j}^{im} + V_i^{im} V_j^{im} Y_{i,j}^{re} \tag{7.4}$$

$$F_{i,j}^{im} = V_i^{im} V_j^{re} Y_{i,j}^{re} - V_i^{im} V_j^{im} Y_{i,j}^{im} - V_i^{re} V_j^{re} Y_{i,j}^{im} - V_i^{re} V_j^{im} Y_{i,j}^{re} \tag{7.5}$$

The AC feasibility problem is formulated as follows.

ACF:

$$\min_{P,Q,V,F,s} \quad s^2 \tag{7.6}$$

$$\text{s.t.} \quad \sum_{g\in\mathcal{G}_i} \underline{G}_g^{\text{real}} - D_i^{\text{real}} \leqslant P_i \leqslant \sum_{g\in\mathcal{G}_i} \bar{G}_g^{\text{real}} - D_i^{\text{real}} \qquad \forall i \in \text{BUS} \tag{7.7}$$

$$\sum_{g\in\mathcal{G}_i} \underline{G}_g^{\text{imag}} - D_i^{\text{imag}} \leqslant Q_i \leqslant \sum_{g\in\mathcal{G}_i} \bar{G}_g^{\text{imag}} - D_i^{\text{imag}} \qquad \forall i \in \text{BUS} \tag{7.8}$$

$$-\bar{F}_{i,j} \leqslant F_{i,j}^{\text{real}} \leqslant \bar{F}_{i,j} \qquad \forall (i,j) \in \text{LINE} \tag{7.9}$$

$$(\underline{V}_i)^2 \leqslant (V_i^{\text{real}})^2 + (V_i^{\text{imag}})^2 \leqslant (\bar{V}_i)^2 \qquad \forall i \in \text{BUS} \tag{7.10}$$

$$\sum_{g\in\mathcal{G}_i} (G_g^0 - \Delta_g) - s^2 \leqslant P_i \leqslant \sum_{g\in\mathcal{G}_i} (G_g^0 + \Delta_g) + s^2 \qquad \forall i \in \text{BUS} \tag{7.11}$$

$$\text{and } (7.2), (7.3), (7.4). \tag{7.12}$$

A note on symbols and notation: $\mathcal{B}$ denotes the set of buses and $\mathcal{L}$ denotes the set of lines. $\mathcal{G}_i$ is the set of generators attached to bus i, symbols G and D represent the generation and demand parameters, parameter G^0 is the base-case real power generation which is passed in from the master solution, Δ represents generators' ramping radii (5-minute response time is used in this chapter) and s models the violation in the ramping constraint (7.11).

Equations (7.2), (7.3) and (7.4) can be substituted into (7.7), (7.8) and (7.9) to eliminate variables P, Q and F^{real} from the model. All equations in the reduced model, including the objective function and each constraint, are quadratic functions of the variables V and s. An equivalent SDP formulation can then be derived. To do so, we arrange the scalar variables in a vector x of size $2n+1$,

$$x = [V_1^{\text{re}} \ \dots \ V_n^{\text{re}} \ V_1^{\text{im}} \ \dots \ V_n^{\text{im}} \ s]$$

then define a matrix W by

$$W = x^T x = \begin{bmatrix} (V_1^{re})^2 & \cdots & V_1^{re}V_n^{re} & V_1^{re}V_1^{im} & \cdots & V_1^{re}V_n^{im} & V_1^{re}s \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ V_n^{re}V_1^{re} & \cdots & (V_n^{re})^2 & V_n^{re}V_1^{im} & \cdots & V_n^{re}V_n^{im} & V_n^{re}s \\ V_1^{im}V_1^{re} & \cdots & V_1^{im}V_n^{re} & (V_1^{im})^2 & \cdots & V_1^{im}V_n^{im} & V_1^{im}s \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ V_n^{im}V_1^{re} & \cdots & V_n^{im}V_n^{re} & V_n^{im}V_1^{im} & \cdots & (V_n^{im})^2 & V_n^{im}s \\ sV_1^{re} & \cdots & sV_n^{re} & sV_1^{im} & \cdots & sV_n^{im} & s^2 \end{bmatrix}$$

W is positive semidefinite and will serve as the variable in the SDP formulation. To construct an SDP formulation, we need to replace each of the quadratic equations in the ACF model with an SDP-type of equation and ultimately form a model of the following form:

ACF-SDP:

$$\begin{aligned} \underset{W \succeq 0}{\text{Min}} \quad & A_0 \bullet W \\ \text{s.t.} \quad & \sum_{g \in \mathcal{G}_i} \underline{G}_g^{real} - D_i^{real} \leqslant A_{1i} \bullet W \leqslant \sum_{g \in \mathcal{G}_i} \bar{G}_g^{real} - D_i^{real} && \forall i \in BUS \\ & \sum_{g \in \mathcal{G}_i} \underline{G}_g^{imag} - D_i^{imag} \leqslant A_{2i} \bullet W \leqslant \sum_{g \in \mathcal{G}_i} \bar{G}_g^{imag} - D_i^{imag} && \forall i \in BUS \\ & -\bar{F}_{i,j} \leqslant A_{3ij} \bullet W \leqslant \bar{F}_{i,j} && \forall (i,j) \in LINE \\ & (\underline{V}_i)^2 \leqslant A_{4i} \bullet W \leqslant (\bar{V}_i)^2 && \forall i \in BUS \\ & \sum_{g \in \mathcal{G}_i} (G_g^0 - \Delta_g) \leqslant A_{5i} \bullet W \leqslant \sum_{g \in \mathcal{G}_i} (G_g^0 + \Delta_g) && \forall i \in BUS \end{aligned}$$

where A_0, A_{1i}, A_{2i}, A_{4i}, A_{5i} for each $i \in \mathcal{B}$ and A_{3ij} for each $(i,j) \in \mathcal{L}$ are all matrices of the same size as W. The operator $\bullet$ stands for the entry-wise product, e.g., $A \bullet B = \sum_{i,j} A_{i,j}B_{i,j}$ for matrices $A, B \in \mathbb{R}^{m \times n}$. Note that $A \bullet B = \text{Tr}(A^T B) = \text{Tr}(B^T A)$, where $\text{Tr}(\cdot)$ is the matrix trace operator. ACF-SDP is a relaxation of the

ACF model, since the requirement that W must be a rank 1 matrix is dropped.

The remaining task in completing the SDP relaxation is to determine the parameter matrices A so that the calculation results match those in ACF. For example, $A_0 \bullet W$ should equal to s^2 and $A_{1i} \bullet W$ should equal to the right-hand side of (7.2), and so forth. Detailed composition of these A matrices is omitted here but readers could find a similar exercise in Bai et al. (2008).

Benders' Cut Generated by the SDP Subproblem

ACF-SDP is a convex optimization problem and so is its dual. Let the scalar quantities u_{1i}, u_{2i}, u_{4i}, u_{5i} for each $i \in \mathcal{B}$ and u_{3ij} for each $(i,j) \in \mathcal{L}$ be the dual variables corresponding to the upper bound inequalities in (7.7) to (7.11) and let v_{1i}, v_{2i}, v_{4i}, v_{5i} for each $i \in \mathcal{B}$ and v_{3ij} for each $(i,j) \in \mathcal{L}$ be the dual variables corresponding to the lower bound inequalities in (7.7) to (7.11). The dual of ACF-SDP is then given below.

ACF-SDP-Dual:

$$\begin{aligned}
\max_{u,v} \quad & \sum_{i\in \text{BUS}} [v_{1i}(\sum_{g\in\mathcal{G}_i} \underline{G}_g^{\text{real}} - D_i^{\text{real}}) + u_{1i}(\sum_{g\in\mathcal{G}_i} \bar{G}_g^{\text{real}} - D_i^{\text{real}}) \\
& \qquad + v_{2i}(\sum_{g\in\mathcal{G}_i} \underline{G}_g^{\text{imag}} - D_i^{\text{imag}}) + u_{2i}(\sum_{g\in\mathcal{G}_i} \bar{G}_g^{\text{imag}} - D_i^{\text{imag}}) \\
& \qquad + v_{4i}(\underline{V}_i)^2 + u_{4i}(\bar{V}_i)^2 + v_{5i}\sum_{g\in\mathcal{G}_i}(G_g^0 - \Delta_g) + u_{5i}\sum_{g\in\mathcal{G}_i}(G_g^0 + \Delta_g)] \\
& + \sum_{(i,j)\in\text{LINE}} (-v_{4ij}\bar{F}_{i,j} + u_{4ij}\bar{F}_{i,j}) \qquad (7.13) \\
\text{s.t.} \quad & \sum_{i\in\text{BUS}} [(u_{1i}+v_{1i})A_{1i} + (u_{2i}+v_{2i})A_{2i} + (u_{4i}+v_{4i})A_{4i} + (u_{5i}+v_{5i})A_{5i}] \\
& + \sum_{(i,j)\in\text{LINE}} (u_{4ij}+v_{4ij})A_{4ij} \preceq A_0 \qquad (7.14) \\
& u_{1i}, u_{2i}, u_{3i}, u_{5i} \leqslant 0, \ \forall i \in \text{BUS} \qquad (7.15) \\
& v_{1i}, v_{2i}, v_{3i}, v_{5i} \geqslant 0, \ \forall i \in \text{BUS} \qquad (7.16) \\
& u_{4ij} \leqslant 0, v_{4ij} \geqslant 0, \ \forall (i,j) \in \text{LINE} \qquad (7.17)
\end{aligned}$$

Using the convex ACF-SDP as the feasibility subproblem, subgradient inequalities (i.e., Benders' cuts) can be derived. Specifically, let $\nu(G^0)$ be the optimal value of ACF-SDP-Dual given the real power generation G^0 of the base-case. For a particular base-case solution G^{0*}, let $\nu^* = \nu(G^{0*})$ and let ν^*_{5i} and u^*_{5i}, for $i \in \mathcal{B}$, be the values of ν^*_{5i} and u^*_{5i} at the optimal solution of ACF-SDP-Dual(G^{0*}). For convenience, let us condense the objective function (7.13) as

$$\max_{u,\nu} \sum_{i\in\mathcal{B}} \sum_{g\in\mathcal{G}_i} (\nu_{5i} + u_{5i}) G^0_g + C \tag{7.18}$$

where C captures all extra terms needed to equate (7.18) with (7.13). Then, we have

$$\begin{aligned}
\nu(G^0) = \quad & \max_{u,\nu} \sum_{i\in BUS} \sum_{g\in\mathcal{G}_i} (\nu_{5i} + u_{5i}) G^0_g + C \geqslant \sum_{i\in BUS} \sum_{g\in\mathcal{G}_i} (\nu^*_{5i} + u^*_{5i}) G^0_g + C \\
\nu(G^{0*}) = \quad & \sum_{i\in BUS} \sum_{g\in\mathcal{G}_i} (\nu^*_{5i} + u^*_{5i}) G^{0*}_g + C
\end{aligned}$$

Combining the above two lines, we have the typically-called subgradient inequality:

$$\nu(G^0) \geqslant \nu(G^{0*}) + \sum_{i\in\mathcal{B}} \sum_{g\in\mathcal{G}_i} (\nu^*_{5i} + u^*_{5i})(G^0_g - G^{0*}_g)$$

In order to achieve $\nu(G^0) \leqslant 0$, it is necessary to enforce,

$$\nu(G^{0*}) + \sum_{i\in\mathcal{B}} \sum_{g\in\mathcal{G}_i} (\nu^*_{5i} + u^*_{5i})(G^0_g - G^{0*}_g) \leqslant 0 \tag{7.19}$$

The linear inequality (7.19) of G^0 is the Benders' cut to be used in the master model, in which G^0 is a decision variable.

AC Feasibility of the Base Case

By default the base case, i.e., g_0 and h_0 in (7.1), is modeled using linearized DC equations in the master problem, the rationale being that the base case represents a planning problem for which a linear approximation is acceptable. Nevertheless,

it would be more realistic if the base-case dispatch was also performed on the AC network representation. From the modeling perspective, it is straightforward to swap the LP master model with an NLP (full ACOPF) model without affecting the validity of the Benders' cuts. An apparent issue with this is the difficulty in global optimization of a nonconvex NLP, in a market context where a global optimum is essential.

We propose a convex approach to shepherd the base-case solution toward AC feasibility. Specifically, we treat the base-case network constraints as a special contingency (call it contingency zero) and construct an SDP subproblem for it. The subproblem is treated as one of the many subproblems to be solved in parallel, hence its introduction adds little computational cost. Furthermore, since the network constraints are now handled by the subproblem, we can remove them from the master model. In the end, the bare master model only consists of a convex (in our case, linear) objective function (of base-case injection P) and bound constraints on P; all other constraints come from cuts.

7.3 Numerical Experiments

Since contingency response is an operation-level action aimed at achieving a feasible AC operating point, we compare the SCED model with SDP subproblems and the SCED model with DC-based linear subproblems (see the previous chapter for its formulation) by evaluating their solutions in the full AC context. Given the base-case solution from a SCED model, we run an AC power flow model to identify a feasible post-contingency AC operating point for each contingency case covered by the SCED model. If an AC operating point is found for a contingency, it means that the SCED solution is indeed secure to this contingency; otherwise, the SCED solution is false secure hence not reliable.

We use the following model ACF-PS to find a feasible AC power flow from a base-case dispatch G^{0*}. For each inequality constraint in the AC power flow model, a pair of nonnegative artificial variables are introduced to allow for violation in the positive and negative directions, respectively. By minimizing the total violation,

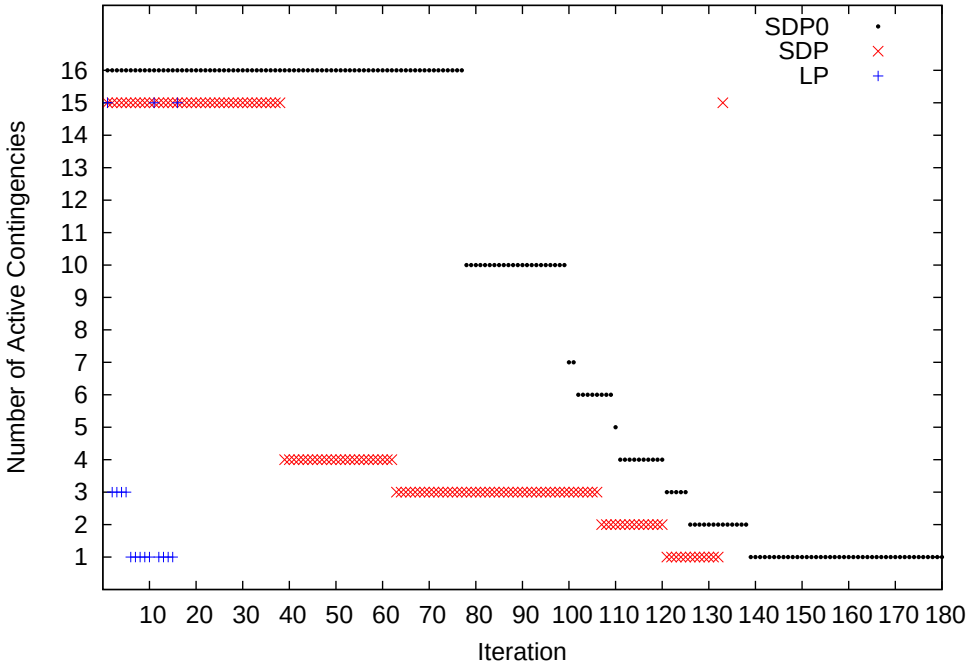


Figure 7.1: Benders' iterations of different models for the 118-bus case.

the model seeks a feasible AC power flow solution. Note that the set of lines and the line limits in this model are written as $\mathcal{L}'$ and $\bar{F}'$, respectively, to reflect the post-contingency setting, i.e., a line is removed from the original network topology and the limits of remaining lines are relaxed to the 5-minute short-term emergency (STE) rating.

The ACF-PS is a nonconvex model and obtaining a global solution can be time-consuming. We adopt a two-step procedure: first solve ACF-PS using the local solver CONOPT, if the optimal value is 0, a feasible AC solution is found; otherwise, if either the problem is infeasible or the optimal value is positive, then solve it again using the global solver GLOMIQO by taking the solution from CONOPT as the starting point. This method is more efficient than using GLOMIQO in all cases, as in most cases CONOPT can find a global solution (one with zero optimal value) in a few seconds. The computer used for experiments is an HP Z400 workstation with Intel Xeon W3520 CPU @2.67GHz and 8GB memory. All models and algorithms

are coded in GAMS (version 24.3.3 for Windows 64-bit) and the SDP problems are solved by the MOSEK solver within GAMS.

ACF-PS:

$$\begin{aligned}\min_{P,Q,V,F,s\cdot} \quad & \sum_{i\in\mathcal{B}}(s_i^{P+}+s_i^{P-}+s_i^{Q+}+s_i^{Q-}+s_i^{V+}+s_i^{V-} \\ & +s_i^{R+}+s_i^{R-})+\sum_{(i,j)\in\mathcal{L}'}(s_{i,j}^{F+}+s_{i,j}^{F-})\end{aligned}$$

Subject to

$$\sum_{g\in\mathcal{G}_i}\underline{G}_g^{re}-D_i^{re}\leqslant P_i-s_i^{P+}+s_i^{P-}\leqslant\sum_{g\in\mathcal{G}_i}\bar{G}_g^{re}-D_i^{re},\ \forall i$$

$$\sum_{g\in\mathcal{G}_i}\underline{G}_g^{im}-D_i^{im}\leqslant Q_i-s_i^{Q+}+s_i^{Q-}\leqslant\sum_{g\in\mathcal{G}_i}\bar{G}_g^{im}-D_i^{im},\ \forall i$$

$$(F_{i,j}^{re})^2-s_{i,j}^{F+}+s_{i,j}^{F-}\leqslant(\bar{F}'_{i,j})^2,\ \forall(i,j)\in\mathcal{L}'$$

$$(\underline{V}_i)^2\leqslant(V_i^{re})^2+(V_i^{im})^2-s_i^{V+}+s_i^{V-}\leqslant(\bar{V}_i)^2,\ \forall i$$

$$\sum_{g\in\mathcal{G}_i}(G_g^{0*}-\Delta_g)\leqslant P_i-s_i^{R+}+s_i^{R-}\leqslant\sum_{g\in\mathcal{G}_i}(G_g^{0*}+\Delta_g),\ \forall i$$

$$s^{P+},\ s^{P-},\ s^{Q+},\ s^{Q-},\ s^{V+},\ s^{V-},\ s^{F+},\ s^{F-},\ s^{R+},\ s^{R-}\geqslant 0$$

and (7.2), (7.3), (7.4).

Several IEEE test cases Christie (1993) are used in the experiments. To make numerical results informative for analysis, i.e., to avoid cases in which all contingencies are trivially feasible or trivially uncontrollable, we scaled up/down all nodal loads by some fixed factor. Specifically, for the 14-, 30-, 57- and 118-bus cases the loads are scaled by 2.1x, 1.5x, 0.5x and 1.5x, respectively. In each case, we compare three SCED models: (1) LP, model with a linear base case and linear contingency subproblems; (2) SDP, model with a linear base case and SDP contingency subproblems; (3) SDP0, model with both base case and contingency cases modeled as SDP subproblems.

The experimental results are listed in Table 7.1. The table is to be read as follows. The "Tabu" column lists the number of uncontrollable contingencies reported by the

Table 7.1: Solution Comparison of Three SCED Models

Case	Cont	Solution				Performance		
		Model	Tabu	Cost	Time	IF	FS	FT
14	20	LP	0	13253.3	4.2	12	12	0
		SDP	6	16065.8	18.4	6	0	0
		SDP0	6	16003.4	11.9	6	0	0
30	40	LP	0	582.0	4.0	1	1	0
		SDP	1	585.0	20.1	1	0	0
		SDP0	1	600.5	22.1	1	0	0
57	20	LP	0	12508.0	1.9	1	1	0
		SDP	1	12508.0	13.2	1	0	0
		SDP0	1	12560.0	50.9	1	0	0
118	15	LP	0	139716.8	54.0	16	16	0
		SDP	0	141372.2	2414.3	1	1	0
		SDP0	0	144220.1	11951.1	0	0	0

SCED model (with LP subproblem or SDP subproblem, identified by the "Model" column). A contingency is deemed uncontrollable if either one of the following situation has arisen in the SCED run: (1) its corresponding subproblem becomes infeasible; (2) it has contributed conflicting cuts that renders the master problem infeasible Liu et al. (2014). The "Cost" is the base-case generation cost and the "Time" marks total solution time in seconds. Once we obtain the base-case dispatch, we evaluate it by running ACF-PS for all contingencies one by one. The dispatch is marked infeasible for a given contingency if the corresponding ACF-PS is either infeasible or has a positive optimal value. The number of contingencies for which the base-case dispatch is InFeasible is presented in the "IF" column. FS (False Secure) represents the number of contingencies allegedly secured (i.e., not in Tabu) by SCED but turn out to be infeasible in ACF-PS, whereas FT (False Tabu) indicates the number of contingencies reported as uncontrollable but are feasible in ACF-PS.

We can see that SCED with SDP subproblems provides a more reliable solution, despite coming with a higher cost. In the 14-bus case, the Tabu set (including contingency 3, 13, 14, 15, 17 and 20) determined by the SCED is identical to the IF set resulted from ACF-PS, which means that all contingencies reported by the SCED as

uncontrollable are indeed uncontrollable and all contingencies reportedly secured by the SCED are indeed secure. This observation holds for all cases tested. In comparison, the model with LP subproblems does not always guarantee a feasible AC power flow in a contingency. For example, in the 14-bus case the model deemed all contingencies to be controllable, while 12 contingencies turn out to be AC infeasible (including the 6 identified by SDP as uncontrollable). Nonetheless, the extent of infeasibility in FS cases is not very pronounced, e.g., the ACF-PS objectives are in the order of 10^{-1} and 10^{-3} in the 14- and 118-bus cases, respectively. Since a problem cannot be feasible if its convex relaxation is infeasible, the SDP and SDP0 are by design immune to the false tabu error, as corroborated by the numerical results. Note that in the 118-bus case the solution from SDP model is not AC feasible for the base case while the solution from SDP0 is, which demonstrates the value of SDP0 over SDP. However, SDP0's robustness comes with a significantly higher computational cost. Figure 7.1 compares the number of Benders' iterations needed for convergence in different models. The plot is truncated at the 180th iteration as SDP0 actually took 1000+ iterations and most iterations beyond the 139th one are used to neutralize a single difficult contingency (the base case). This suggests that a more efficient treatment of difficult contingencies could significantly improve the solution speed.

7.4 Conclusion

We have developed a novel approach to solve the AC-based SCED problem. The main novelty resides in the use of semidefinite programming as a convex relaxation of the AC feasibility problem and the development of Benders' cut based on the SDP subproblem. Experiments have shown superior solution quality of the new approach over the LP-based approach. Further algorithmic improvements, in particular a better treatment of the base case and difficult contingency cases, are needed in order to solve large instances faster and more reliably.

8 CONCLUSION

New technologies are rapidly changing the way electric energy is generated, transmitted and consumed. These changes in turn drive upgrades in the policy framework and operational standards within the power industry. Backed by rigorous analyses, concrete examples and abundant numerical results, this dissertation has contributed original design ideas and solution methodology to several important issues within the contemporary wholesale electricity markets in the United States. In summary, the dissertation has provided answers to the following questions.

- What do optimization modelers need to know (at a minimum) about the physics and maths that govern the operation of a transmission network? Chapter 1 has provided an overview of the power engineering basics, including a description of different formulations of the AC power flow problem, economic dispatch problem and unit commitment problem, an introduction of commonly used data formats and an industrial case study at ISO New England, Inc.

- What is the role of demand response in the electricity markets and how should system operators efficiently implement the FERC Order 745? The lack of demand-side participation hinders realization of the economic efficiency purported by the two-sided, competitive market design. Many policy-making initiatives have attempted to mend this issue and FERC Order 745 is one of them. Since its enactment in 2011, the Order has been widely criticized and challenged. One of the leading accusation against the Order is premised on equating the act of demand response to a unrightful sale of energy. To rationalize the Order and hence implement it in an economically efficient way, Chapter 2 has provided an alternative economic interpretation of DR: demand response can be treated as an organized trade of "consuming rights" among electricity consumers. Based on this interpretation, a compliant market model has been developed and a three-phase solution procedure involving the

joint use of nonlinear and mixed integer solvers as well as bound-tightening techniques has been devised.

- What are the flaws of the payment rule in the context of unit commitment and how can they be mended? In the current design of U.S. electricity spot markets, the generation dispatch mechanism and the payment rule are incompatible with each other. The uniform-price auction format predicated on a two-sided market design with marginal pricing is flawed, since the supply and the demand are not treated equitably and discrete decisions, such as unit commitment, are inevitable. Chapter 3 has proposed a pay-as-bid scheme as an better alternative and has demonstrated the merit of pay-as-bid via a bidding behavior model and simulation experiments.

- What are the limitations within the existing bidding structure and how should the limitations be relaxed in order to benefit the market? Existing bid formats are all separable over time. However, a significant and growing segment of demand can be shifted across time and therefore has no way to bid its true valuation of consumption. Chapter 4 has proposed additional bid types that allow deferrable, adjustable and storage-type loads to better express their value, and thus elicit demand response in the most natural way - via direct participation in the market. The additional bid types have been shown to be easily incorporated into the existing market structure with no technological barrier and able to substantially increase social welfare.

- How is the stochastic programming technique being used in an ISO's daily operations and what is its practical effectiveness? Chapter 5 has presented a stochastic unit commitment problem formulated for ISO New England's reserve adequacy analysis. Due to the large problem size and computational constraints, a stochastic RAA model cannot take as many scenarios as one would wish – in reality only a small number of scenarios will be incorporated into the stochastic model. This chapter has proposed a Derand method that makes informed guesses based on partitioning and properties of conditional

expectation and has demonstrated a substantial performance boost of RAA by this method.

- How could recent advancements in optimization and high-performance computing extend the capability of dispatch operations within a power system? Chapter 6 has devised and implemented a series of algorithmic enhancements based on the Benders' decomposition method. These enhancements have ameliorated the computational difficulty arising from a security-constrained economic dispatch model that, for an increased economic efficiency, considers multiple stages of rescheduling. In addition, Chapter 7 has proposed a novel approach based on semidefinite programming (SDP) to solve the model in the nonlinear AC setting. The key point is to approximate the nonconvex AC feasibility problem with its SDP relaxation and use these SDP models as a convex subproblem within a Benders' decomposition framework. Numerical experiments have demonstrated the superior solution quality of the approach and its tractability for IEEE test cases.

To fully realize the social benefits of open-access, competitive electricity markets, which now serve two thirds of U.S. electricity consumers, two prerequisites must be in place - a set of fair and efficient market rules and an adequate population of knowledgable individuals willing and able to participate in both sides of the supply and demand equation. Tremendous research has focused on the former, including my work in Chapter 2, 3 and 4, while the latter has received little. In particular, few in the general public know or have a venue to exercise their trading rights in the bulk power markets. This lack of visibility and direct participation from the masses has caused many economic and reliability issues. Examples include exploitative trading by large institutions, inefficient demand response and sluggish day-ahead preparation in the event of disastrous weather such as Hurricane Sandy.

A meaningful future work is to set up an education project that brings speculative electricity trading onto everyone's financial radar and to create a proxy program that facilitates the willing public to trade in the bulk power markets, through an aggregation and exchange engine of virtual offers and bids. A core

part of this program will be a market simulator that claws in real-time market data, synthesizes them with "client"-provided bid data, executes dispatch and pricing algorithms and renders authentic market results, all in real time. Comprehensive research of the policy, regulation and ISO-specific market rules is needed along with sophisticated modeling, architecting and implementation. This project will pave the way for the ongoing policy reform in the set direction. First, increased public participation will leave little room for inefficiency or loophole in market rules. Second, by financially engaging retail users in wholesale spot markets, the program will melt the regulatory barrier between the two physical markets and promote retail-level competition. Furthermore, the implementation exercise can serve as a test-bed for sprouting ideas in the field of large-scale optimization and big-data analytics.

REFERENCES

Anderson, Edward J., Pär Holmberg, and Andrew B. Philpott. 2013. *Mixed strategies in discriminatory divisible-good auctions*.

Andersson, Göran. 2008. *Modelling and analysis of electric power systems*. Eidgenössische Technische Hochschule Zürich. Lecture 227-0526-00, ITET ETH Zürich.

Arroyo, José Manuel, and Antonio J. Conejo. 2002. Multiperiod auction for a pool-based electricity market. *IEEE Transactions on Power Systems* 17(4):1225–1231.

Bai, Xiaoqing, Hua Wei, Katsuki Fujisawa, and Yong Wang. 2008. Semidefinite programming for optimal power flow problems. *International Journal of Electrical Power & Energy Systems* 30(6¨C7):383 – 392.

Bard, J. F. 1998. *Practical bilevel optimization: Algorithms and applications*, vol. 30 of *Nonconvex Optimization and its Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Boisvert, Richard N., Peter A. Cappers, and Bernie Neenan. 2002. The benefits of customer participation in wholesale electricity markets. *The Electricity Journal* 15(3):41 – 51.

Borghetti, A., G. Gross, and C.A. Nucci. 2002. Auctions with explicit demand-side bidding in competitive electricity markets. In *The next generation of electric power unit commitment models*, ed. BenjaminF. Hobbs, MichaelH. Rothkopf, RichardP. O'Neill, and Hung-po Chao, vol. 36 of *International Series in Operations Research & Management Science*, 53–74. Springer US.

Borlick, Robert. 2011. Paying for Demand-Side Response at the Wholesale Level: The Small Consumers' Perspective. *The Electricity Journal* 24:8–19.

Brandimarte, Paolo. 2014. *Handbook in monte carlo simulation: Applications in financial engineering, risk management and economics*. WILEY.

Bussieck, Michael R., Michael C. Ferris, and Timo Lohmann. *GUSS: Solving collections of data related models within GAMS*.

Cain, Mary B., Richard P. O'Neill, and Anya Castillo. 2012. History of optimal power flow and formulations. FERC staff paper.

Capitanescu, F., M. Glavic, D. Ernst, and L. Wehenkel. 2007a. Contingency filtering techniques for preventive security-constrained optimal power flow. *Power Systems, IEEE Transactions on* 22(4):1690–1697.

Capitanescu, F., J.L. Martinez Ramos, P. Panciatici, D. Kirschen, A. Marano Marcolini, L. Platbrood, and L. Wehenkel. 2011. State-of-the-art, challenges, and future trends in security constrained optimal power flow. *Electric Power Systems Research* 81(8):1731 – 1741.

Capitanescu, F., and L. Wehenkel. 2007. Improving the statement of the corrective security-constrained optimal power-flow problem. *Power Systems, IEEE Transactions on* 22(2):887–889.

———. 2008. A new iterative approach to the corrective security-constrained optimal power flow problem. *Power Systems, IEEE Transactions on* 23(4):1533–1541.

Capitanescu, Florin, Mevludin Glavic, Damien Ernst, and Louis Wehenkel. 2007b. Interior-point based algorithms for the solution of optimal power flow problems. *Electric Power Systems Research* 77(5):508 – 517.

Chao, Hung-po. 2010. *Demand management in restructured wholesale electricity markets*. ISO New England.

Chao, Hung-po. 2011. Demand response in wholesale electricity markets: the choice of customer baseline. *Journal of Regulatory Economics* 39:68–88.

Christie, Rich. 1993. Power systems test case archive. Http://www.ee.washington.edu/research/pstca.

Colson, Benoît, Patrice Marcotte, and Gilles Savard. 2007. An overview of bilevel optimization. *Annals of Operations Research* 153:235–256.

Conejo, Antonio J, Enrique Castillo, Raquel García-Bertrand, and Roberto Mínguez. 2006. *Decomposition techniques in mathematical programming: engineering and science applications*. Springer Berlin.

Constantinescu, EM, VM Zavala, M Rocklin, S Lee, and M Anitescu. 2009. Unit commitment with wind power generation: integrating wind forecast uncertainty and stochastic programming. Tech. Rep., Argonne National Laboratory (ANL). ANL/MCS-TM-309.

Cooper, H.J., G.C. Goodwin, A. Feuer, and M.G. Cea. 2012. Design of scenarios for constrained stochastic optimization via vector quantization. In *American control conference (acc), 2012*, 1865–1870.

Cramton, Peter, and Steven Stoft. 2007. Why we need to stick with uniform-price auctions in electricity markets. *The Electricity Journal* 20(1):26 – 37.

Daryanian, B., R.E. Bohn, and R.D. Tabors. 1989. Optimal demand-side response to electricity spot prices for storage-type customers. *IEEE Transactions on Power Systems* 4(3).

Dobson, Ian, Scott Greene, Rajesh Rajaraman, Christopher L. DeMarco, Fernando L. Alvarado, Mevludin Glavic, Jianfeng Zhang, and Ray Zimmerman. 2001. *Electric power transfer capability: Concepts, applications, sensitivity, uncertainty*. Power Systems Engineering Research Center, PSERC Publication 01-34. Http://www.pserc.cornell.edu/tcc/info.html.

Dommel, Hermann W., and William F. Tinney. 1968. Optimal power flow solutions. *IEEE Transactions on Power Apparatus and Systems* PAS-87(10).

Durrett, Rick. 2010. *Probability: Theory and examples*. 4th ed. Cambridge University Press.

ISO New England, Inc. *ISO New England operating procedures 19 – transmission operations*.

———. 2012. Market Rule 1 Appendix E. Http://www.iso-ne.com/regulatory/tariff/sect_3/.

Federico, Giulio, and David M. Rahman. 2001. Bidding in an electricity pay-as-bid auction. *Journal of Regulatory Economics* (2):175–211.

Feng, Y., and S. M. Ryan. 2014. Solution sensitivity-based scenario reduction for stochastic unit commitment. *Computational Management Science* 1–34. DOI: 10.1007/s10287-014-0220-z.

FERC. 2008. *Wholesale competition in regions with organized electric markets*. Federal Energy Regulatory Commission. Http://www.ferc.gov/whats-new/comm-meet/2008/101608/E-1.pdf.

———. 2011. *Demand response compensation in organized wholesale energy markets*. Http://www.ferc.gov/EventCalendar/Files/20110315105757-RM10-17-000.pdf.

Fernández-Blanco, Ricardo, José Manuel Arroyo, and Natalia Alguacil. 2012. A unified bilevel programming framework for price-based market clearing under marginal pricing. *Power Systems, IEEE Transactions on* 27(1):517 –525.

Ferris, Michael C. 2011. *Coupled optimization models for planning and operation of power systems on multiple scales*.

Ferris, Michael C., Rishabh Jain, and Steven Dirkse. 2011. *Gdxmrw: Interfacing gams and matlab*. Http://www.gams.com/dd/docs/tools/gdxmrw.pdf.

Fischetti, Matteo, Domenico Salvagnin, and Arrigo Zanette. 2010. A note on the selection of benders' cuts. *Mathematical Programming* 124(1-2):175–182.

Fliscounakis, S., P. Panciatici, F. Capitanescu, and L. Wehenkel. 2013. Contingency ranking with respect to overloads in very large power systems taking into account uncertainty, preventive, and corrective actions. *Power Systems, IEEE Transactions on* 28(4):4909–4917.

Galbraith, J.K. 1980. *American capitalism: The concept of countervailing power*. Classics In Economics Series, Transaction Pub.

GAMS. Scenred2 user manual. Http://www.gams.com/dd/docs/solvers/scenred2.pdf.

Gersho, Allen, and Robert M Gray. 1992. *Vector quantization and signal compression*. Springer.

Goodwin, Graham C, Jan Østergaard, Daniel E Quevedo, and Arie Feuer. 2009. A vector quantization approach to scenario generation for stochastic NMPC. In *Nonlinear model predictive control*, 235–248. Springer.

Heitsch, Holger, and Werner Römisch. 2003. Scenario reduction algorithms in stochastic programming. *Computational optimization and applications* 24(2-3):187–206.

Hogan, William W. 2009. *Providing incentives for efficient demand response*. PJM Demand Response. FERC Docket EL09-68-000.

———. 2010a. *Demand response pricing in organized wholesale markets*. ISO/RTO Council. FERC Docket RM10-17-000.

———. 2012. *Economists' brief on ferc order 745 regarding demand response compensation*. District Court of Columbia Circuit.

J. Himelic, F. Novachek. 2011. *Sodium sulfur battery energy storage and its potential to enable further integration of wind (wind-to-battery project)*. Xcel Energy.

Jacobs, Jonathan M. 1997. Artificial Power Markets and Unintended Consequences. *IEEE Transactions on Power Systems* 12:968 – 972.

Jiang, Q., and K. Xu. 2013. A novel iterative contingency filtering approach to corrective security-constrained optimal power flow. *Power Systems, IEEE Transactions on* PP(99):1–11.

Johnson, Raymond B., Shmuel S. Oren, and Alva J. Svoboda. 1997. Equity and efficiency of unit commitment in competitive electricity markets. *Utilities Policy* 6(1):9 – 19.

Joskow, Paul L. 2001. California's electricity crisis. *Oxford Review of Economic Policy*.

Kahn, Alfred E., Peter C. Cramton, Robert H. Porter, and Richard D. Tabors. 2001. Uniform pricing or pay-as-bid pricing: A dilemma for california and beyond. *The Electricity Journal* 14(6):70 – 79.

Kirschen, Daniel S. 2003. Demand-side view of electricity markets. *IEEE Transactions on Power Systems* 18(2):520–527.

Kleywegt, Anton J., and Alexander Shapiro. 2001. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 502.

Krall, Eric, Michael Higgins, and Richard P. O'Neill. 2012. *RTO unit commitment test system*. Federal Energy Regulatory Commission.

Lavaei, J., and S.H. Low. 2012. Zero duality gap in optimal power flow problem. *Power Systems, IEEE Transactions on* 27(1):92–107.

Lesieutre, B.C., D.K. Molzahn, A.R. Borden, and C.L. DeMarco. 2011. Examining the limits of the application of semidefinite programming to power flow problems. In *Communication, control, and computing (allerton), 2011 49th annual allerton conference on*, 1492–1499.

Li, Fangxing, and Rui Bo. 2007. DCOPF-based LMP simulation: Algorithm, comparison with ACOPF, and sensitivity. *Power Systems, IEEE Transactions on* 22(4): 1475–1485.

Li, Yuan, and J.D. McCalley. 2009. Decomposed SCOPF for improving efficiency. *Power Systems, IEEE Transactions on* 24(1):494–495.

Liu, Yanchao, Michael C. Ferris, and Feng Zhao. 2014. Computational study of security constrained economic dispatch with multi-stage rescheduling. *Power Systems, IEEE Transactions on* PP(99):1 – 10.

Lloyd, S. 1982. Least squares quantization in PCM. *Information Theory, IEEE Transactions on* 28(2):129–137.

Lubin, Miles, Cosmin G. Petra, and Mihai Anitescu. 2011a. The parallel solution of dense saddle-point linear systems arising in stochastic programming. *Optimization Methods and Software*.

Lubin, Miles, Cosmin G. Petra, Mihai Anitescu, and Victor M. Zavala. 2011b. Scalable stochastic optimization of complex energy systems. In *Sc'11 proc. of 2011 international conference for high performance computing, networking, storage and analysis*. ACM, Seattle, Washington: ACM.

Mankiw, N. Gregory. 2011. *Principles of Economics*. Sixth ed. South-Western College Pub.

Molzahn, Daniel. 2013. *Estimating line-flow limits*. Private communication.

Molzahn, D.K., J.T. Holzer, B.C. Lesieutre, and C.L. DeMarco. 2013. Implementation of a large-scale optimal power flow solver based on semidefinite programming. *Power Systems, IEEE Transactions on* 28(4):3987–3998.

Momoh, James A., M. E. El-Hawary, and Ramababu Adapa. 1999. A review of selected optimal power ftow literature to 1993. *IEEE Transactions on Power Systems* 14(1).

Monticelli, A., M. V F Pereira, and S. Granville. 1987a. Security-constrained optimal power flow with post-contingency corrective rescheduling. *Power Systems, IEEE Transactions on* 2(1):175–180.

Monticelli, A., M.V.F. Pereira, and S. Granville. 1987b. Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Transactions on Power Systems* PWRS-2(1).

Newell, Sam, and Attila Hajos. 2010. *Demand response in the midwest iso - an evaluation of wholesale market design*. The Brattle Group.

O'Neill, Richard P., Paul M. Sotkiewicz, Benjamin F. Hobbs, Michael H. Rothkopf, and William R. Stewart. 2005. Efficient market-clearing prices in markets with nonconvexities. *European Journal of Operational Research* 164(1):269–285.

Papadaskalopoulos, D., P. Mancarella, and G. Strbac. 2011. Decentralized, agent-mediated participation of flexible thermal loads in electricity markets. In *Intelligent system application to power systems (ISAP), 2011 16th international conference on*, 1–6.

Papavasiliou, A., and S. Oren. 2013. Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network. *Operations Research* 61(3):578–592.

Phan, Dzung, and J. Kalagnanam. 2014. Some efficient optimization methods for solving the security-constrained optimal power flow problem. *Power Systems, IEEE Transactions on* 29(2):863–872.

Pinto, Herminio, and Brian Stott. *Security constrained economic dispatch with post-contingency corrective rescheduling*. FERC Conference, June 23-24, 2010, Washington DC.

Ruff, Larry E. 2002. *Economic principles of demand response in electricity*. Edison Electric Institute,Washington D. C.

Schweppe, Fred C., Michael C. Caramanis, Richard D. Tabors, and Roger E. Bohn. 1988. *Spot pricing of electricity*. Kluwer international series in engineering and computer science: Power electronics & power systems, Kluwer Academic.

Shapiro, Alexander, Darinka Dentcheva, and Andrzej Ruszczynski. 2009. *Lectures on stochastic programming: Modeling and theory*. SIAM.

Spees, Kathleen, and Lester B. Lave. 2007. Demand response and electricity market efficiency. *The Electricity Journal* 20:69–85.

Stigler, George J. 1954. The economist plays with blocs. *The American Economic Review* 44(2):pp. 7–14.

Stoft, S. 2002. *Power system economics: Designing markets for electricity*. IEEE Press, IEEE Press.

Stott, B., J. Jardim, and O. Alsac. 2009. Dc power flow revisited. *Power Systems, IEEE Transactions on* 24(3):1290–1300.

Strbac, G., and D. Kirschen. 1999. Assessing the competitiveness of demand-side bidding. *Power Systems, IEEE Transactions on* 14(1):120–125.

Su, Chua-Liang, and Daniel Kirschen. 2009. Quantifying the effect of demand response on electricity markets. *IEEE Transactions on Power Systems* 24:1199 – 1207.

Walawalkar, Rahul, Jay Apt, and Rick Mancini. 2007. Economics of electric energy storage for energy arbitrage and regulation in New York. *Energy Policy* 35(4):2558 – 2568.

Wang, Hongye, Carlos E. Murillo-Sánchez, Ray D. Zimmerman, and Robert J. Thomas. 2007. On computational issues of market-based optimal power flow. *IEEE Transactions on Power Systems* 22(3):1185–1193.

Wang, Qianfan, Jianhui Wang, and Yongpei Guan. 2013. Stochastic unit commitment with uncertain demand response. *Power Systems, IEEE Transactions on* 28(1): 562–563.

Wellinghoff, Hon. Jon, and David L. Morenoff. 2007. Recognizing the importance of demand response: The second half of the wholesale electric market equation. *Energy Law Journal* 28(2).

Wright, Stephen. 2009. *Sparse optimization methods*. Presentation slides.

Yu, Yaowen, P.B. Luh, E. Litvinov, Tongxin Zheng, Feng Zhao, and Jinye Zhao. 2013. Markov-based stochastic unit commitment considering wind power forecasts. In *Power and energy society general meeting (PES), 2013 IEEE*, 1–5.

Zimmerman, R. D., C. E. Murillo-Sánchez, and R. J. Thomas. 2011. Matpower: Steady-state operations, planning and analysis tools for power systems research and education. *IEEE Transactions on Power Systems* 26(1).