OPTIMIZATION AND EQUILIBRIUM METHODS IN POWER SYSTEMS

by

Jesse T. Holzer

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

Doctor of Philosophy
(Mathematics)

at the
UNIVERSITY OF WISCONSIN–MADISON
2014

Date of final oral examination: 8/27/2014

The dissertation is approved by the following members of the Final Oral Committee:

Michael C. Ferris, Professor, Computer Sciences
Alexander J. Nagel, Professor, Mathematics
Jordan S. Ellenberg, Professor, Mathematics
Thomas F. Rutherford, Professor, Agricultural and Applied Economics
Bernard C. Lesieutre, Associate Professor, Electrical and Computer Engineering

For Regina. Let’s go on a ’dventure.

ACKNOWLEDGMENTS

Many people and institutions have have helped me immensely in my work as a graduate student. The Departments of Mathematics and Computer Sciences at the University of Wisconsin-Madison supported me generously through TA and Student Lecturer positions. The Air Force Office of Scientific Research and the Department of Energy supported me through research grants. It is plain reality that no one can spend a decade in graduate school without financial support. I sincerely hope to give back to the larger community that has made a remarkable and risky investment in me.

I have felt good will and encouragement from so many people. The one thing that so many people gave so constantly more than any other thing was incredible patience while I brought myself around to getting this thesis done. For this I thank especially Regina, Michael, Mom, Papa, Maria, Sai, Ord, Steve, Andrew.

Michael was stoic through the most inexplicable delays on my part. He really went beyond what any advisor should be asked. When it was time for a meeting and I had nothing much to talk about, his sheer enthusiasm for optimization took over. I have learned so much from him, and I know I have much more to learn.

Troy told me to join the club of people who have PhD's, and it was just what I needed to hear when I didn't know if I could find a job, much less write a thesis. I'm grateful to Troy and Peabody for the opportunity to do academic work on an industrial problem.

Best friends the Olsons, the Spooner-Harveys, the Crafty Bitches, the Seahorses, the German students, the Akerliefs, Roddingtons, Boggrigottas and Guenettes made this university and this city of Madison a home.

Colleagues Lisa, Yanchao, Youngdae, Aditya, Charlie, Drew, Jagdish, Eric and Krishna shared late nights of work, became friends to me and then uncles and auntie to Malcolm.

Dear children Malcolm and Michaela and so many others, turned my life upside down and then showed me what it's really all about.

And I cannot say enough how thankful I am for Regina's incredible support. She gave so much of her time and emotional energy, so much that it really was a sacrifice. She talked with me and listened and waited and waited She helped me see that this thesis was possible when I thought it wasn't, and that I really did want to finish it. I could not and would not have done it without her. I would have given up and always regretted it.

And I am forever thankful, to the universe, for the chances I have had and amazingly still have.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Efficient operation and planning of electric power systems promise huge gains to society, from reduction of global poverty to avoidance of drastic climate change. There is a great need for mathematical and economic modeling to support these efficiency efforts. This thesis explores optimization methods to directly improve planning and operation and equilibrium methods to understand how markets composed of multiple economic agents carry out these optimal plans or in some cases thwart them. The challenges faced by this thesis are rooted not only in economic behavior and engineered systems but also in mathematical complexity.

We characterize the mode of divergence of agent-decomposable iterative algorithms for equilibrium problems. These problems pose a challenge for agent-decomposable algorithms because the applications where equilibrium seems to be most necessary for a faithful model have strong interactions between agents. We introduce an equilibrium model for understanding the effect of risk aversion in investment in power grid components.

We develop a novel global optimization technique for a nonconvex problem arising from coal mine quality planning. Our technique relies on isolating the nonconvexity to a low dimensional structure,

which is then approximated by a discrete grid. The low dimensionality keeps the computational cost manageable.

We model the effect of unit commitment on the behavior of strategic power suppliers under various market rules by computing discrete approximations of mixed Nash equilibria in continuous spaces. We are able consider rules such as a single price, a price with an uplift, and pay-as-bid.

We also document our contributions to collaborative work on semidefinite programming relaxations of nonconvex power flow problems and on

Efficient operation and planning of electric power systems promise huge gains to society, from reduction of global poverty to avoidance of drastic climate change. There is a great need for mathematical and economic modeling to support these efficiency efforts. This thesis explores optimization methods to directly improve planning and operation and equilibrium methods to understand how markets composed of multiple economic agents carry out these optimal plans or in some cases thwart them. The challenges faced by this thesis are rooted not only in economic behavior and engineered systems but also in mathematical complexity. We develop a novel global optimization technique for a nonconvex problem arising from coal mine quality planning. We model the effect of unit commitment on the behavior of strategic power suppliers under various market rules by computing discrete approximations of mixed Nash equilibria in continuous spaces. We characterize the mode of divergence of agent-decomposable iterative algorithms for equilibrium problems. We introduce an equilibrium model for understanding the effect of risk aversion in investment in power grid components. We also document our contributions to collaborative work on semidefinite programming relaxations of nonconvex power flow problems and and on the social benefit of expanded bidding structures in wholesale power markets.

Chapter 1

Optimization and equilibrium, convexity and monotonicity

The research documented in this thesis began with a question about mathematical problems with a certain structure. Could we use this structure to design algorithms to solve these problems more efficiently than standard algorithms that do not use such structure? We found that the structure we had in mind was no help without a certain more basic assumption that did not apply to the problems that we posed as interesting examples of the structure. And indeed our examples were challenging to standard algorithms as well. So our research, and this thesis, came to focus on ways of solving the problems at hand, and the structural assumption fell away.

The problems that we set out to solve are equilibrium, complementarity, and variational inequality problems, all generalizations of the basic optimization problem of minimizing an objective function over a set of feasible points. In an equilibrium problem there are a number of agents, each solving their own optimization problem by a choice of some strategy. The twist is that the strategy choice of one agent affects the objective value of the other agents. To solve the problem, we must find a choice of strategy for each agent so that no agent can do better by deviating from their assigned strategy. Complementarity problems and their generalization to variational inequalities arise as mathematical characterizations of solutions to optimization and equilibrium problems based only on local knowledge of the objectives and feasible sets.

The structure that we focus on is essentially that of equilibrium itself, with multiple agents each optimizing over their own set of variables, potentially influencing the outcomes of the other agents. As a structural assumption in the class of equilibrium problems, this is not very specific. But in a variational inequality or a complementarity problem, the structure that we identify is the mapping from variables to agents, and the objective functions of the agents. In a general

variational inequality this structure might not exist, or the information that it represents might have been discarded in the process of going from an equilibrium problem to a variational inequality characterizing its solutions. So the problem of our thesis might be said to be variational inequalities arising from equilibrium problems.

The basic idea of an algorithm using this structure is already obvious, and indeed without further explanation of the rather abstract concept of variational inequalities, it might be the only possible algorithm we can imagine. In this algorithmic idea, we begin with any choice of strategies for all the agents, not necessarily an equilibrium. We focus on one of the optimizing agents at a time. Holding fixed the variables belonging to the other agents, we update the variable of this one agent by solving his or her problem. Then we move on to another agent and repeat this updating process. We continue in this way until the strategies of all the agents appear to converge, and then we have found an equilibrium.

There are variations on this idea of single-agent optimization. For example, we may update all the agents simultaneously, rather than one at a time, but the algorithm behaves similarly, converging rather slowly, if at all. Or else rather than stepping in the direction of the optimal reaction we might step in the direction of the negative gradient of each agent. All these methods rely on an assumption that the agent optimization problems are convex and that the resulting variational inequality is monotone. Neither assumption is enough on its own, and without them we may see these methods fail to converge. And monotonicity cannot be ensured by reference to the individual optimization problems of the different agents, so it is not easy to design equilibrium models to have this property. In Chapter 2 we consider some small illustrative examples of this failure to converge without the right assumptions. We then consider a large-scale example, drawn from a model of investement in the electric power grid, of the failure of agent-based methods to converge.

In chapters 3, 4, and 5 we describe research we have done that overcomes the problems of nonconvexity in optimization and nonmonotonicity and nonexistence of equilibrium. Chapter 3 describes a nonconvex optimization problem that we encountered in collaboration with industry. We handled nonconvexity by enumerating and evaluating a dense sample of feasible points and

modeling the problem using integer programming to restrict the solution to lie on the curve containing these points. In a general context this approach is intractable, but we identify a feature of the problem that allows us to use this technique in just a low dimensional setting, keeping the number of sample points low and the model manageable. The resulting algorithm is compared to standard approaches and is shown to outperform them, either in solution quality or computational time.

In chapter 4 we review the theory of noncooperative games, or equilibrium problems, including mixed strategy equilibrium, games with continuous strategy spaces, and games where some players may face discontinuous or nonconvex optimization problems. Mixed equilibrium is a generalization of equilibrium that was introduced to ensure existence of at least some kind of solution to equilibrium problems that actually do not have a solution. But with continuous strategy spaces and nonconvex agent problems, the concept of mixed equilibrium can help us find a solution to an equilibrium problem that may have a pure equilibrium that we could not find by an agent-based method because of either nonconvexity or nonmonotonicity. In this context the crucial technique for both computation and proof of existence is discretization of the strategy space, and this will be familiar from our use of sampling to handle nonconvexity in optimization.

In Chapter 5 we describe an application of mixed equilibrium to a game model of a wholesale electric power market. The goal is to analyze several different market structures and rules, to see which leads to the lowest cost of generation to meet demand. We use a game model because we want to account for the fact that suppliers set their prices and production strategically, not reflecting only their own cost of production but also accounting for how the market will react to their choices. This strategic behavior, as opposed to competitive behavior, tends to increase the total cost of generation, and some market rules bring about more strategic behavior than others. We must use mixed equilibrium because the suppliers' profit functions turn out to be nonconcave and discontinuous when the unit commitment decision of which power plants to turn on is included in the analysis. It is important to include unit commitment partly because some of the market rules under consideration were designed specifically to account for it. Other researchers have considered only some of these market rules, or have neglected unit commitment, or have neglected strategic

behavior. Our contribution is to systematically include all of these features. Mixed equilibrium in a discrete approximation of a continuous strategy space is the correct tool for this task.

Chapter 6 contains a published paper [35] documenting joint work with Dan Molzahn, Chris DeMarco and Bernie Lesieutre. The paper reports on a decomposition scheme for solving semidefinite programming (SDP) relaxations of nonconvex quadratically constrained programming (QCP) formulations of the AC optimal power flow problem. We have long been interested in techniques of convex relaxation of nonconvex optimization problems, so we were able to provide general perspective on how this decomposition scheme fits into the overall literature on optimization. In particular, in our investigation of the StratPlan problem described in chapter 3, we considered SDP relaxation of a formulation using quadratic constraints. As a result we were also able to provide modeling and formulation techniques using semidefiniteness constraints. Furthermore we provided guidance on choosing numerical tolerances for convergence in the SDP solvers that were used and for tests of the rank of the matrices returned as solutions of the SDP relaxation.

We also built early versions of several parts of the Matlab code used to conduct the numerical experiments in the paper, including a code to formulate the SDP relaxation for a given power flow instance, a code testing different SDP solvers on the relaxation, and an implementation of Prim's algorithm to find a minimum spanning tree in a graph representing the sparsity pattern of the SDP relaxation.

Chapter 7 contains a paper documenting joint work with Yanchao Liu and Michael Ferris. The paper presents an optimization modeling framework for evaluating the social benefit of expanded bidding structures in wholesale power markets. We provided the standard econonic result on integrability of a partial equilibrium model showing that the natural description of the electric power market as an interaction among a number of optimizing agents is equivalent to the optimization problem of allocating power production and consumption among the agents so as to maximize the net social surplus.

we contributed the point that participants whose bid structure does not allow truthful bidding must bid falsely. For example if a participant has a downward sloping demand curve (as is typical) but only vertical fixed-quantity bids are allowed, then that participant must guess what the price p

will be and bid a quantity q so that the outcome (q, p) will be on its demand curve. The error in this guess adds up to a net social loss that can be eliminated by allowing sloping bids.

We provided the interpretation of convex cost or benefit functions as equivalent to monotone supply or demand functions. In this interpretation convex bid structures can be viewed as requiring that participants choose a supply or demand function from the a parametrized class of monotone functions, so no participant is able to bid exactly truthfully. But just some very expansions of allowable bid structures all quite close approximation to any monotone function, because in general monotone functions are well approximated by polyhedral monotone functions with few faces.

Chapter 2

Multiple Optimization Problems with Equilibrium Constraints

2.1 Introduction

In this chapter we formally introduce a class of mathematical programming problems that we and a growing number of researchers call multiple optimization problems with equlibrium constraints (MOPEC). We give the standard formulation of a MOPEC as a variational inequality. We identify the characteristics of the desired type of algorithm for MOPEC. We identify certain prototypical algorithms for variational inequalities and one for MOPEC specifically. And we show, by both numerical and theoretical means, that none of these prototypical algorithms has all the desired characteristics. Each of these algorithms either applies only to MOPEC problems satisfying the restrictive condition of monotonicity or else is not agent-decomposable. This suggests that an algorithm with all the desired characteristics may not exist, despite resonable conjectures that it should.

In the next section we describe a MOPEC model of electric power grid investment. We show how this model satisfies the restrictive condition of monotonicity and is thus able to be solved by a prototypical agent-decomposable algorithm. However as soon as we try to make the model a bit more interesting by adding a representation of risk-aversion this algorithm fails to converge. The divergence is quite similar to what is observed in one of the nonmonotone examples in section 2.2. We are still able to solve the model with an algorithm that is not agent-decomposable.

2.2 MOPEC: Theory, algorithms, and small examples

2.2.1 Problem definition

Suppose that a number of agents all make decisions simultaneously, and each agent's decision affects not only that agent but also all the others. How can we predict the decisions of all the agents? More precisely, what condition must the decisions of all agents collectively satisfy in a reasonable outcome? One standard condition is that the decision of each agent should be optimal for that agent under the assumption that the decisions of all the other agents are fixed. This condition is known as Nash equilibrium and MOPEC is a generalization of this concept.

Suppose agent i faces an optimization problem

$$\begin{array}{ll} \min_{x_i \in X_i} & f_i(x) \\ \text{s.t.} & g_i(x) \leq 0 \end{array} \tag{2.1}$$

Here $x_i \in X_i$ is the decision variable of agent i, and the objective and constraints of agent i are parametrized by the decision variables x_{-i} of the other agents. An equilibrium is a point $x \in X = \Pi_i X_i$ describing the decisions of all agents so that for all i, x_i is optimal for (2.1).

Introducing multipliers y_i in the negative orthant Y_i on the constraints of agent i, we may write the first order optimality conditions of (2.1) as a variational inequality

$$0 \in H_i(z) + N_{Z_i}(z_i) \tag{2.2}$$

Here $z_i = (x_i, y_i) \in Z_i = X_i \times Y_i$ and $z \in Z = \Pi_i Z_i$ and

$$H_i(z) = (F_i(z), G_i(z)) = (\nabla_{x_i} f_i(x) - d_{x_i} g_i(x)^T y_i, g_i(x)) \tag{2.3}$$

and (2.2) is a variational inequality in z_i parametrized by z_{-i}. Under an appropriate constraint qualification, (2.2) is a necessary condition for optimality in (2.1), and if X_i, f_i and all components of g_i are convex then it is sufficient. Henceforth we assume the appropriate convexity so that our equilibrium problem can be reformulated as the variational inequality

$$0 \in H(z) + N_Z(z) \tag{2.4}$$

And when we consider variational inequalities in general we also now assume that Z is closed, convex and nonempty.

In many applications the agents' optimization problems depend further on some parameter p that is in turn determined by the agents' decisions x_i. To accomodate this situation we assume that p is related to (x, y) by some other equilibrium problem

$$0 \in W(x, y, p) + N_P(p) \tag{2.5}$$

in p and parametrized by (x, y). And we generalize the optimization problem of agent i to

$$\begin{array}{ll} \min_{x_i \in X_i} & f_i(x, p) \\ \text{s.t.} & g_i(x, p) \leq 0 \end{array} \tag{2.6}$$

The problem (2.6, 2.5) is called a *multiple optimization problem with equilibrium constraints* (MOPEC) and is formulated as a variational inequality (2.4) with $z = \Pi_i(x_i, y_i) \times p \in Z = \Pi_i(X_i \times Y_i) \times P$ and

$$H(z) = \Pi_i(F_i(z), G_i(z)) \times W(z) \tag{2.7}$$

where

$$F_i(z) = \nabla_{x_i} f_i(x, p) - d_{x_i} g_i(x, p)^T y_i \tag{2.8}$$

and

$$G_i(z) = g_i(x, p) \tag{2.9}$$

2.2.2 Existence

Our major concern is with the convergence of algorithms for (2.4). But when we test our algorithms on a problem we must be sure that a solution exists. The existence theorem that we will use for this purpose holds even if Z is not convex:

Theorem 2.1 If Z is compact and nonempty, and H is continuous then (2.4) has a solution.

Though we do not need it, we record for completeness the other broadly applicable existence theorem for variational inequalities:

Theorem 2.2 If Z is closed, convex and nonempty, and H is strongly monotone then (2.4) has a solution.

Theorems 2.1 and 2.2 can both be generalized considerably, but they serve to illustrate the two broad categories of existence results for variational inequalities.

2.2.3 Properties of iterative solution methods

An iterative method for (2.4) begins with a point $z_0 \in \mathbb{R}^n$, and for each iterate z_k, a subproblem is solved to obtain the next iterate z_{k+1}. If each subproblem is easily solved then the algorithm may be less expensive than a direct approach to (2.4). And we are unaware of any direct methods for nonlinear variational inequalities.

An iterative algorithm for MOPEC is said to be *agent-decomposable* if each subproblem can be solved by solving one smaller subproblem for each agent in only the variables belonging to that agent. Agent-decomposability is a desirable property as makes the subproblem less expensive. First it allows the subproblem to be solved in a distributed fashion by processing different single-agent subproblems in parallel. Second, even if the single-agent subproblems are solved in sequence, if the full subproblem is too large for the available memory but the single-agent subproblems are small enough, then agent decomposability enables us to solve an otherwise unsolvable problem.

An agent-decomposable iterative algorithm for MOPEC is *optimization-preserving* if each single-agent subproblem corresponding to an optimization agent is itself an optimization problem. This property is desirable as both theory and algorithms are somewhat more advanced for optimization than they are for variational inequalities.

And the most important property that an iterative algorithm may have is convergence. This is really two properties. First, the iterates z_k should converge to a point z^*, and second, z^* should be a solution of (2.4). Of course convergence is dependent also on the problem at hand. Our goal in this research was to find an agent-decomposable optimization-preserving iterative method and a significant class of MOPEC problems treatable by this method.

2.2.4 Five iterative solution methods

We consider five iterative solution methods for variational inequalities, of which three apply in general, one applies with polyhedral Z, and one applies to MOPEC specifically.

We begin with the algorithm of iterated optimal reaction, which we define in the case of MOPEC. Given an iterate $z = (x, y, p)$, evaluate an optimal reaction $z' = (x', y', p')$ satisfying

$$\begin{aligned} 0 &\in F_i(x_i', y_i', x_{-i}, y_{-i}, p) + N_{X_i}(x_i') \\ 0 &\in G_i(x_i', y_i', x_{-i}, y_{-i}, p) + N_{Y_i}(y_i') \end{aligned} \tag{2.10}$$

for all i and

$$0 \in W(x, y, p') + N_P(p') \tag{2.11}$$

Then define the next iterate $\hat{z}$ by taking a step in the direction of z' with step length multiplier $\tau > 0$:

$$\hat{z} = z + \tau(z' - z) \tag{2.12}$$

This is the most obvious candidate for the type of algorithm we are interested in. It is clearly agent-decomposable and optimization preserving. We will see by numerical investigation that the method of iterated optimal reaction has similar convergence behavior to the next method, which is more easily studied analytically.

The projected gradient method is defined by the recursion

$$z_{k+1} = \Pi_Z(z_k - \tau H(z_k)) \tag{2.13}$$

where $\Pi_Z : \mathbb{R}^n \to \mathbb{Z}$ is the metric projection onto Z and $\tau > 0$ is a step multiplier. It is agent-decomposable, and although it is not optimization preserving, it is arguably even better, as the single-agent subproblem requires only the evaluation of ∇f, ∇g and g, and this is less expensive than solving a single-agent optimization problem. Projected gradient has a very narrow convergence guarantee:

Theorem 2.3 Suppose H is strongly monotone and Lipschitz continuous. Then there exists $\tau > 0$ so that the projected gradient method converges to a solution.

The extragradient method is defined by

$$\begin{aligned} z'_k &= P_Z(z_k - \tau H(z_k)) \\ z_{k+1} &= P_Z(z_k - \tau H(z'_k)) \end{aligned} \tag{2.14}$$

with step multiplier $\tau > 0$. This method is agent-decomposable and better than optimization-preserving, like the projected gradient method. It is essentially twice as expensive as projected gradient, but has a slightly broader convergence guarantee:

Theorem 2.4 Suppose H is monotone and Lipschitz. Then there exists $\tau > 0$ so that the extragradient method converges to a solution.

A well-known example where extragradient converges but projected gradient does not is

Example 2.5

$$0 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.15}$$

The proximal point method is defined implicitly by

$$z_k \in z_{k+1} + \sigma H(z_{k+1}) + N_Z(z_{k+1}) \tag{2.16}$$

where $\sigma > 0$ is a regularization parameter. This method is not agent-decomposable, and its convergence theory is not much better than that of the extragradient method:

Theorem 2.6 Suppose H is monotone. Then for all $\sigma > 0$ the proximal point method converges to a solution.

It does converge more rapidly than the decomposable methods, but here we are concerned with whether a method converges or not.

If Z is polyhedral, then we may linearize H to obtain an affine variational inequality. The essential idea of the PATH solver is to take this affine problem as a subproblem:

$$0 \in dH(z_k)(z_{k+1} - z_k) + N_Z(z_k) \tag{2.17}$$

PATH solves (2.17) by a pivoting algorithm and uses a merit function to damp the iterates in order to promote global convergence. It approximates the interaction between agents more accurately than the agent decomposable methods of projected gradient and extragradient and optimal reaction, which essentially ignore this interation. This greater accuracy yields more robust convergence but comes at a computational cost, ultimately because the interactions among all n agents number about n^2.

The convergence results for projected gradient and extragradient can be strengthened in several ways. Of particular relevance for us, the convergence guarantee holds for any step multiplier less than τ. In practice a good step multiplier may not be known in advance, so a sequence of step multipliers τ_k, slowly decreasing to 0, may be used instead. E.g. $\tau_k = 1/k$ ensures that eventually the multipliers will be small enough to ensure convergence, and the fact that $\sum_k \tau_k = \infty$ ensures that the limit point still is a solution. In our computational experiments this is the step multiplier sequence that we use.

A complete discussion of these algorithms requires some mention of the termination conditions. Generally an iterative algorithm would include a bound on the iterates z_k that signals divergence if it is exceeded, a tolerance on $z_{k+1} - z_k$ signalling convergence, and some resource and computation limits. Also commonly used is a merit function for $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ such that $\Phi(z) = 0$ if and only if z is a solution of (2.4). When $\Phi(z_k)$ is below a predetermined tolerance the algorithm stops and declares that z_k solves (2.4) adequately. In principle this is different from convergence, but for well-behaved problems these concepts are equivalent. One such merit function is given by

$$\Phi(z) = \|z - \Pi_Z(z - H(z))\| \tag{2.18}$$

and this is the merit function that we use in our numerical experiments.

2.2.5 Exploring convergence in a nontrivial MOPEC class

In this section we look for a nontrivial class of MOPEC problems in which an agent-decomposable optimiztion-preserving iterative method converges to a solution if a solution exists. Essentially we find that there is no such class. For this we perform numerical experiments using the methods of

projected gradient, extragradient, and optimal reacction on a class of MOPEC problems that seems highly favorable to any algorithm. We find counterexamples to convergence of each method. We also perform numerical experiments to compare the convergence behaviors of the methods of optimal reaction and projected gradient, and we observe that they converge or diverge together.

The convergence results we have so far for agent-decomposable algorithms all require some degree of monotonicity. But it is hard to build a relevant MOPEC model that is monotone. Monotonicity requires essentially that the influence of each agent's variables on its objective be greater than those of all the other agents combined. But in all the applications we are aware of where the MOPEC structure *per se* is necessary for a faithful model of the phenomenon of interest, the other agents are highly influential. This is not to say that such an application is definitely not monotone, just that the only easy method of verifying monotonicity is not applicable.

There is an extreme case where a variational inequality is easily seen to be monotone, and this naturally applies to the variational inequality formulation of a MOPEC.

Definition 2.7 An optimization problem $\mathcal{P}$ is said to be an integral of a variational inequality $\mathcal{V}$ if $\mathcal{V}$ is the first-order optimality conditions of $\mathcal{P}$.

This definition can be generalized to allow for permutation of variables and multipliers, or even further, to allow for rotations in the space $\mathbb{R}^n$ of variables and multipliers. We have:

Theorem 2.8 If the variational inequality (2.4) has a convex integral, then H is monotone.

If an integral can be found for a model that is naturally a MOPEC, then this may provide insight into the meaning of the model, but we gain nothing by treating it as a MOPEC for computation, as an optimization solver can do just as well.

So what can we require of a MOPEC short of monotonicity to promote convergence of an agent-decomposable optimization-preserving iterative method without monotonicity? We prefer requirements that are easy to verify in particular MOPEC models. And for the purpose of computer coding and ese of understanding we prefer very simple MOPEC problems. And of course we must ensure existence of a solution.

Definition 2.9 We say that a MOPEC satisfies the *strong assumptions* if

- there is no equilbrium agent p, and there are no contraints g_i, so that we may identify z_i with x_i;
- each X_i is the compact interval $[-1, 1] \subset \mathbb{R}$;
- each objective function is a homogeneous quadratic $f_i(x) = 0.5x^T Q_i x$; and
- the objective of each agent is strongly convex with respect to the decision variables of that agent, i.e. the submatrix $Q_{ix_ix_i}$ is a positive scalar.

These strong assumptions meet all our requirements, ensuring in particular that 0 is a solution, and appear to be as favorable to convergence of iterative algorithms as possible, short of requiring monotonicity.

It has been asserted that an agent-decomposable iterative method should converge even without monotonicity if the agents' objectives are convex and their decision sets are compact and convex. We now formulate a precise conjecture representing this assertion. Then we give a simple counterexample to this conjecture.

Conjecture 2.10 Under the strong assumptions, with iteration-dependent step multiplier $\tau_k = 1/k$, from any starting point z_0, the extragradient method given by

$$\begin{aligned} z'_k &= P_Z(z_k - \tau_k H(z_k)) \\ z_{k+1} &= P_Z(z_k - \tau_k H(z'_k)) \end{aligned} \tag{2.19}$$

satisfies $\Phi(z_k) \to 0$.

We may make a similar conjecture for the projected gradient method and the iterated reaction method, and our counterexample applies for these too, but we emphasize the extragradient method as it has the strongest theoretical guarantee of convergence.

Here is a counterexample:

Example 2.11 Three agents $i = 1, 2, 3$. Agent i solves

$$\min_{x_i} 0.5(x_i + 3x_{i+1})^2 \tag{2.20}$$

where the index $i + 1$ is interpreted modulo 3.

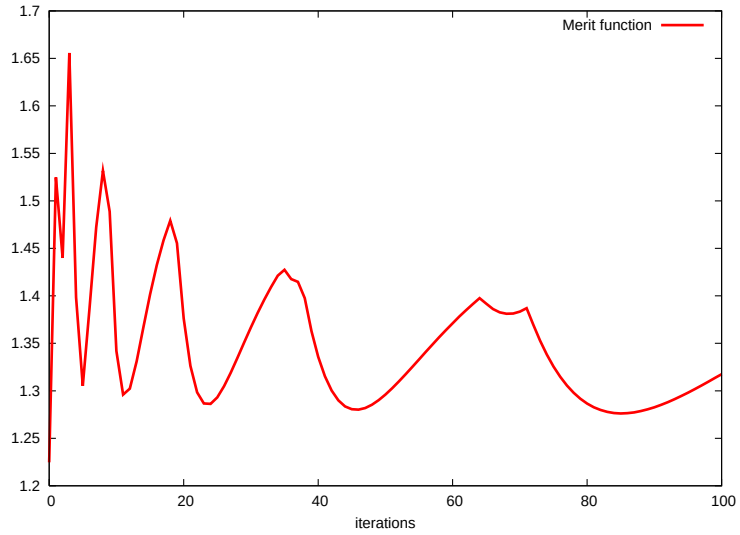

Figure 2.1 The iterated reaction method applied to problem (2.20)

We ran the three prototypical agent-decomposable methods on this example, and the value of $\Phi(z_k)$ is plotted against k in figures (2.1), (2.2), (2.3). All three methods show no sign of convergence.

Here is an example where extragradient converges, but the other two methods do not:

Example 2.12 Three agents $i = 1, 2, 3$. Agent i solves

$$\min_{x_i} 0.5(x_i + 2x_{i+1})^2 \tag{2.21}$$

The projected gradient method, shown in figure (2.5), and the iterated reaction method, shown in figure (2.4), fail to converge to a solution, while the extragradient method, shown in figure (2.6), does converge to a solution. This example mimics example 2.5, which takes place in $\mathbb{R}^2$ and is integrable but does not satisfy the stronconvexity requirement of the strong assumptions. In fact there is no such example in $\mathbb{R}^2$ satisfying the strong assumptions:

Theorem 2.13 Under the strong assumptions, suppose there are two optimizing agents. Then the projected gradient method and the extragradient method both converge to a solution from any starting point.

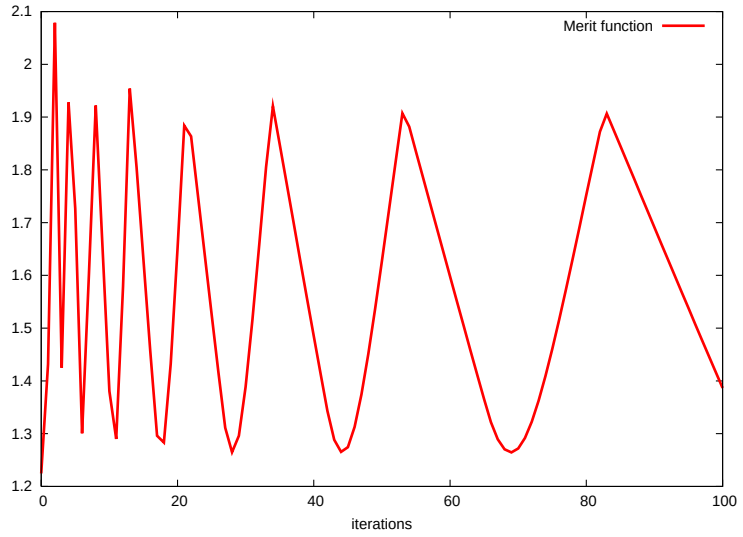

Figure 2.2 The projected gradient method applied to problem (2.20)

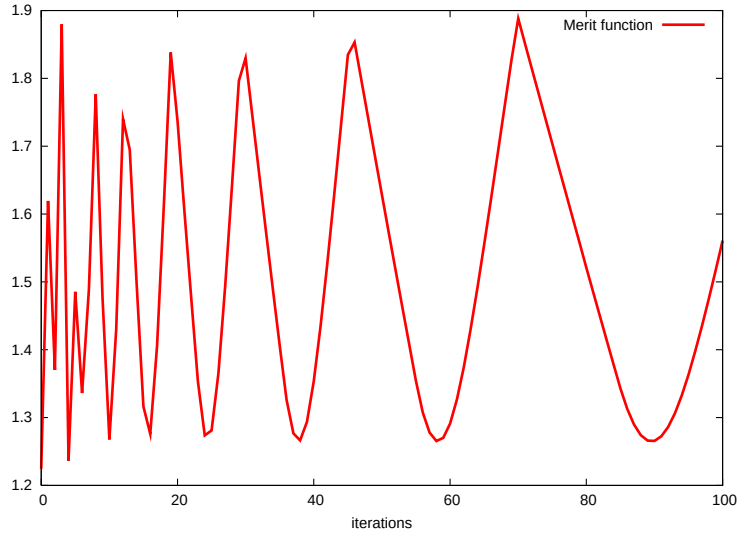

Figure 2.3 The extragradient method applied to problem (2.20)

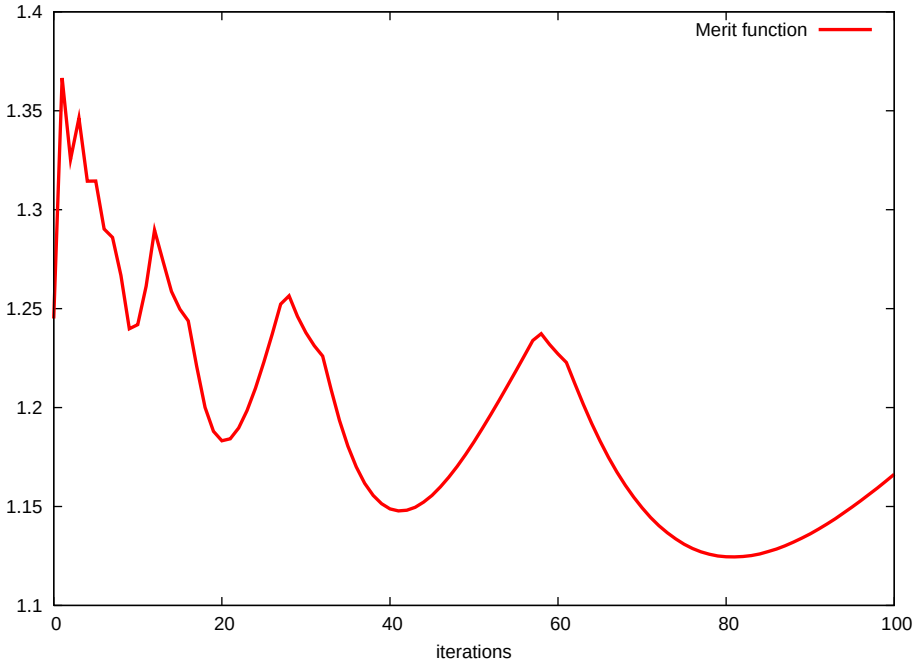

Figure 2.4 The iterated reaction method applied to problem (2.21)

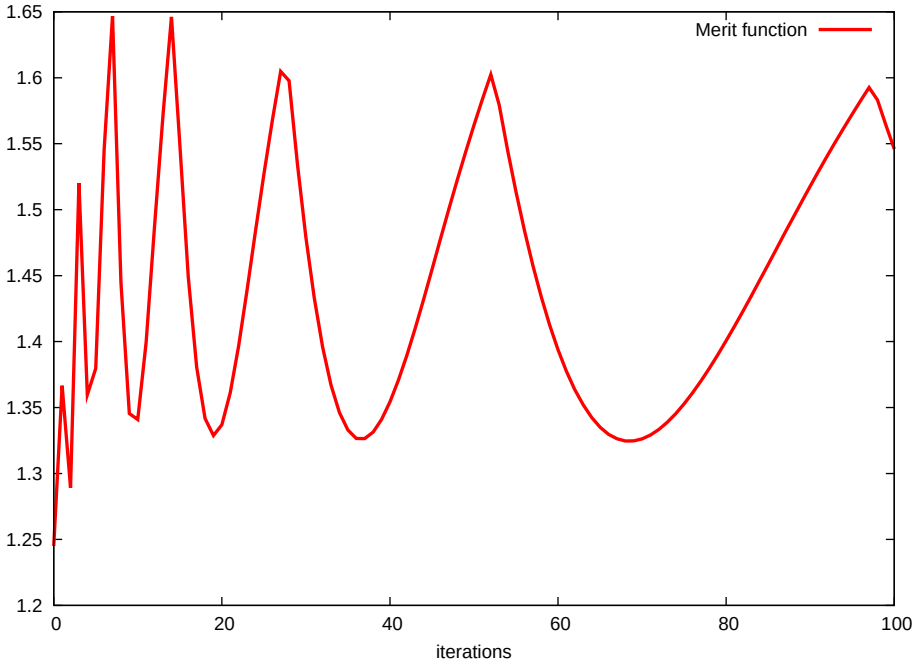

Figure 2.5 The projected gradient method applied to problem (2.21)

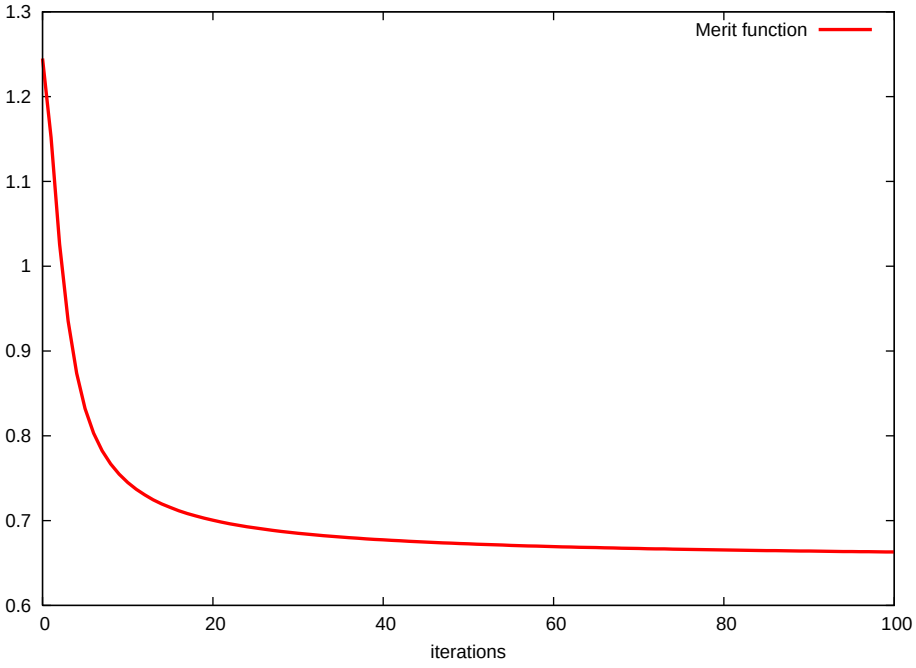

Figure 2.6 The extragradient method applied to problem (2.21)

Proof: We prove convergence for the projected gradient method. The assumptions guarantee that the MOPEC is of the form $0 \in Az + N_{[-1,1]^2}(z)$ where

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{2.22}$$

with $a, d > 0$. The intuitive idea of the proof is that the iterates z_k follow the explicit Euler method for the ordinary differential equation

$$z' = Az \tag{2.23}$$

as long as z_k remains in the interior of the domain $[-1, 1]^2$. The origin is an equilibrium of (2.23), so we refer to the classification of phase portraits of linear ODEs in $\mathbb{R}^2$. In all possible phase portraits except for a source (case 1), a source with rotation (case 2), and a cycle (case 3), convergence is clear, without any need for the strong assumptions. The eigenvalues of A are negative in case (1), are complex conjugates with negative real parts in case (2) and are pure imaginary numbers summing to 0 in case (3). In each case

$$\lambda_1 + \lambda_2 \leq 0 \tag{2.24}$$

But since $a, d > 0$, the eigenvalues satisfy

$$\lambda_1 + \lambda_2 = \operatorname{tr}(A) = a + d > 0 \tag{2.25}$$

□

In view of Theorem 2.13 it is easy to see one reason why Conjecture 2.10 and more general versions of it might be intuitively appealing. Limited mainly to 2 dimensions in our imagination, we are unable to easily visualize anything like Example 2.20. And the proof of Theorem 2.13 reveals exactly how we might have expected the compact domain Z to ensure convergence despite nonmonotonicity, In particular, when A has a negative eigenvalue, the boundary of Z blocks the iterates z_k from diverging along the vector field $-H(z)$. But this topological condition is unique to 2-dimensional space.

Finally to compare the convergence properties of the projected gradient method and the optimal reaction method, we test the methods on a set of 10 randomly generated MOPEC problems. We

ensure that all the problems satisfy the strong assumptions as otherwise an equilibrium may not exist or the variational inequality formulation may not be equivalent to the MOPEC itself. The problems all have three optimizing agents, as we know the convergence behavior analytically from theorem 2.13. The quadratic coefficients Q are uniformly distributed on $[-3, 3]$ except for the diagonal of each Q_i, which is uniform on $[0, 1]$. Thus each agent's objective is strongly convex with probability 1. On 3 instances both methods failed and on 7 instances both methods succeeded. On the basis of this evidence and extensive practical experience with these methods, we conclude that these methods converge or diverge together. The results of one of the failure instances are shown in figures (2.8) and (2.7).

2.2.6 Conclusion

In this section we have indroduced the MOPEC problem, its variational inequality formulation, and basic existence results. We have characterized the type of algorithm we hoped to find in this research as iterative, agent-decomposable, and optimization-preserving. We have described the class of MOPEC problems that seemed likely to enable both relevant MOPEC modeling and solution by the methods of interest. And we gave numerical evidence suggesting that the appealing method of iterated optimal reaction is about as likely to converge on any given MOPEC problem as the simpler method of projected gradient. And we showed that even in the most favorable of circumstances these methods might very well fail on nonmonotone problems. In the next section we will see in a large scale MOPEC model of electric power grid investment with risk aversion the very same mode of nonconvergence that we identified in these minimal examples.

2.3 Nonconvergence: An example from a large scale model of power grid investment

2.3.1 Introduction

We introduce a model of investment in, and operation of, productive capacity for a single good. There are standard models that treat investment in risky assets and models of a portfolio of production technologies to meet demand for a good. Our model links these two standard models,

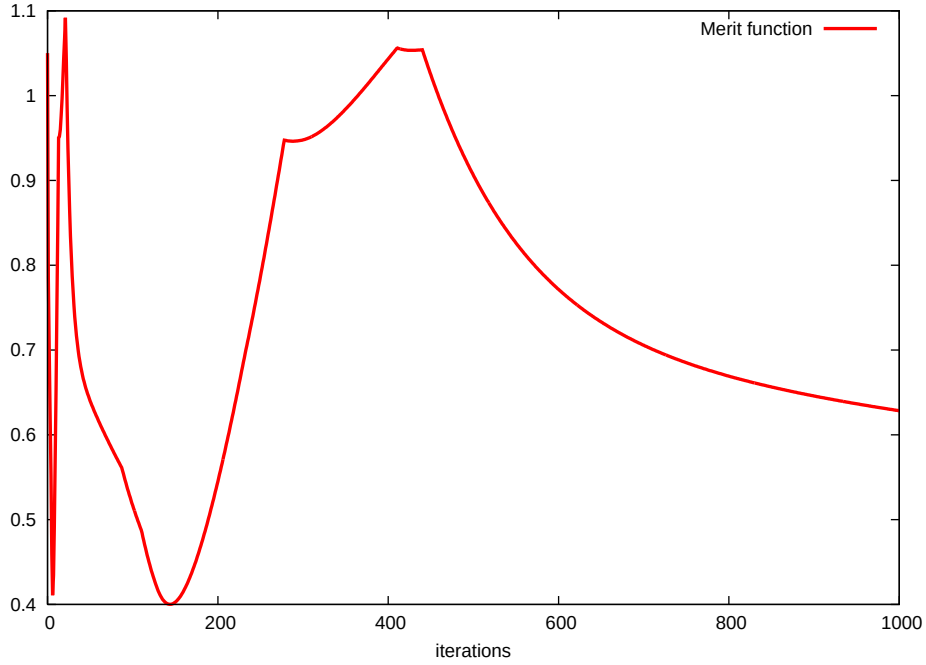

Figure 2.7 Failure of the iterated reaction method on a random problem

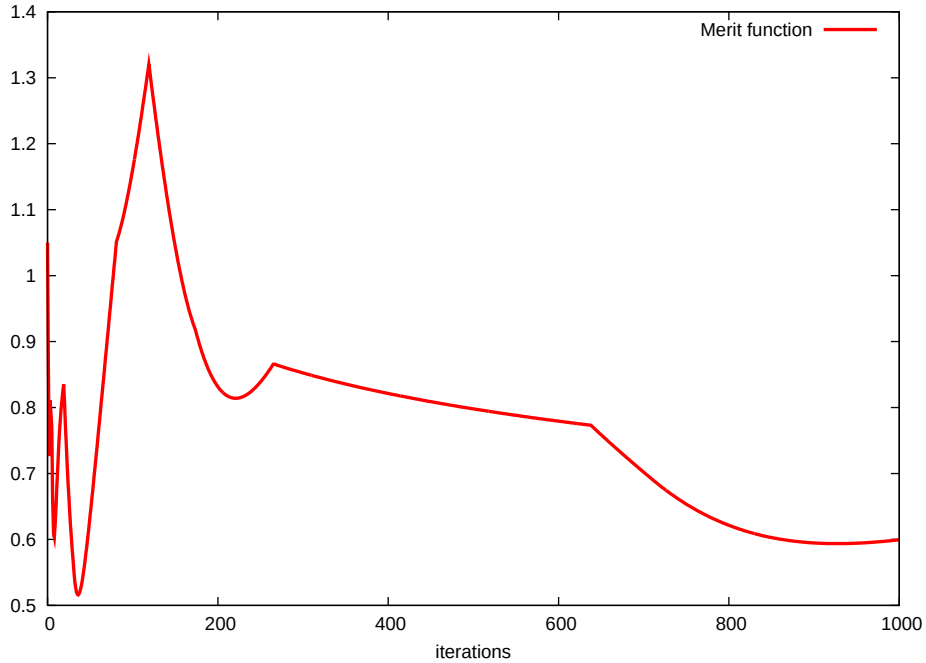

Figure 2.8 Failure of the projected gradient method on a random problem

allowing interaction between investment and operation. Broadly, there is a negative feedback, which supports an equilibrium: Higher investment leads to higher capacity, a lower output price, a lower return on investment, and thus lower investment.

The standard model of investment is the Markowitz portfolio selection model. The levels of investment in assets are chosen to minimize a convex combination of the expected value of the negative of the rate of return and the variance. The joint distribution of the rates of return of the individual assets is exogenous. The weight γ on the variance is a proxy for risk aversion; with $\gamma = 0$ investors are risk neutral; with $\gamma > 0$ they are risk averse.

The standard model of operation of productive capacity for a single good is the partial equilibrium model. Activity levels of production technologies of given capacity and marginal cost are selected to minimize the total cost of production subject to the constraint that total supply exceed demand. The price of the good is given by the Lagrange multiplier on the supply constraint. Demand may be elastic, in which case the consumer's surplus is subtracted from the objective. Stochastic demand may be considered by formulating the model for a number of different demand scenarios, yielding a distribution on the good price and the output of each production technology.

The investment model yields capacity as a function of returns. The operation model yields production and price as a function of capacity, which, together with fixed costs, give returns as a function of capacity. Our model combines the investment model and the operation model, yielding capacity, production, price and returns simultaneously.

The equilibrium model can be contrasted with two stochastic optimization models of investment and operation of productive capacity. The first optimization model selects investment and operation so as to minimize the expected total cost subject to the demand requirement. This model represents risk neutral system-optimal investment. The second optimization model minimizes a specified convex risk measure of total cost. The first model lacks risk-aversion, which is an essential feature of the investment environment. The second model distorts operation in ways that depend on exactly what risk measure is used. Neither optimization model adequately represents the independence of the decision-makers on the levels of investment and operation.

The model can be easily implemented numerically using the EMP feature of GAMS. relatively small instances can then be solved by PATH. We also consider solution of the model by iterative methods focusing on the vector of investment levels. Such iterative methods are more suitable for large-scale models. Sensible model results require a rather rich representation of demand uncertainty, which results in a large-scale model, requiring an iterative solution algorithm.

We consider the sensitivity of model results to the degree of investor risk aversion and to the marginal value and elasticity of demand. Generally we expect that greater risk aversion leads to lower investment in each technology. The sensitivity to marginal value of demand is somewhat ambiguous as a lower marginal value leads to greater variance of revenue but may also lead to greater mean revenue to production.

We consider a specific application of this model to the wholesale electric power market. Here demand is essentially inelastic, and load shedding plays the role of marginal value of demand. Demand uncertainty represents both the load-duration curve and unexpected outages of various grid components.

We also consider energy storage, which can be viewed as a production technology with a more complicated operation model. The operation model of energy storage introduces a time interval, allowing extra energy to be stored in one time period to be used in a later time period, as long as the net energy taken out of storage over the full interval is 0. Thus each demand scenario must specify demand as a function of time over the interval.

2.3.2 Initial model

Let J be the set of production technologies for a given good, defined by fixed costs c_{0j}, variable costs c_{1j} and legacy capacities $\underline{x}_j > 0$. Let Ω be the set of scenarios, with probabilities $\pi_\omega > 0$. Demand $q_{d\omega}$ for the good is given by a linear demand curve with reference quantity $q_{d\omega 0} > 0$ and reference price c_{1d} and slope $-c_{2d}$. Then c_{1d} is the marginal value of demand at the reference quantity and c_{2d} is comparable to the inverse elasticity of demand. Capacity x_j in technology j is bounded below by legacy capacity:

$$x_j \geq \underline{x}_j \tag{2.26}$$

Production $q_{j\omega} \geq 0$ in technology j is bounded by capacity:

$$q_{j\omega} \leq x_j \tag{2.27}$$

Demand is met by production:

$$\textstyle\sum_j q_{j\omega} = q_{d\omega} \quad (\perp p_\omega) \tag{2.28}$$

with Lagrange multiplier p_ω giving the market price of the good. Production and demand are chosen so as to minimize the social loss:

$$\min_{q_\omega \geq 0} \sum_j c_{1j} q_{j\omega} - \left(c_{1d} + c_{2d} q_{d\omega 0}\right) q_{d\omega} + \frac{1}{2} c_{2d} q_{d\omega}^2 \tag{2.29}$$

Returns $r_{j\omega}$ to investment in technology j in scenario ω are given by the short-term profit minus the capacity cost, divided by the capacity cost:

$$r_{j\omega} = \frac{(p_\omega - c_{1j}) q_{j\omega} - c_{0j} x_j}{c_{0j} x_j} = \frac{(p_\omega - c_{1j}) q_{j\omega}}{c_{0j} x_j} - 1 \tag{2.30}$$

Given returns r, investment in capacity x is perceived to incur a stochastic loss

$$z_\omega = -\sum_j r_{j\omega} c_{0j} x_j \tag{2.31}$$

Then x is selected so as to minimize a specified convex risk measure $\rho(z) = \rho(x, r)$. The full model is formulated as an equilibrium problem:

$$\begin{array}{ll} \min_{q_\omega \geq 0} & \textstyle\sum_j c_{1j} q_{j\omega} + c_{1s} q_{s\omega} \\ \text{s.t.} & \textstyle\sum_j q_{j\omega} + q_{s\omega} = d_\omega \quad (\perp p_\omega) \\ & q_{j\omega} \leq x_j \\ \min_{x \geq \underline{x}} & \rho(x, r) \\ r_{j\omega} = & \frac{(p_\omega - c_{1j}) q_{j\omega}}{c_{0j} x_j} - 1 \end{array} \tag{2.32}$$

2.3.3 Alternate models based on stochastic optimization

The EMP model and its formulation as a VI are not simple. The theory of existence of solutions for such models is narrow, and algorithms for numerical solution are not robust. In contrast

linear and nonlinear programming have a complete existence theory and there are many mature algorithms and computer codes for numerical solution. So as a modeler, we must show that the real-world phenomenon we want to model cannot be modeled by LP or NLP. In practice this comes down to two arguments: (1) All reasonable optimization models fail to represent some important feature that we want to model. (1) The equilibrium model does represent the features that we want to model and is not integrable as the first-order optimality conditions of an optimization problem.

For our problem, there are two reasonable optimization models representing investment and operation of productive capacity. Here we introduce these models and explain their deficiencies as compared to the equilibrium model.

A risk-neutral LP model:

$$\begin{array}{lll} \min_{x\geq\underline{x},q\geq 0} & \sum_\omega \pi_\omega \left(\sum_j \left(c_{0j}x_j + c_{1j}q_{j\omega}\right) + c_{1s}q_{s\omega}\right) & \\ \text{s.t.} & \pi_\omega \left(\sum_j q_{j\omega} + q_{s\omega} - d_s\right) = 0 & (\perp p_\omega) \\ & q_{j\omega} \leq x_j & \end{array} \tag{2.33}$$

A risk-averse NLP:

$$\begin{array}{lll} \min_{x\geq\underline{x},q\geq 0} & \sum_\omega \pi_\omega L\left(\sum_j \left(c_{0j}x_j + c_{1j}q_{j\omega}\right) + c_{1s}q_{s\omega}\right) & \\ \text{s.t.} & \pi_\omega \left(\sum_j q_{j\omega} + q_{s\omega} - d_s\right) = 0 & (\perp p_\omega) \\ & q_{j\omega} \leq x_j & \end{array} \tag{2.34}$$

where L is a convex function.

The LP model fails to represent risk-aversion in the investment decision. Risk-aversion arises fundamentally from the inability of investors to raise short-term funds to meet any possible daily loss. Even if the expected daily loss is zero, a positive probability of a single day's loss exceeding a certain threshold poses a risk that cannot be hedged. More commonly, larger losses require larger loans to meet in the short term, and these require higher interest rates, so the marginal cost of a loss is increasing. In the LP model we see that all technologies earn 0 expected return and incur a nonzero variance of return. In a risk-averse investment environment, the variance must be compensated by a positive expected return.

The deficiencies of the NLP model are more subtle. First, there is a principled argument that the objective under consideration is incorrect: The revenue to investment is represented only indirectly,

by the consumer's surplus, rather than by the output price and quantity. But it is not possible to include the output price in the objective if it is to be the Lagrange multiplier on the market equation. The NLP model attempts to combine the investors's objective with the dispatcher's objective, but these two agents face fundamentally different incentives.

In our first application to the elctricity market we consider the effect of the load-shedding price κ. Without this price cap, efficient investment leads the technology with the highest marginal cost to build capacity beyond its break-even point, as it is called on to meet even the highest level of demand throughout the year. With the price cap this costly technology is able build less capacity and earn the maximum price in more hours out of the year. In the risk neutral LP model this is exactly what happens. In figure (2.9 we see increasing investment in high marginal cost natural gas turbine technology as κ increases. In the risk-averse EMP model we see this behavior at first in figure (2.10), but as the load-shedding price increases, the investment risk becomes high enough to depress investment in all generation technologies, and more and more of the load is met by 0-fixed cost load shedding. And as the risk-aversion parameter γ increases, we see in figure (2.11) that the amount of load shedding increases. This model has the disconcerting result that risk aversion, together with the behavior of other investors, can yield an outcome that is substantially worse for society at large than a risk-neutral investment decision. It makes sense to treat society at large as risk-neutral because it is large enough to hedge the risk that smaller investment agents must be averse to.

We have also tried solving the EMP model with the iterated optimal reaction algorithm and the projected gradient algorithm. We tried a range of fixed step sizes and dynamic step size methods and were simply unable to obtain convergence with positive risk aversion γ. We believe that these algorithms display the same cyclic behavior on the EMP model as we observed on the counterexamples in the previous section. However this is a large scale model with real-world interest and where MOPEC makes a contribution that LP and NLP could not make. We are able to solve the model only because PATH represents the interactions between agents accurately.

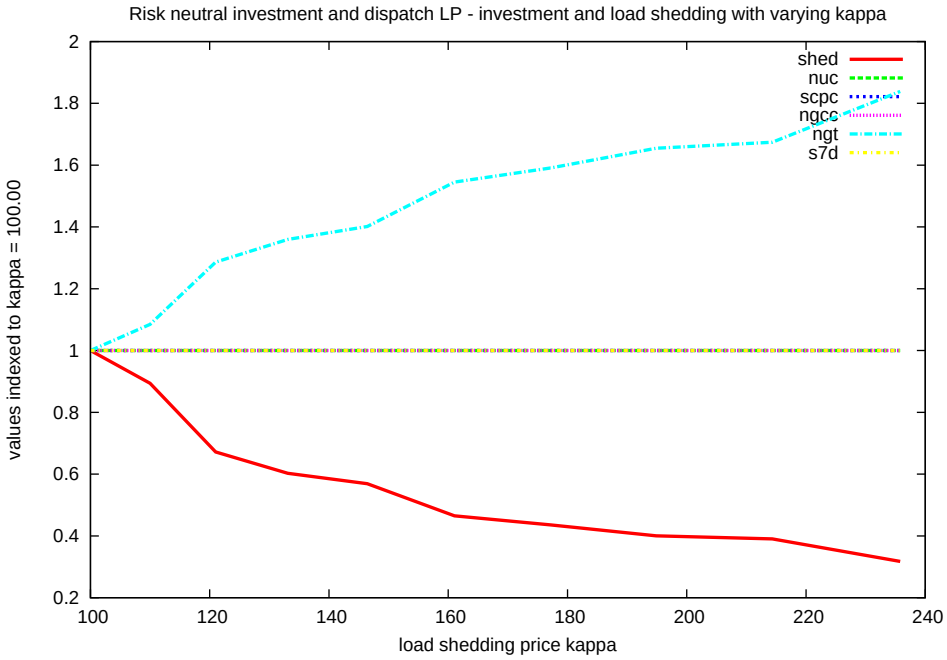

Figure 2.9 Risk-neutral investment at different values of κ

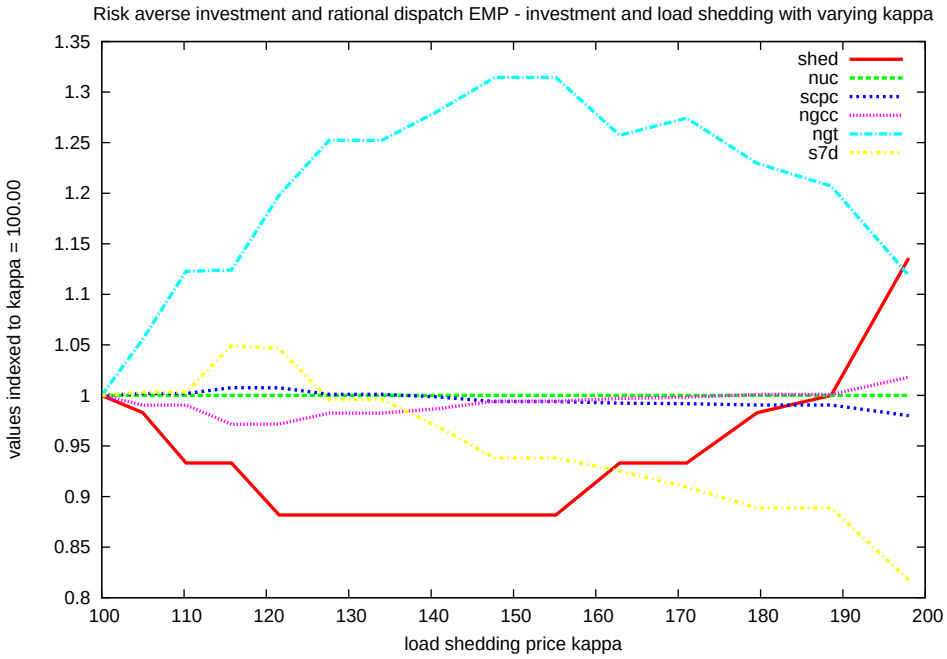

Figure 2.10 Risk-averse equilibrium investment at different values of κ

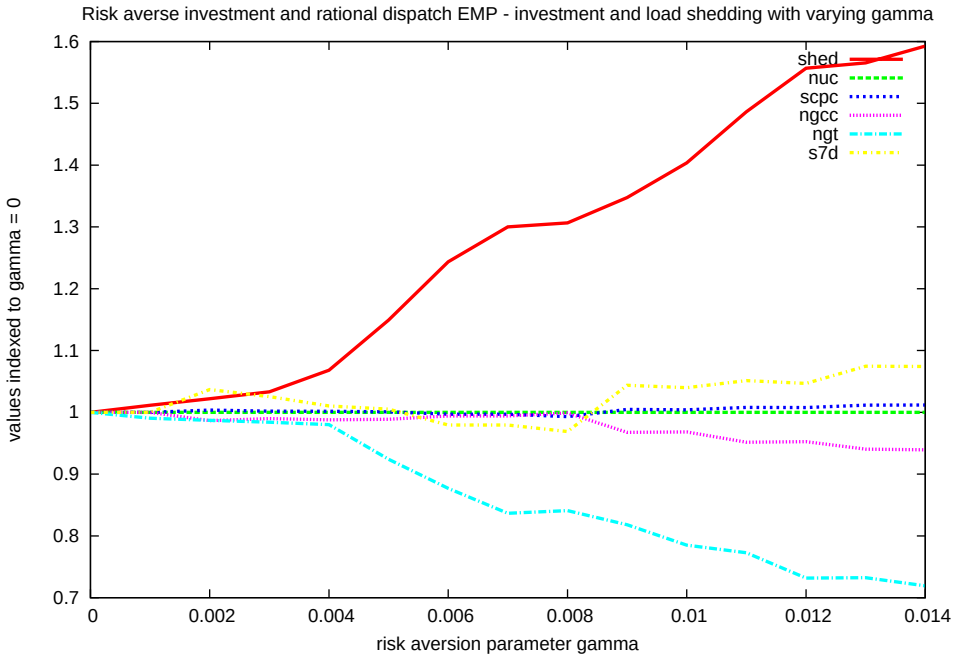

Figure 2.11 Risk-averse equilibrium investment at different values of γ

2.4 Conclusion

Without an agent-decomposable algorithm with wide applicability, we are left to find more exotic ways of handling nonmonotone MOPEC problems. Our most versatile tool for nonmonotone problems is discretization, in which essentially the whole domain of the problem is approximated by a discrete grid and each point evaluated separately. This discretization technique is computationally feasible in only low dimensional domains, so any nonmonotone feature of a problem must be confined to a low-dimensional structure or else the whole problem must be represented in a low dimensional space. In the next chapter we give an example of the former approach in the context of optimization, where the concept analgous to monotonicity is convexity. And in the following two chapters we give an example of the latter approach.

Chapter 3

Nonconvexity resolved by discretization: An example from industry

3.1 Introduction

In this chapter we describe an original method for the global solution of a nonconvex optimization problem with a particular structure. The structural assumtion is that the nonconvexity is confined to a small number of low-dimensional constraints. We approximate each such constraint by a piecewise linear constraint constructed on a grid. We thus convert a problem of nonconvex nonlinear programming, for which general-purpose global solvers are not robust, to a problem of integer linear programming, for which there are many robust and sophisticated solvers. We give conditions under which our approximation method gives rise to both a relaxation and a primal heuristic, and we show how to refine our approximation so that the gap between the relaxation and the heuristic can be made arbitrarily small, thus solving the problem.

We then describe the application of our method to a large nonconvex optimization problem coming from the strategic planning of a coal mine. The nonconvexity in this problem arises from the definition of coal quality and its value to customers, and this interpretation facilitates the verification of the assumptions of our solution method. We show that our method outperforms the available general purpose solution methods on several large instances constructed from data provided by a large US mining firm. On these instances, the general purpose global solution methods are unable handle the size of the problem, and make no discernible progress in any reasonable amount of time. On some of these instances, local solution methods fail to find a feasible point, and on others the local methods terminate at a point that is not globally optimal. Our method solves

all these instances quickly, in some cases improving the objective value obtained by other methods by 10%, or about half a million dollars over five years.

In the remainder of this section we briefly review nonconvex optimization and solution methods. In the next section we describe our method. In the section after that we describe the application to strategic planning of a coal mine. And lastly we indicate future directions of this research.

3.1.1 Nonconvex optimization and solution methods

An optimization problem $\max_{x \in X} f(x)$ in which f is not concave or X is not convex is called a nonconvex optimization problem. Nonconvex optimization problems are difficult to solve because there are no criteria for a solution that can be easily checked. E.g. if $X = \mathbb{R}^n$, then the natural solution criterion that is $df(x) = 0$, i.e. that x be a stationary point of f. If f is not concave, then this first order stationarity condition is neither necessary nor sufficient for a solution. Points satisfying first order stationarity may be local maxima but not global maxima.

Generally methods for solution of nonconvex optimization problems fall into one of two categories: local, and global. Local methods attempt to find a stationary point. These methods may get stuck at a local maximizer or may even fail to find a feasible point. Global methods are guaranteed to find a global maximizer but they are typically much slower than local methods, as they work by subdividing the feasible set and applying a local method and a bounding method on each subdivision.

One case in which nonconvex optimization is quite approachable, practically if not theoretically, is integer linear programming (MIP). In a MIP problem the objective f is linear, and the feasible set X is given by linear constraints and the requirement that certain variables take integer values. The feasible set is thus nonconvex, but the nonconvexity is highly structured. Solvers for MIP have been improving for decades and perform quite well in practice. They can handle much larger problems than general purpose global solvers for nonconvex optimization without this special structure.

3.2 Global solution of a nonconvex optimization problem by grid approximation

Consider an abstract optimization problem $s = \max_{x \in X} f(x)$ with closed bounded domain $X \subset \mathbb{R}^n$ with nonempty interior and continuous objective $f : X \to \mathbb{R}$. Knowing nothing more about the problem, the only sensible solution method is to evaluate f on a grid of points in $\mathbb{R}^n$, testing each point for membership in X, and returning the feasible point with maximal objective value. The observed variation in values of f and in membership in X can be used to decide whether to repeat the procedure on a finer grid. No guarantee of optimality or even an upper bound on s can be given, and the overall method is computationally intensive for large n. Quantitative complexity estimates can be given in terms of Lipschitz regularity of f and the boundary of X but these constants are typically not known.

3.2.1 Approximating a low-dimensional nonconvex structure

The grid approximation method is computationally feasible when the dimension n is relatively small. In a high-dimensional optimization problem, if the nonconvexity is confined to a low-dimensional structure, then that structure can be approximated by points on a low-dimensional grid. We show now how to incorporate this approximation into a MIP formulation of the overall problem.

Consider an optimization problem of the form

$$\begin{aligned} s = \quad & \max \quad f(w, x, y, z) \\ & \text{s.t.} \quad (w, x, y, z) \in C \\ & \qquad z \leq g(x, y) \end{aligned} \tag{3.1}$$

where $C \subset \mathbb{R}^n$ is a closed bounded nonempty convex set, x, y, z are scalar variables, w is a vector variable, $f : C \to \mathbb{R}$ is a concave function, and $g : \mathbb{R}^2 \to \mathbb{R}$ is continuous. With the exception of the constraint $z \leq g(x, y)$, (3.1) is a convex optimization problem, and this constraint involves only three scalar variables. To solve this problem our strategy is to replace this constraint by a piecewise linear approximation generated by evaluating g on a grid. The low dimensionality

of this constraint will enable this approximation to attain sufficient accuracy with a manageable number of evaluations of g.

We begin by defining grid points x_i^*, y_j^* with $x_i^* < x_{i+1}^*$ and $y_j^* < y_{j+1}^*$. Then let $z_{ij}^* = g(x_i^*, y_j^*)$ and

$$\overline{z}_{ij} = \max\{z_{ij}^*, z_{i,j+1}^*, z_{i+1,j}^*, z_{i+1,j+1}^*\} \tag{3.2}$$

A simple approximation of (3.1) is given by introducing continuous variables x_i, y_j, z_{ij} and binary variables $\mu_i, \lambda_j, \theta_{ij}$:

$$\begin{aligned}
\tilde{s} = \quad & \max \quad f(w,x,y,z) \\
& \text{s.t.} \quad (w,x,y,z) \in C \\
& \qquad x = \textstyle\sum_i x_i \\
& \qquad y = \textstyle\sum_j y_j \\
& \qquad z = \textstyle\sum_{ij} z_{ij} \\
& \qquad \lambda_i x_i^* \le x_i \le \lambda_i x_{i+1}^* \\
& \qquad \mu_j y_j^* \le y_j \le \mu_j y_{j+1}^* \\
& \qquad z_{ij} \le \theta_{ij} \overline{z}_{ij} \\
& \qquad \theta_{ij} = \lambda_i \mu_j \\
& \qquad \textstyle\sum_i \lambda_i = 1 \\
& \qquad \textstyle\sum_j \mu_j = 1 \\
& \qquad \lambda_i, \mu_j, \theta_{ij} \in \{0,1\}
\end{aligned} \tag{3.3}$$

The bilinear constraint $\theta_{ij} = \lambda_i \mu_j$ can be enforced by linear inequalities, given that $\lambda_i, \mu_j, \theta_{ij} \in \{0,1\}$. Assuming that the convex domain C and the concave objective f can be modeled adequately by linear constraints, (3.3) is a MIP model.

The constraints ensure that $\lambda_i > 0$ for exactly one $i = i^*$ and $\mu_j > 0$ for exactly one $j = j^*$, that $x_{i^*}^* \le x \le x_{i^*+1}^*$ and $y_{j^*}^* \le y \le y_{j^*+1}^*$, and that $z \le \overline{z}_{i^*j^*}$, and this justifies our characterization of (3.3) as an approximation of (3.1). For this, define the mesh size of the grid by

$$\epsilon = \max\{\max_i (x_{i+1} - x_i), \max_j (y_{j+1} - y_j)\} \tag{3.4}$$

and assume a constraint qualification such as

$$\text{int}\{(w, x, y, z) \in C : z \leq g(x, y)\} \neq \emptyset \tag{3.5}$$

Then over a sequence of grids with $\epsilon \to 0$, we have $\tilde{s} \to s$.

As we have defined a version of z for each grid cell, we have greatly increased the number of variables. But in so doing we enable an approximation of the nonconvex constraint to be incorporated into a MIP formulation of the problem. In the practice of optimization modeling using MIP, a formulation using a large number of variables to express a complicated constraint on a small number of variables is called an extensive formulation. Extensive formulations typically give rise to very tight bounds in MIP solvers leading to good performance despite the large number of variables [53]. This practical experience with MIP extensive formulations partly motivates our MIP extensive approximation of (3.1).

In the more general case of multiple low-dimensional nonconvex constraints in a single problem, we can still apply this technique, simply by approximating each constraint independently of the others. All our further analysis applies mutatis mutandis. The computational burden here can be rather large, but it is still substantially less than that of a naive application of a gridded approximation ignoring the low-dimensional structure. Indeed suppose there are K constraints $z_k = g_k(x_k, y_k)$, and each one requires a grid Q_k containing L points. The naive technique evaluates f and checks feasibility at each point in the Cartesian product $\Pi_k Q_k$, doing at least L^K evaluations. By taking advantage of the low-dimensional structure of the problem our technique uses only $O(LK)$ function evaluations. The general case covers a wide variety of applications, including the example that we treat later in this chapter.

3.2.2 Relaxation under an extreme value property

With only an abstract convergence result it may be difficult to know how fine a grid is needed in (3.3) to obtain an acceptable approximation of (3.1). An a priori estimate can be given in terms of the Lipschitz constants of f and g but these are typically unknown. For practical use of this

method, we now give conditions under which the basic approximation (3.3) is actually a relaxation of (3.1), i.e. $s \leq \tilde{s}$.

Let us assume that g has a certain extreme value property, i.e. that for all grids in x and grids in y and all (x, y) with $x_i^* \leq x \leq x_{i+1}^*$ and $y_j^* \leq x \leq y_{j+1}^*$ we have $g(x, y) \leq \overline{z}_{ij}$. Under this extreme value property, we denote the optimal value of (3.3) by $\overline{s}$, and it is immediate that $s \leq \overline{s}$.

Many functions that are useful for modeling satisfy this extreme value property, including, on the positive orthant, all homogeneous quadratics $g(x, y) = ax^2 + bxy + cy^2$ and the rational function $g(x, y) = x/y$. These examples certainly are not concave in general, so (3.1) is indeed a nonconvex optimization problem.

3.2.3 A primal heuristic under a monotonicity property

In addition to a relaxation of (3.1) it would also be useful to have a primal heuristic, i.e. a method of obtaining a feasible point for (3.1), whose objective value $\underline{s}$ consequently satisfies $\underline{s} \leq s$. We now give conditions under which (3.3) can be modified to obtain a primal heuristic.

Of course without any special conditions, not even the extreme value property, (3.3) provides a candidate for a primal heuristic. Given a solution (w, x, y, z) of (3.3) consider the point $(w, x, y, g(x, y))$ and $\underline{s} = f(w, x, y, g(x, y))$. Certainly $\underline{s} \leq s$, but there is no guarantee that a solution defined in this way is feasible for (3.1), i.e. it may not lie in C, even if the extreme value property is assumed. More generally one might replace z with a sequence of values less than $g(x, y)$ until a feasible point is found, but this would be quite challenging with more than one low-dimensional nonconvex constraint. Instead we give a sufficient condition and a modification of (3.3) so that the candidate solution $(w, x, y, g(x, y))$ is feasible.

We assume that $f(w, x, y, z)$ is nondecreasing in z, and define

$$\underline{z}_{ij} = \min\{z_{ij}^*, z_{i,j+1}^*, z_{i+1,j}^*, z_{i+1,j+1}^*\} \tag{3.6}$$

and replace $\overline{z}$ in (3.3) with $\underline{z}$:

$$\begin{aligned}
\max \quad & f(w,x,y,z) \\
\text{s.t.} \quad & (w,x,y,z) \in C \\
& x = \textstyle\sum_i x_i \\
& y = \textstyle\sum_j y_j \\
& z = \textstyle\sum_{ij} z_{ij} \\
& \lambda_i x_i^* \le x_i \le \lambda_i x_{i+1}^* \\
& \mu_j y_j^* \le y_j \le \mu_j y_{j+1}^* \\
& z_{ij} \le \theta_{ij} \underline{z}_{ij} \\
& \theta_{ij} = \lambda_i \mu_j \\
& \textstyle\sum_i \lambda_i = 1 \\
& \textstyle\sum_j \mu_j = 1 \\
& \lambda_i, \mu_j, \theta_{ij} \in \{0,1\}
\end{aligned} \tag{3.7}$$

Then for a solution (w,x,y,z) of (3.7), the point $(w,x,y,g(x,y))$ is feasible for (3.1) and $\underline{s} = f(w,x,y,g(x,y))$ satisfies $\underline{s} \le s$, giving a primal heuristic.

This monotonicity property on the objective f is quite common in applications. Essentially it says that z represents a quantity of something that the decision maker would rather have more of under any circumstances. This is precisely the definition of an economic good. In this interpretation, $g(x,y)$ represents the definition of the amount of this good that is made available by a decision defined by (x,y). From a modeling perspective, an equation $z = g(x,y)$ might be more natural than the inequality $z \le g(x,y)$, but under the monotonicity property, only the inequality needs to be enforced, which is important for the validity of our primal heuristic. With this monotonicity property, the constraint $z_{ij} \le \theta_{ij}\underline{z}_{ij}$ is essentially pessimistic, while the corresponding constraint $z_{ij} \le \theta_{ij}\overline{z}_{ij}$ in (3.3) is optimistic.

3.2.4 Combining the relaxation and heuristic and refining the grid

Under both assumptions, the extreme value property of g and the monotonicity assumption on f, we may obtain both an upper bound $\overline{s}$ and a lower bound $\underline{s}$ on the optimal value s and a feasible point. If the estimate $\overline{s} - \underline{s}$ of the optimality gap of the feasible point is small enough then we simply take this feasible point as an adequate solution of (3.1) with a rigorous upper bound on the optimality gap. If the bounds $\underline{z}_{ij}$ and $\overline{z}_{ij}$ are close enough then the optimality gap will be small, and indeed this method is particularly successful when these bounds are close just by the nature of the function g.

If the gap is too large then the grid in x and y can be refined, and a new feasible point and objective bounds computed. This refinement process can be carried out to any desired optimality tolerance. Refinement of the grid can be done in a number of different ways, but we have obtained the best results from dividing in half those grid cells containing the values of the solution of either the relaxation or the approximation. This method keeps the size of the MIP subproblems moderate. Another subdivision scheme that we considered is to divide a grid cell at exactly the point where the solution of the relaxation or the approximation falls. Under this scheme the algorithm stalled, making no further progress in the optimality gap.

Thus our method solves the nonconvex optimization problem (3.1) by solving a sequence of MIP problems. It is able to give a rigorous bound on the optimality gap of the solution it returns, and it is able to drive that bound arbitrarily close to 0, so it is a global solution method. As solvers for MIP are more robust than general purpose global solvers for nonconvex optimization, our method can handle much larger problem instance than other global solvers.

3.3 Application to the StratPlan problem

We now apply of our method to a nonconvex optimization problem for the strategic planning of a coal mine with quality incentives. This problem arose through work with Peabody Energy, a large coal mining firm. At Peabody this problem is known as StratPlan. During this work the problem was modeled in a natural way using mixed integer nonlinear programming (MINLP). MINLP is

a very general algebraic class of nonconvex optimization for which a number of global and local solvers are available. It became evident that these general purpose solvers were unable to handle the nonconvexity and size of the problem at hand. This work motivated the development of the present algorithm for nonconvex optimization with low-dimensional nonconvexity structure.

In this section we describe the StratPlan problem, give its natural MINLP formulation, and detail the formulation using our discrete approximation method of the previous section. We devote particular attention to how nonconvexity arises from quality incentives and how this nonconvexity is formulated in a low-dimensional structure that meets the assumptions of our solution method as well as the intuitive condition for good performance of our method.

3.3.1 The StratPlan problem in brief

The goal of the StratPlan problem is to plan several years of extraction of coal from pits of varying quality and mining cost to meet customer contracts paying quality-dependent prices. In each year, the amount of coal that is extracted from each pit and allocated to each contract determines the mining cost, the contract quantity, quality, price and revenue, and net profit. The solution must observe bounds on pit quantity and contract quantity and quality. The objective is to maximize net profit. As the dependence of contract price on quality are at the root of the nonconvex difficulty of the problem we emphasize that modeling this dependence correctly is crucial to ensuring that the high-quality coal goes to those contracts that pay the most for it under the overal objective of profit maximization.

StratPlan might be solved once per year. The contract prices and quantity and quality bounds might be updated each year as new contracts are struck. Mining costs might be updated as equipment is moved into or out of specific pits or the mine as a whole. These contract and equipment decisions and indeed the overall plan of the order in which to mine the different areas of each pit are made at a higher level and are considered fixed in the context of StratPlan. On the other hand the decisions taken by StratPlan concerning the quantity of material to allocate from each pit to each contract in each year and the resulting quality estimates for each contract are used to guide the shorter term blending problem faced when a given train car destined for a given contract must

be filled from silos of coal of given qualities that have already been mined. StratPlan is thus a problem of medium-term quality and quantity optimization.

To demonstrate the general solution method of the previous section we define a simplified version of the StratPlan problem. The simplified problem includes only one of the several qualities (e.g. BTU but not sulfur) and does not consider quality tracking (the requirement for some contracts that the quality be roughly constant over the planning horizon). Furthermore the simplified problem considers only qualities that are defined as the ratio of quality-bearing material to total material in a given sample. The general problem includes more complex qualites, e.g. the product of two simple qualities.

The main difficulties posed by nonconvexity in the natural MINLP formulation are still present in the simplified problem, and the general solution method remains valid in the full problem. But the computer coding is much less arduous in the simplified problem. And we wish at this time only to provide an demonstration of our method at industrial scale on a problem with real data.

3.3.2 A natural MINLP formulation with SOS2 constraints and quadratic equations

Here we describe the natural algebraic formulation of the StratPlan problem. This formulation was natural to the author because it is essentially an algebraic transcription of the problem as posed by Peabody. It uses the combinatorial constraints known as special ordered sets of type 2 (SOS2) to express the cost incurred and quality-bearing material contained in a quantity of total material mined from a given pit in a given year. Quality is then expressed as the ratio of quality-bearing material to total material using quadratic constraints. The SOS2 constraints are already nonconvex, but they are handled with good practical performance by MIP solvers. It is the quadratic constraints that make the model nonconvex in the sense of nonlienar programming and that require a MINLP solver.

The fundamental decision variables are the quantities (material) M_{pcy} mined from pit p and allocated to contract c in year y. Pit contract quantities are subject to lower bounds

$$M_{pcy} \geq 0 \tag{3.8}$$

From the pit-contract quantities the pit quantities M_{py} and contract quantities M_{cy} are given by

$$M_{py} = \sum_{c} M_{pcy} \tag{3.9}$$

and

$$M_{cy} = \sum_{p} M_{pcy} \tag{3.10}$$

Pit and contract quantities are subject to given lower and upper bounds

$$\underline{M}_{py} \leq M_{py} \leq \overline{M} \tag{3.11}$$

and

$$\underline{M}_{cy} \leq M_{cy} \leq \overline{M}_{cy} \tag{3.12}$$

The cumulative pit quantities M'_{py} satisfy

$$M_{py} = M'_{py} - M'_{p,y-1}|_{y>1} \tag{3.13}$$

Cumulative pit quality-material QM'_{py} and cumulative pit cost-material CM'_{py} are piecewise linear functions of cumulative pit material. To model this relation, first cumulative pit-block quantities M'_{pb} are computed from given pit-block quantities $M_{pb'}$ by

$$M'_{pb} = \sum_{b' \leq b} M_{pb'} \tag{3.14}$$

Then cumulative pit quantity is interpolated to the computed values of cumulative pit-block quantities by

$$M'_{py} = \sum_{b} M'_{pb} T_{pby} \tag{3.15}$$

where the interpolation coefficients T_{pby} satisfy

$$T_{pby} \geq 0 \tag{3.16}$$

and

$$1 = \sum_{b} T_{pby} \tag{3.17}$$

and for each (p, y) there are at most two blocks b with $T_{pby} > 0$ and these blocks are consecutive. This combinatorial requirement identifies the set $\{T_{pby} : b\}$ as special ordered set of type 2 (SOS2). Then cumulative pit quality-material and cumulative pit cost-material are interpolated to the given values QM'_{pb} and CM'_{pb} of cumulative pit-block quality-material and cumulative pit-block cost-material by

$$QM'_{py} = \sum_b QM'_{pb} T_{pby} \tag{3.18}$$

and

$$CM'_{py} = \sum_b CM'_{pb} T_{pby} \tag{3.19}$$

Then pit quality-material QM_{py} and pit cost-material (gross cost) CM_{py} are given by the differences of successive values of the corresponding cumulative pit quality-material and cumulative pit cost-material:

$$QM_{py} = QM'_{py} - QM'_{p,y-1}|_{y>1} \tag{3.20}$$

and

$$CM_{py} = CM'_{py} - CM'_{p,y-1}|_{y>1} \tag{3.21}$$

Contract quality Q_{cy} is the average of the pit qualities in year y weighted by the corresponding pit-contract quantities. To model this contract quality definition, first pit-contract material is expressed as a share S_{pcy} of pit material by the quadratic equation

$$M_{pcy} = S_{pcy} M_{py} \tag{3.22}$$

with

$$1 = \sum_c S_{pcy} \tag{3.23}$$

and

$$S_{pcy} \geq 0 \tag{3.24}$$

Then pit-contract quality-material is expressed as a share of pit quality-material by the quadratic equation

$$QM_{pcy} = S_{pcy} QM_{py} \tag{3.25}$$

And contract quality Q_{cy} is defined by the quadratic equation

$$QM_{cy} = Q_{cy}M_{cy} \tag{3.26}$$

Contract qualities are subject to given lower bounds

$$Q_{cy} \geq \underline{Q}_{cy} \tag{3.27}$$

These quadratic constraints, and others that will be described later, are what make StratPlan a challenging problem of nonconvex optimization.

The price paid by each contract is a piecewise linear function of the contract quality. To model this relation, first contract quality is interpolated to given contract quality tiers Q_{cty} by

$$Q_{cy} = \sum_t Q_{cty}T_{cty} \tag{3.28}$$

where the interpolation coefficients T_{cty} satisfy

$$1 = \sum_t T_{cty} \tag{3.29}$$

and

$$T_{cty} \geq 0 \tag{3.30}$$

and for each (c, y) there are at most two tiers t with $T_{cty} > 0$ and these tiers are consecutive. Then the contract price P_{cy} are interpolated to given values of contract-tier prices P_{cty} by

$$P_{cy} = \sum_t P_{cty}T_{cty} \tag{3.31}$$

The revenue (price-material) PM_{cy} paid by contract c in year y is given by the quadratic equation

$$PM_{cy} = P_{cy}M_{cy} \tag{3.32}$$

Finally the objective is to maximize profit Z, formulated as the difference between contract revenues and pit gross costs:

$$Z = \sum_{cy} PM_{cy} - \sum_{py} CM_{py} \tag{3.33}$$

Thus the natural MINLP formulation of the StratPlan problem is

$$\begin{aligned} \max \quad & Z \\ \text{s.t.} \quad & (3.8), (3.9), \ldots, (3.33) \\ & \forall (p, y), \{T_{pby} : b\} \text{isSOS2} \\ & \forall (c, y), \{T_{cty} : t\} \text{isSOS2} \end{aligned} \tag{3.34}$$

3.3.3 Modeling SOS2 constraints with binary variables

The piecewise linear functions involved in the natural algebraic formulation require that the interpolation coefficients T_{pby} satisfy certain combinatorial constraints. Specifically for each (p, y) there are at most two blocks b with $T_{pby} > 0$ and these blocks are consecutive, so that T_{pby} form a special ordered set of type 2 (SOS2) with respect to b. Similarly the interpolation coefficients T_{cty} are SOS2 with respect to t. These SOS2 constraints can be handled directly by some but not all solvers, so I discuss here a method of enforcing them by means of some additional constraints and binary variables. This alternative formulation in terms of binary variables is useful for the formulation as a MIP relaxation and primal heuristic that is used in our eventual solution method.

An interval indicator V_{pby} selects the block b in which cumulative pit quantity M_{pby} lands by

$$T_{pby} \leq V_{p,b-1,y} + V_{pby} \tag{3.35}$$

The interval indicators satisfy

$$\sum_b V_{pby} = 1 \tag{3.36}$$

and

$$V_{pby} \in \{0, 1\} \tag{3.37}$$

Similarly, contract quality tier selection is modeled with interval indicators V_{cty} satisfying

$$T_{cty} \leq V_{c,t-1,y} + V_{cty} \tag{3.38}$$

and

$$\sum_t V_{cty} = 1 \tag{3.39}$$

and

$$V_{cty} \in \{0, 1\} \tag{3.40}$$

Thus the alternative MINLP formulation of StratPlan problem using binary variables is

$$\begin{aligned} \max \quad & Z \\ \text{s.t.} \quad & (3.8), (3.9), \ldots, (3.40) \end{aligned} \tag{3.41}$$

3.3.4 MIP extensive formulation

We now describe the MIP formulation of both the relaxation and the primal heuristic used in our solution method. The formulation is considered extensive as it introduces, for each (p, b, b, c, t, y) a decision variable $M_{pbb'cty}$ taking the value of M_{pcy} if mining in pit p in year y begins in block b and ends in block b' and the quality attained by conctract c in year y falls in tier t. These extensive quantities, defined over $b \leq b'$, are bounded below by

$$M_{pbb'cty} >= 0 \tag{3.42}$$

If $M_{pbb'cty} > 0$ then the quality contained in that quantity is known with great precision as long as the blocks b and b' are small enough. Thus the block b plays the role of a single x-cell in the abstract description of our method, and the block b' plays the role of a y-cell.

For each (p, y) there is exactly one (b, b') so that $M_{pbb'cty} > 0$ for some (c, t). That is, for each pit and year there is a unique block in which mining begins and a unique block in which mining ends. Similarly, for each (c, y) there is exactly one t so that $M_{pbb'cty} > 0$ for some (p, b, b'). To enforce these combinatorial requirements on $M_{pbb'cty}$, first introduce tier aggregations $M_{pbb'cy}$, contract-tier aggregations $M_{pbb'y}$, and pit-block aggregations M_{cty}, defined by

$$M_{pbb'cy} = \sum_{t} M_{pbb'cty} \tag{3.43}$$

$$M_{pbb'y} = \sum_{c} M_{pbb'cy} \tag{3.44}$$

and

$$M_{cty} = \sum_{pbb'} M_{pbb'cty} \tag{3.45}$$

An upper bound $\overline{M}_{pbb}$ on $M_{pbb'y}$ may be computed from the cumulative pit-block quantities by

$$\overline{M}_{pbb} = M'_{pb'} - M'_{p,b-1} \tag{3.46}$$

Then introducing pit-block pair indicators $V_{pbb'y}$, the combinatorial constraints are enforced by

$$M_{pbb'y} \leq \overline{M}_{pbb} V_{pbb'y} \tag{3.47}$$

$$M_{cty} \leq \overline{M}_{cy} V_{cty} \tag{3.48}$$

$$\sum_{bb'} V_{pbb'y} = 1 \tag{3.49}$$

and

$$V_{pbb'y} \in \{0, 1\} \tag{3.50}$$

The pit-block pair indicators are related to the the pit-block indicators by

$$V_{pby} = \sum_{b' \leq b} V_{pb'by} \tag{3.51}$$

$$V_{pby} = \sum_{b' \geq b} V_{pbb'} \tag{3.52}$$

$$V_{p,b,y-1}|_{y>1} + V_{pb'y} \leq V_{pbb'y} + 1|_{y>1} \tag{3.53}$$

$$V_{pbb'y} \leq V_{p,y-1,b} \tag{3.54}$$

for $y > 1$ and

$$V_{pbb'y} \leq V_{pyb'} \tag{3.55}$$

The material variables indexed by block pairs are related to the further aggregated material variables by

$$M_{pcy} = \sum_{bb'} M_{pbb'cy} \tag{3.56}$$

For each (p, b, b') the maximum possible quality $\overline{Q}_{pbb'}$ of material contained in $M_{pbb'y}$ can be computed from given quality data by

$$\overline{Q}_{pbb'} = \max\{Q_{pbb'}, Q_{p,b,b'-1}, Q_{p,b-1,b'}, Q_{p,b-1,b'-1}\} \tag{3.57}$$

where

$$Q_{pbb'} = QM_{pbb'}/M_{pbb'} \tag{3.58}$$

$$M_{pbb'} = M'_{pb'} - M'_{pb} \tag{3.59}$$

and

$$QM_{pbb'} = QM'_{pb'} - QM'_{pb} \tag{3.60}$$

A tight relaxation of contract-tier quality-material QM_{cty} can then be enforced by

$$QM_{cty} \leq \sum_{pbb'} \overline{Q}_{pbb'} M_{pbb'cty} \tag{3.61}$$

This constraint plays the role of the constraint $z_{ij} \leq \theta_{ij}\overline{z}_{ij}$ in the abstract description of our method. And here we see that the extreme value property holds ultimately because quality is defined as the ratio of quality-material to material.

Similarly a minimum quality value $\underline{Q}_{pbb'}$ can be computed by

$$\underline{Q}_{pbb'} = \min\{Q_{pbb'}, Q_{p,b,b'-1}, Q_{p,b-1,b'}, Q_{p,b-1,b'-1}\} \tag{3.62}$$

and can be used to constrain contract-tier quality-material by a primal heuristic

$$QM_{cty} \leq \sum_{pbb'} \underline{Q}_{pbb'} M_{pbb'cty} \tag{3.63}$$

instead of the relaxation. All contract prices are nondecreasing in quality essentially because quality is unconditionally an economic good. Hence the profit objective satisfies the monotonicity property that guarantees feasibility of the primal heuristic. And given this monotonicity property we see that the gap between the objective values in the relaxation and in the primal heuristic will be narrow as long as the pit blocks are narrow enough or quality itself does not vary too much from block to block. In our application both these favorable conditions hold so that little refinement is needed, but we can refine the blocks if desired by splitting them at the midpoint.

Membership in the correct contract quality tier is enforced by

$$QM_{cty} \geq Q_{c,t-1,y} M_{cty} \tag{3.64}$$

and

$$QM_{cty} \leq Q_{cty} M_{cty} \tag{3.65}$$

Since for each (c, y) there is at most one t with $M_{cty} > 0$, the same holds for QM_{cty}.

The contract revenue (price-tons) from tier t of contract c in year y is denoted by PM_{cty} and is defined by the linear equation

$$PM_{cty} = P_{c,t-1,y} M_{cty} + \frac{P_{cty} - P_{c,t-1,y}}{Q_{cty} - Q_{c,t-1,y}} \left(QM_{cty} - Q_{c,t-1,y} M_{cty} \right) \tag{3.66}$$

where Q_{cty} is the quality at the right endpoint of quality tier t for contract c in year y, P_{cty} is the price given quality Q_{cty}. Note that for each (c, y) there is at most one t with PM_{cty} nonzero, so that PM_{cy} can be defined by

$$PM_{cy} = \sum_{t} PM_{cty} \tag{3.67}$$

This completes the description of the MIP extensive formulation that is adaptable to either a relaxation or a primal heuristic for the StratPlan problem and can be refined to arbitrary accuracy.

Specifically, the MIP relaxation is obtained by removing the quadratic equations from the binary MINLP and adding the extensive equations using the optimistic bound $\overline{Q}_{pbb'}$ on $Q_{pbb'}$:

$$\begin{aligned} \max \quad & Z \\ \text{s.t.} \quad & (3.8), (3.9), \ldots, (3.21), (3.23), \\ & (3.24), (3.27), \ldots, (3.31), \\ & (3.33), \ldots, (3.61)(3.64), \ldots, (3.67) \end{aligned} \tag{3.68}$$

And the MIP primal heuristic is the same as 3.68 except that it uses the pessimistic bound $\underline{Q}_{pbb'}$ on $Q_{pbb'}$:

$$\begin{aligned} \max \quad & Z \\ \text{s.t.} \quad & (3.8), (3.9), \ldots, (3.21), (3.23), \\ & (3.24), (3.27), \ldots, (3.31), \\ & (3.33), \ldots, (3.63)(3.64), \ldots, (3.67) \end{aligned} \tag{3.69}$$

3.4 Numerical results for StratPlan

We have tested our solution method and several other methods on a number of instances of StratPlan. We constructed a few small instances to catch coding errors. Several large scale instances were constructed from a data set for a single large mine provided by Peabody in an Excel spreadsheet. The data set spans the years 2006-2010, contains 3 pits, 57 contracts, 850 pit-block pairs, and 5 quality tiers per quality. The data contained information on several qualities, but in each instance we focused on only on the quality of BTU, or energy density. The data set contains three alternative pricing structures, generally characterized by low, medium, and high prices respectively. We used each pricing sturucture to create a separate instance. In the original data the contract quantity lower bounds are all equal to the corresponding upper bounds. It is unrealistic that all contracts would have this characteristic /exclude Troy Ball In particular contracts representing coal to be sold on the spot market are typically created with lower bounds of 0 on quantity, and many contracts have some degree of flexibility as large electrical utilities may do their own blending. Thus we have also created different instances by manipulating a few of the lower or upper bounds on contract quantity. We have also created still more different instances by manipulating some of the bounds on pit quantity.

We focus on 6 instances, named $A_1, A_2, A_3, A_4, A_5, A_6$. All of the instances except A_2 and A_6 use the medium price structure. Instance A_2 uses high prices and A_6 uses low prices. Instances A_1 and A_2 both use the original quantity bounds. Instances A_3, A_4, A_5 have lower bounds on quantity set to 0 for contracts c_{56} and c_{57}. In A_4 and A_5 the upper bounds on quantity for these two contracts are doubled. In A_5 the lower bounds on quantity are also moved to 0. In A_6 we have moved the lower bounds on quantity to 0 for 10 of the 57 contracts.

We tested our method on these instances, along with the global MINLP solver BARON and the two local MINLP solvers SBB and DICOPT. For the local solvers we used the SOS2 formulation. BARON was not configured to handle SOS2 variables, so we used the MINLP formulation with indicator variables. The MIP subproblems used by our custom method were solved by CPLEX. The models were written in the modeling language GAMS. All computational tests were performed

on a 64-bit linux server oxon.cs.wisc.edu. In all the instances our method solved the problem to high accuracy in under a minute. Each of the general purpose methods fails in one way or another on at least one instance.

To begin with, on all six instances, both BARON and DICOPT were unable to find a feasible point after 1000 seconds. This occurred with the infeasible default starting point of all 0 values, and also with infeasible random starting points. In some cases the solver terminated sooner than the time limit, still with no feasible point.

BARON did somewhat better when given a feasible starting point E.g. on A_1, using two random starting points, two feasible points with objective 5176413 and 5153914 were found by SBB. Each of these points were then passed to BARON as a starting point. Starting from the first point, BARON found it to be infeasible and returned no feasible solution after 1000 seconds. BARON recognized the second point as feasible and declared it optimal, which is correct under the 1% relative optimality tolerance we used. These results demonstrate that BARON is dependent on being able to use a local method to find a feasible solution from the starting point. In general BARON is unable to solve this problem.

We should note though that BARON is able to handle small instances of this problem that we constructed to check the validity of our code. On large scale instances the number of linear programming subproblems that must be solved by BARON becomes to large.

In most cases SBB found a feasible point within 60 seconds and concluded that this point was optimal to 1% within 200 seconds. However this assessment of the optimality gap is based on a nonconvex NLP relaxation, which is solved by a local method, and thus may be incorrect. In fact when randomizing the starting point, the objective value determined by SBB varied randomly so that in some cases an upper bound was claimed that was violated by a feasible point found with a different starting point. On instances A_1, A_2, A_3, A_4, A_5 the variation was within 3% of the optimal value. However on instance A_6 SBB returned objective values varying by up to 20% of the optimal value, and with some starting points the initial nonconvex NLP relaxation termminated with local infeasibility. This susceptibility of SBB to local optima appears to be exacerbated in instances where relatively low prices and consequently low profit margins make it all the more important to

allocate high quality coal to those contracts that pay the most for it. It is not hard to imagine that such low profit margins will become more common as coal loses market share to natural gas as a fuel for electric power generation. Thus it is important to have a solution method that reliably obtains a globally optimal solution.

Our method does just that. For example, on instance A_5, with 6 rounds of refinement taking 46 seconds, the MIP relaxation model obtained a dual bound of 5380616, and the MIP heuristic model obtained a primal bound of 5380382. By contrast SBB took over 200 seconds to find a feasible point with objective 5369496, and declared an upper bound of 5375833 on the true optimal objective. This declared upper bound is derived from local solutions of a nonconvex NLP relaxation and may be invalid if some of those local solutions are not globally optimal. Our feasible solution with objective 5380382 indicates that this upper bound does suffer from this problem of nonglobality.

We now consider a test of our MIP-based solution method on all six instances. The results of a single iteration (i.e. with no refinement) are shown in table (3.1). For each instance we display the objective of the feasible solution obtained by the primal heuristic, the rigorous upper bound obtained by the relaxation, the relative percent optimality gap between the heuristic and the relaxation, and the execution time Δt in seconds. On all instances except A_6, we achieved a

	heur	relax	gap	Δt
A_1	5168995	5190506	0.416	6.86
A_2	6872116	6893627	0.313	6.65
A_3	5168995	5190506	0.416	7.30
A_4	5358569	5382657	0.449	3.23
A_5	5367742	5393315	0.476	3.45
A_6	1292343	1308598	1.257	3.85

Table 3.1 One iteration of the MIP-based algorithm

relative optimality gap of less than 0.5% in 3 to 7 seconds. On instance A_6 the gap is somewhat larger, at 1.25%, but still quite reasonable, and the execution time is still only 3 seconds.

For comparison, the results of the SBB algorithm are shown in table (3.2). Three different methods of choosing a starting point were used. Method LL chooses a point to the lower left of the feasible set, i.e. the origin. Method UR chooses a point to the upper right. Method RAND chooses a starting point at random, typically not a feasible point. The objective value of the first feasible point found by SBB, the relative percent gap to the feasible solution obtained by one iteration of our method, and the execution time Δt required by SBB to obtain a first feasible point are given for each instance. On most instances, SBB had some trouble even finding a feasible point, and when

	LL			UR			RAND		
	obj	gap	Δt	obj	gap	Δt	obj	gap	Δt
A_1	5153914	0.291	39	5156777	0.236	41	5153914	0.291	35
A_2	6857062	0.219	36	6857035	0.219	31	$-\infty$	∞	38
A_3	$-\infty$	∞	10	5082107	1.680	35	5082107	1.680	53
A_4	5117503	4.498	43	5117503	4.498	41	$-\infty$	∞	14
A_5	5066765	5.607	38	5068728	5.570	34	$-\infty$	∞	21
A_6	1256581	2.767	33	1262511	2.308	37	$-\infty$	∞	18

Table 3.2 Results using the SBB algorithm

it was able to find a feasible point, this point was significantly worse than the solutions obtained by our method, and it took a much longer time to find this point than our method. Our method is always able to find a feasible point.

To show the results of our refinement method, we consider instance A_6, which appears to be the most challenging instance for our method. The results of 4 iterations of our MIP-based algorithm (i.e. with 3 rounds of refinement) are shown in table (3.3). We display the objective of the feasible solution obtained by the primal heuristic, the rigorous upper bound obtained by the relaxation, the relative percent optimality gap between the heuristic and the relaxation, and the cumulative time

Δt from the start of the algorithm to the end of each iteration. At each iteration both the primal

iter	heur	relax	gap	Δt
1	1292343	1308598	1.257	3.85
2	1295929	1303520	0.585	10.03
3	1297644	1301552	0.301	18.07
4	1298537	1300884	0.180	29.32

Table 3.3 Refinement in the MIP-based algorithm on instance A_6

heuristic solution and the relaxation are improved, the gap decreasing by approximately a factor of 2. Each iteration takes longer than the one before but not prohibitively so.

Our method performs at least as well as all the general pupose methods surveyed here on all six instances and in many cases performs substantially better. It gives rigorous upper and lower bounds on the optimal value and thus avoids the local minima and local infeasibility that affect methods such as SBB and DICOPT. And because it relies on a sequence of relatively few MIP subproblems it can take advantage of robust MIP solvers and handle much larger problems than methods such as BARON and is faster than methods relying on NLP solvers, as SBB does.

3.5 Conclusions and further work

One natural step to take for further work in the direction of this research is to quantify the nonconvex difficulty of the problem. This can be done by evaluating the objective and feasibility at a random sample of points in $\{(w, x, y) : (w, x, y, f(x, y)) \in C\}$. Quantitative measures of local nonconvexity and Lipschitz continuity can be computed for each such point by means of its nearest neighbors. These measure might give insight into the circumstances where one solution method or another performs particularly well. Indeed random sampling can be seen as a solution method in its own right as a randomized version of the naive grid evaluation method and should be considered alongside the other methods.

Also, the poor performance of general purpose global solution methods bears further investigation. In principle these methods use similar ideas to our own method, essentially solving convex relaxations to obtain bounds and then branching on spatial refinements of the problem.

The algorithm would be much improved by efficient solution of the refined MIP subproblem following the solution of the previous relaxation or primal heuristic. This would require adding a certain number of rows and columns to a previously optimized MIP formulation. We are unaware of any technique for efficient reoptimization of a MIP on adding both rows and columns, and such a technique would have substantial value beyond our method for structured nonconvex optimization problems.

Lastly we have not treated the original application to strategic planning of coal mine quantity and quality in all its complexity. In general the problem has multiple qualities, some with more complicated definitions than the ratio we have used here, constraints enforcing quality consistency over time.

In this chapter we have demonstrated a class of structured nonconvex optimization problems that can be treated advantageously by a discretization technique. We gave conditions guaranteeing that this technique gives rise to both a relaxation and a primal heuristic, and we showed how refinement of the discrete grid drives the resulting optimality gap to 0. We then applied this method to a problem of practical importance and showed that our method can signifcantly outperform general purpose methods for nonconvex optimization.

Chapter 4

Mixed equilibrium for nonconcave payoffs in continuous strategy spaces

4.1 Introduction

This chapter gives an introduction to Nash equilibrium in games. We focus on the fundamental concepts of equilibrium in pure strategies and in mixed strategies, with finite or continuous strategy sets, and with payoff functions that may be continuous or otherwise and concave or otherwise. We give particular attention to conditions for existence of equilibrium and numerical methods for computing or approximating an equilibrium. The theoretical material is well-established, and we follow closely the treatment in [14].

Our goal in a later chapter is to develop an equilibrium model to analyze a particular phenomenon in electricity markets. This model can be seen as a generalization to discontinuous and nonconcave payoffs of the classical Cournot equilibrium model. We therefore use the Cournot model in this chapter to illustrate some of the concepts that we review here.

Finally, we lay out the numerical method that we have developed for the purpose of solving this equilibrium model of electicity markets with discontinuous and nonconcave payoffs. As the theoretical results guarantee the existence of only a mixed equilibrium, this is the kind of equilibrium that we seek in our numerical method. Our numerical method is essentially an implementation of the method of proof of the most nearly applicable existence result.

4.2 Basic concepts, definitions and notation

In a noncooperative game, each of finitely many players seeks to maximize a payoff. Each player chooses a strategy, and the payoff depends on the strategies chosen by all players. An equilibrium strategy profile is a specification of the strategies chosen by all players so that no player can increase its payoff by deviating from its chosen strategy while the other players maintain their chosen strategies.

A typical application is the Cournot equilibrium model. Several suppliers of a single good each choose a quantity to produce and sell, knowing that the more they produce, the lower will be the market price, and the higher will be their total cost of production. Thus the quantity chosen by each supplier is its strategy, and the profit, consisting of the revenue from sale minus the cost of production, is the payoff. This model of market equilibrium differs from that of competitive equilibrium, in which suppliers do not consider the effect of their supply quantity on the market price but rather take the price as given. The Cournot model is the starting point for study of market power, as larger suppliers are able to exercise greater influence on the price than smaller suppliers are. In the competitive equilibrium model, the outcome, consisting of the quantity supplied by each supplier and the total quantity consumed, is optimal in that it achieves the maximum net social benefit to all suppliers and consumers. The Cournot outcome can be contrasted with this optimal, competitive outcome to quantify the social cost imposed by market power.

Let I denote the (finite) set of players. For each $i \in I$, let S_i denote the set of strategies available to player i. Let $S = \Pi_i S_i$ denote the set of strategy profiles. Let $S_{-i} = \Pi_{s' \neq i} S_{i'}$ denote the set of strategy profiles with the strategy of player i unspecified. For any strategy profile $s \in S$ let $s_{-i} = (s_{i'})_{i' \neq i}$ denote the corresponding strategy profile with the strategy of player i unspecified.

We will need to consider probability distibutions on S, so we make some mathematical assumptions to that end. Assume the strategy sets S_i are topological spaces. A mixed strategy σ_i of player i is a Borel probability measure on S_i. Let Σ_i denote the set of mixed strategies of player i. Define the set of mixed strategy profiles to be set of product measures $\Sigma = \Pi_i \Sigma_i$, so that $\sigma(s) = \Pi_i \sigma_i(s_i)$.

Let $\Sigma_{-i} = \Pi_{i' \neq i} \Sigma_i$. For any $\sigma \in \Sigma$ define $\sigma_{-i} = (\sigma_{i'})_{i' \neq i}$. Thus Σ is contained in the set of Borel probability measures on S, and Σ_{-i} is contained in the set of Borel probability measures on S_{-i}.

Let $u_i : S \to \mathbb{R}$ denote the payoff function of player i. Define $u : S \to \mathbb{R}^I$ by $u = (u_1, \ldots, u_I)$. Extend the payoff functions to functions $\omega_i : \Sigma \to \mathbb{R}$ by $\omega_i(\sigma) = E_s u_i(s)\sigma(s)$. And define $\omega : \Sigma \to \mathbb{R}^I$ by $\omega = (\omega_1, \ldots, \omega_I)$.

Define the set-valued pure reaction function $r_i : S_{-i} \to S_i$ of player i by

$$r_i(s_{-i}) = \operatorname{argmax}_{s_i} u_i(s_i, s_{-i}) \tag{4.1}$$

Define $r : S \to S$ by $r(s)_i = r_i(s_{-i})$. Define the set-valued mixed reaction function $\rho_i : \Sigma_{-i} \to \Sigma_i$ of player i by

$$\rho_i(\sigma_{-i}) = \operatorname{argmax}_{\sigma_i} \omega_i(\sigma_i, \sigma_{-i}) \tag{4.2}$$

Define $\rho : \Sigma \to \Sigma$ by $\rho(\sigma)_i = \rho_i(\sigma_{-i})$.

A pure equilibrium is a fixed point of r. A mixed equilibrium is a fixed point of ρ. Every pure equilibrium is obviously a mixed equilibrium, so an equilibrium, with out any further specification, will refer to a mixed equilibrium. By a theorem of J. Nash (1950), if S is finite, then the game has an equilibrium.

To put the problem of Nash equilibrium into the language of mathematical programming, a pure equilibrium is a point $s \in S$ so that each s_i solves the optimization problem

$$\max_{s_i \in S_i} u_i(s_i, s_{-i}) \tag{4.3}$$

and a mixed equlibrium is a point $\sigma \in \Sigma$ so that each σ_i solves the optimization problem

$$\max_{\sigma_i \in \Sigma_i} \omega_i(\sigma_i, \sigma_{-i}) \tag{4.4}$$

In the example of Cournot equilibrium, the strategy s_i of supplier i is the quantity of the good that it produces and sells. If $p(q)$ denotes the inverse demand function and $c_i(s_i)$ is the cost of production of supplier i, then the payoff to supplier i is the profit

$$u_i(s_i) = p\left(\sum_{i'} s_{i'}\right) s_i - c_i(s_i)$$

4.3 Finite strategy spaces and equilibrium computation

If S is finite then the mixed equilibrium condition can be formulated using complementarity for convenient computation by the methods of mathematical programming. In this case the mixed strategy set of player i is a polyhedron $\Sigma_i = \{\sigma_i \in \mathbb{R}_+^{S_i} : 1^T\sigma_i = 1\}$. And the mixed utility function of player i is $\omega_i(\sigma) = \sum_s u_i(s)\sigma(s)$. Define $\phi : \Sigma \to \mathbb{R}^S$ by

$$\phi_{is_i}(\sigma) = -\frac{d\omega_i}{d\sigma_i(s_i)} = -\sum_{s_{-i}} u_i(s_i, s_{-i})\sigma_{-i}(s_{-i})$$

Then σ is an equilibrium if and only if it is a solution of the variational inequality $VI(\phi, \Sigma)$, i.e.

$$0 \in \phi(\sigma) + N_\Sigma(\sigma) \tag{4.5}$$

Since Σ is polyhedral, the VI can be formulated as a complementarity problem (CP). For this introduce a Lagrange multiplier λ_i for the normalization constraint of player i:

$$\begin{array}{rlcll} 0 \leq & \phi_{is_i}(\sigma) - \lambda_i & \perp & \sigma_i(s_i) & \geq 0 \\ 0 = & 1^T\sigma_i - 1 & \perp & \lambda_i & \text{free} \end{array} \tag{4.6}$$

The (CP) formulation (4.6) can be solved numerically, e.g. by PATH.

The data requirements of this formulation may be substantial: $u(s)$ is an array of $|I| \times |S| = |I| \times \Pi_i|S_i|$ real numbers, each of which may need to be computed by a simulation or an optimization. Indeed in practice this computation has been more time-consuming than the subsequent computation of an equilibrium.

4.3.1 2-person special case

When $|I| = 2$ a number of special conditions hold and allow a simpler treatment, smaller formulation, or more reliable algorithm. First ϕ is linear, and PATH is guaranteed to solve the problem, though it may run in exponential time. When $|I| > 2$, ϕ is nonlinear, so (4.6) is solved iteratively by linearizing. The progress toward a solution of the nonlinear problem is guided by a merit function such as the Fischer-Burmeister function. This merit function may be nonconvex, so PATH may terminate without a solution to (4.6), even though a solution does exist. To guarantee

finding a solution with $|I| > 2$, one must use a triangulation algorithm to compute a fixed point of ρ.

Second if the problem is symmetric, then there is a symmetric solution, and a symmetric CP formulation is half the size of the nonsymmetric formulation. Suppose that the strategy spaces are equal, i.e. $S_1 = S_2$, and that $u_1(s_1, s_2) = u_2(s_2, s_1)$. Then the problem is said to be symmetric. And there exists a mixed equilibrium σ that is symmetric in that $\sigma_1 = \sigma_2$. And we can give a CP formulation of the necessary and sufficient first-order optimality conditions for a symmetric mixed equilibrium.

For this, let S_0 denote the common value of S_1 and S_2. Let Σ_0 be the space of probability measures on S_0. Define $u_0 : S_0 \to \mathbb{R}$ by setting $u_0(s_0)$ to be the common value of $u_1(s_0, s_0)$ and $u_2(s_0, s_0)$. Define $\phi_0 : \Sigma_0 \to \mathbb{R}^{S_0}$ by

$$\phi_{0s_0}(\sigma_0) = -\sum_{s' \in S_0} u_0(s_0, s')\sigma_0(s')$$

Then $\sigma = (\sigma_0, \sigma_0)$ is a symmetric mixed equilibrium if σ_0 solves $VI(\phi_0, \Sigma_0)$, and this VI can be formulated as a CP:

$$\begin{array}{rlcll} 0 \leq & \phi_{0s_0}(\sigma_0) - \lambda_0 & \perp & \sigma_0(s_0) & \geq 0 \\ 0 = & 1^T\sigma_0 - 1 & \perp & \lambda_0 & \text{free} \end{array} \tag{4.7}$$

This CP formulation (4.7) has only half as many variables as the nonsymmetric formulation (4.6).

Finally when $u_1 + u_2$ is constant, the problem is called a zero-sum game, and in this case, with $|I| = 2$, a linear programming formulation can be given. We do not make any particular use of this formulation, as our intended application is not zero-sum, but we record it here for completeness. (Or maybe we should discard this paragraph.) For this, we may assume by addition of appropriate constants, that $u_1 \geq 1$ and $u_2 = -u_1$. Then the following linear program with primal variables σ_1 and dual variables σ_2 is bounded and feasible:

$$\begin{array}{ll} \min_{\sigma_1 \geq 0} & \sum_{s_1} \sigma_1(s_1) \\ & \sum_{s_1} u_1(s_1, s_2)\sigma_1(s_1) \geq 1 \quad (\perp \sigma_2(s_2) \geq 0) \quad \forall s_2 \end{array} \tag{4.8}$$

And any solution (σ_1, σ_2) of (4.8), upon normalization of σ_1 and σ_2 each to have total probability 1, yields a mixed equilibrium.

4.3.2 Nonuniqueness

Nonuniqueness of equilibrium raises several questions, some of which might best be treated in different sections of this chapter. With multiple equilibria, it is not clear how different players would come to agree on which equilibrium to play, so the actual outcome might not be any of the equilibria. And even if we knew that the outcome would be an equilibrium but we did not know which one, the model would still have little predictive value.

For applications, it would be desirable to be able to diagnose nonuniqueness of equilibrium from a solution of the CP formulation. In this finite game context we can easily dispose of two reasonable conjectures in this direction. First, in the CP formulation (4.6) we might look for strict complementarity as either a sufficient or necessary condition for uniqueness of equilibrium. In fact it is not sufficient. For an example, consider the prisoners' dilemma game. Two players each choose between strategies s_1 (cooperation) and s_2 (defection). If both cooperate then both receive a payoff of 0, if both defect then both receive a payoff of -2, and otherwise the defector receives a payoff of -1 while the cooperator receives a payoff of -3. There are two pure equilibria, the first having both players cooperate, and the second having both defect. But each of these equilibria satisfies strict complementarity.

Second, we note that the optimal reaction to a given strategy profile may be nonunique even though there may be only one equilibrium. An example is provided by the game of matching pennies, in which two players each show show one side of a penny, choosing a strategy from the strategy space $\{\text{heads}, \text{tails}\}$. If both players choose the same strategy, then player 1 wins both pennies, otherwise player 2 wins both. A mixed equilibrium is given by $\sigma_1 = (0.5, 0.5)$ and $\sigma_2 = (0.5, 0.5)$, and it is the only equilibrium of this game. However the optimal reaction to this equilibrium is $\rho_1 = \{(1-t, t) : 0 \leq t \leq 1\}$ and $\rho_2 = \{(1-t, t) : 0 \leq t \leq 1\}$.

4.4 Euclidean strategy spaces with convexity: Pure equilibrium

In the Cournot example and in many other models, the strategy spaces are not finite or even discrete. Still, under certain strong conditions, a pure equilibrium is guaranteed to exist, by a

theorem independently credited to Debreu (1952), Glicksberg (1952), and Fan (1952) (F&T Thm. 1.2):

Suppose each S_i is a compact convex nonempty subset of a finite-dimensional Euclidean space, and each u_i is continuous in s and quasiconcave in s_i. Then there exists an equilibrium in pure strategies.

Typically a reasonable interpretation of a model can be given so that the strategy sets may be assumed to be compact. In the Cournot model, we may assume that all firms produce nonnegative quantities and have some finite production capacity. It is also standard to assume that production costs are convex and that the inverse demand function is decreasing. Thus the Cournot model has a pure equilibrium.

4.4.1 An intuitively appealing method for computation of pure equilibrium

A pure equilibrium is a fixed point of the reaction function r, so a reasonable algorithm to find a fixed point is to iterate r, choosing $s^{n+1} \in r(s^n)$. The computation of a point in $r_i(s^n)$ for each player i amounts to solving an optimization problem

$$\max_{s_i} u(s_i, s^n_{-i})$$

Thus this algorithm can take advantage of mature algorithms for optimization and allows obvious parallelization.

But a guarantee of convergence can be given only if r is contractive, and this condition is substantially more restrictive than the main sufficient condition for existence of a solution, which is essentially a set-valued generalization of continuity of r. In a model in which the underlying real-world phenonomenon is the equilibrium result of a process of reaction and strategy adjustment by multiple agents, it would seem reasonable to expect the iterated reaction algorithm to converge, so that a well-designed model of an actual pure equilibrium has to be solvable by this algorithm.

4.4.2 Differentiable payoffs

If the payoff functions u_i are differentiable, then we may write first-order optimality conditions for each player as the variational inequality

$$0 \in -\frac{du_i}{ds_i}(s) + N_{S_i}(s_i) \tag{4.9}$$

where N_X denotes the normal cone operator for a convex set X. These variational inequalities (4.9) are thus a necessary condition for a pure equilibrium. And if the payoffs u_i are concave in s_i, then they are a sufficient condition. In this case we may seek to solve the VI formulation by a mathematical programming algorithm, such as gradient descent or, if the strategy sets S_i are polyhedral, PATH. Without further assumptions neither of these algorithms is guaranteed to converge to a solution. However PATH has good practical performance even on problems not meeting known sufficient conditions for convergence, And the main sufficient condition for convergence of gradient descent, namely monotonicity of the variational inequality (4.9), is closely related to the sufficient condition of contractivity for the iterated reaction algorithm. Thus our comment in that context on the influence of model design on the convergence of iterative algorithms applies here.

4.4.3 Weaker continuity assumptions

A weakening of the continuity assumption in the existence theorem is enabled by a theorem of Dasgupta and Maskin (1986) (F&T Thm 12.3) But the same convexity assumptions are required to ensure a pure equilibrium:

Suppose each S_i is a nonempty convex compact subset of a finite dimensional Euclidean space, and each u_i is quasi-concave in s_i, upper semi-continuous in s, and max-continuous in s_{-i}. Then there exists an equilibrium in pure strategies.

4.5 Euclidean strategy spaces without convexity: Mixed equilibrium

In many models the convexity properties required for existence of a pure equilibrium are lacking. Either the strategy sets S_i are not convex, or the payoff functions u_i are not concave in s_i. But

if the payoff functions are continuous, then a mixed equilibrium is guaranteed (Glicksberg 1952, F&T Thm. 1.3):

Suppose each S_i is a nonempty compact subset of a metric space, and each u_i is continuous. Then there is a mixed equilibrium.

One proof of this result is constructive, so we describe the technique as we will use it to build numerical approximations of mixed equilibrium in this type of model in a later chapter. Essentially the strategy space S is approximated by a finite discrete mesh, and the payoff function u is evaluated at the points of the mesh, generating a finite approximation of the original model which is then guaranteed to have an equilibrium by the Nash theorem. A sequence of meshes, with increasing fineness, generates a sequence of equilibria. This sequence of equilibria has a weakly convergent subsequence in Σ, and the continuity of the payoffs guarantees that the limit point is a mixed equilibrium.

Specifically, for $\epsilon > 0$, we say that a finite set $Z_i \subset S_i$ approximates S_i with tolerance ϵ if every point of S_i is within ϵ of a point of Z_i, and in this case we say that $Z = \Pi_i Z_i$ approximates S with tolerance ϵ. Let $Z^n, n = 1, 2, \ldots,$ be a sequence of finite approximations of S with tolerance converging to 0. Then for each n, the discrete game on strategy set Z^n has a mixed equilibrium σ^n. Then the sequence σ^n has an accumulation point σ^* with respect to weak convergence. And in turn σ^* is an equilibrium of the original game on S.

The assumption of continuity can be weakened, according to a theorem of Dasgupta and Maskin (1986) (F&T Theorem 12.4), and this theorem is proved by the same constructive method as the result of Glicksberg:

Suppose that for each i, S_i is a closed interval of $\mathbb{R}$, and for all $s_{-i} \in S_{-i}$, $u_i(s_i, s_{-i})$ is bounded and weakly lower semi-continuous in s_i. Suppose $\sum_i u_i$ is upper semi-continuous. Suppose that for each pair (i_1, i_2) of players there exist finitely many functions $f^m_{i_1 i_2} : S_{i_1} \to S_{i_2}$, indexed by m, that are one-to-one and continuous, so that for each i the set $s^* * (i)$ of discontinuity points of u_i is contained in the set

$$S^*(i) = \{s \in S : \text{ there exist } i' \neq i \text{ and } m \text{ so that } s_{i'} = f^m_{ii'}(s_i)\}$$

Then there is a mixed equilibrium.

4.6 Numerical algorithm for mixed equilibrium on Euclidean strategy spaces

The constructive method of proof of the existence theorems for mixed equilibrium in games with Euclidean strategy spaces suggests a numerical algorithm to find an approximate equilibrium. We construct finitely many approximations Z^n of the strategy spaces S and solve the complementarity formulations (4.6) for mixed equilibria σ^n of the resulting finite games on strategy spaces Z^n. One of the σ^n will be taken to be an approximate equilibrium of the original game on Euclidean strategy space S. To specify an implementation of this algorithm we need only discuss (1) how to construct the Z^n and (2) how to decide which σ^n is returned as the approximate solution. The solution of the finite game approximations may be carried out by any numerical solver suitable for linear complementarity problems.

In all the implementations that we consider, each approximate strategy space Z is constructed in the same way. We choose $\epsilon > 0$ and construct a finite approximation Z of the strategy space S with tolerance ϵ by taking uniform grids in each coordinate dimension. Specifically, suppose

$$S_i \subset [\underline{s}_i, \overline{s}_i] = \Pi_{j=1}^{n_i} [\underline{s}_{ij}, \overline{s}_{ij}] \tag{4.10}$$

Then define

$$Z_i = S_i \cap \Pi_{j=1}^{n_i} Z_{ij} \tag{4.11}$$

where

$$Z_{ij} = \{z_{ijk} : k = 1, \ldots, K_{ij}\} \tag{4.12}$$

and

$$z_{ijk} = \underline{s}_{ij} + k\epsilon \tag{4.13}$$

and

$$K_{ij} = \lceil (\overline{s}_{ij} - \underline{s}_{ij})/\epsilon \rceil \tag{4.14}$$

In the very simplest implementation we only construct one approximate strategy space Z, and then there is no difficulty selecting an approximate equilibrium. Otherwise, it is sensible to consider a decreasing sequence of values ϵ^n of ϵ, terminating the algorithm with the approximate

solution σ^n when $\|\sigma^n - \sigma^{n-1}\|$ is below a given convergence tolerance. The norm used here is

$$\|\mu\| = \int d|\mu| \tag{4.15}$$

As we have defined them, the approximate equilibria σ^n are finitely supported, and the supports of consecutive σ^n will have few points in common, so this norm will not give a reasonable notion of convergence. A better measure of convergence is obtained by interpreting σ^n as a piecewise uniform probability measure assigning uniform probability on each rectangle with vertices taken from adjacent points in Z^n, so that $\sigma^n(z)$ is the probability assigned to the rectangle whose upper left vertex is z. That is, $\sigma^n(z_{i1k_1}, \ldots, z_{in_ik_{n_i}})$ is the probability assigned to the rectangle $[z_{i,1,k_1-1}, z_{i,1,k_1}] \times \cdots \times [z_{i,n_i,k_i-1}, z_{i,n_i,k_i}]$.

4.7 Conclusion

In this chapter we have reviewed the basic concepts of noncooperative game thoery. We have presented a computational scheme for approximating a mixed equilibrium of a game with continuous strategy spaces and nonconcave and discontinuous payoffs. This scheme is essentially a numerical implementation of the method of proof of existence of an equilibrium in such a game. And we discussed methods to assess the convergence of the algorithm. In the next chapter we will apply this algorithm to a specific problem from the study of market power in wholesale electricity markets with nonconvex costs.

Chapter 5

Application to analysis of electric power market rules with unit commitment and strategic behavior

5.1 Introduction

In this chapter we apply the theory of mixed Nash equilibrium in infinite strategy spaces without convexity to compare various proposed payment rules and market structures in wholesale electric power markets.

Generally in a market for a single good, if all participants - producers and consumers - behave as price-takers, then the outcome will be efficient. That is, the total quantity exchanged of the good will be produced at least cost and consumed to greatest benefit. This price-taking, or competitive, behavior that is crucial to efficiency is a theoretical ideal that may in any given market prevail to a greater or lesser degree. Small producers are practically unable to behave in any way other than as price-takers, simply observing the price and deciding how much to produce so as to maximize their profit. But large producers may be able to predict the effect of their production decision on the market price and include this prediction in their production decision thus increasing their profit beyond the competitive level. The Cournot equilibrium models exactly this strategic behavior. This model is especially appropriate for markets with just a few large suppliers.

Wholesale electric power markets typically have just a few large suppliers, and the Cournot model is good starting point that is often used to quantify market power here [1]. But because of the physical characteristics of electric power as an economic good, most markets have a significant regulatory structure, and suppliers typically have production costs with some degree of nonconvexity. The regulatory structure and nonconvex costs invalidate the Cournot model. We use

the framework of noncooperative games with mixed equilibria to analyze the potential for market power in these more complex settings.

Wholesale markets for electric power show a wide variety of regulatory strucutres, but we focus on an abstract structure that approximates many of the existing markets. In this abstract regulatory structure, a system operator (SO) solicits bids for generation from generation companies (GENCOs) and consumption estimates from distribution companies (DISTCOs). A supply bid is a function indicating, for each level of generation within the physical capability of the GENCO, its marginal cost of generation, or equivalently the total cost of generation. The SO clears the market according to a specified algorithm, dispatches generation and sets prices to compensate GENCOs and charge DISTCOs according to a payment rule. The market algorithm is typically designed so that the dispatch meets the demand at minimum cost. The cost of generation is quantified by the bids submitted by the GENCOs, so if these bids are truthful then the dispatch is socially optimal, i.e. efficient.

With the wide variety of cost structures, bid structures, and payment rules that we wish to consider, we cannot expect the bids to be truthful. Instead we assume that each GENCO bids so as to maximize its profit, leading naturally to a noncooperative game. Each GENCO a player. The bid structure determines the strategy spaces. The true cost functions and the payment rule determine the payoffs.

If the cost functions of the GENCOs are convex, and the marginal cost curves that they bid to the SO are required to be increasing, and the GENCOs are paid at a price equal to the system-wide marginal cost, then the market can be modeled by the Cournot equilibrium. But other payment rules have been considered even in this convex case, such as pay-as-bid, in which GENCOs are paid what their bid indicates is their total cost of generation, not their marginal cost. And GENCOs generally do not have convex costs, and for this reason their bids may be allowed to be nonmonotone. And nonmonotone bids have led to still other payment rules. We are able to analyze the strategic behavior incentive of all these combinations of cost structure, bid structure and payment rule.

In addition to promoting efficient production of electricity in the short term, bid structures and payment rules may be designed with two other objectives in mind. First we might hope to send appropriate signals for long term investment in production capacity. If a payment rule or bid structure gives a systematic advantage to one cost structure over another, then GENCOs of the first type might be expected to invest in more capacity than is socially optimal. Such a systematic advantage could be explored in our mixed equilibrium framework, but we do not do so at present.

Another reasonable objective for a bid structure and payment rule is to promote short term equity. For example, two payment rules may result in the same dispatch from all GENCOs and the same consumption by DISTCOs, but one rule may require much higher payment from each DISTCO and to each GENCO. In this case we would say that the two rules are identical in efficiency but quite different in equity. In our framework we can easily make such equity comparisons. In particular we can assess revenue adequacy, i.e. whether the payments from DISTCOs suffice to cover the payments to GENCOs.

5.2 Mathematical model

We now describe our equilibrium model of GENCO bidding. We begin by formalizing the strategy spaces and payoffs of the game itself, including the cost structure of each GENCO and the allowable bids, the optimal dispatch procedure followed by the SO, and the various payment rules that we will consider.

5.2.1 Market participants

For the purpose of this model, we consider a simplified wholesale electric power market. We assume there are a number of strategic GENCOS making up the players $i \in I$ of a noncooperative game. These GENCOs are strategic in that they are able to bid cost functions that differ from their true costs in order to maximize their profits.

Demand is represented by a fixed consumption quantity. For increased model realism and computational robustness, we allow the demand quantity to be stochastic with a distribution that is known by all participants but is realized only after bids are placed.

We include a so-called competitive fringe of smaller GENCOs that bid their true costs. The elasticity of competitive supply can be seen as subtracting from the otherwise inelastic demand, and this feature can also be expected to make the model more robust.

5.2.2 Generator cost structure

The strategic suppliers i supply quantities q_i subject to capacities $0 \leq q_i \leq \overline{q}_i$. They face true marginal costs

$$MC_i^0(q_i) = a_i^0 + b_i^0 q_i \tag{5.1}$$

and true commitment costs c_i^0 if $q_i > 0$ and 0 otherwise. We assume $b_i^0 > 0$ for partial convexity. To evaluate the importance of nonconvex costs, we may compare with the fully convex case of $c_i^0 = 0$. This cost structure bears some resemblance to that of a single electric generator. In reality a typical GENCO operates many generators so the cost structure of a GENCO is rather more complicated. We focus on this highly simplified structure first to keep the computational requirements to a reasonable level but also to clarify the effects of nonconvex costs and different payment rules on economic efficiency. This twofold justification holds for many simplifications of reality featuring in our model.

5.2.3 Generator bid structure

Strategic suppliers bid marginal costs

$$MC_i(q_i) = a_i + b_i q_i \tag{5.2}$$

and commitment costs c_i with $\underline{a}_i \leq a_i \leq \overline{a}_i$, $\underline{b}_i \leq b_i \leq \overline{b}_i$, and $\underline{c}_i \leq c_i \leq \overline{c}_i$. Thus the strategy of player i is a triple (a_i, b_i, c_i) contained in the strategy space $[\underline{a}_i, \overline{a}_i] \times [\underline{b}_i, \overline{b}_i] \times [\underline{c}_i, \overline{c}_i]$. Note that the strategy spaces are compact convex nonempty subsets of Euclidean spaces. To disallow nonconvex bids, we may set $\underline{c}_i, \overline{c}_i = 0$. To ensure partial convexity we assume $\underline{b}_i > 0$.

5.2.4 Demand and competitive supply

The fixed demand quantity is denoted by q_d and is a normal random variable with mean $\overline{q}_d$ and standard deviation $\hat{q}_d$. The competitive supply fringe has the same cost structure as a single strategic supplier but with no commitment cost and no bounds on supply. For competitive supply q_c, the marginal cost is $MC_c(q_c) = a_c + b_c q_c$ with $q_c > 0$. Typically we set $a_c = 0$ and b_c rather large, so that the competitive dispatch is near 0, Competitive supply thus acts as a regularizer in the model, ensuring that any realized value of demand can be met, though possibly at great cost.

5.2.5 Dispatch procedure

Given supply bids and a specification of competitive supply and a realization of demand, the SO dispatches generation q_i from supplier i and q_c from the competitive fringe and q_d to demand, choosing this dispatch so as to minimize the total apparent cost of generation. That is, the SO solves the following minimum cost unit commitment problem:

$$\begin{array}{ll} \min_{q_i, q_c, x} & \sum_i \left(c_i x_i + a_i q_i + \frac{1}{2} b_i q_i^2 \right) + a_c q_c + \frac{1}{2} b_c q_c^2 \\ \text{s.t.} & \sum_i q_i + q_c = q_d \\ & 0 \leq q_i \leq \overline{q_i} x \forall i \\ & q_c \text{ free} \\ & x_i \in \{0, 1\} \forall i \end{array} \tag{5.3}$$

yielding optimal commitment x^* and dispatch q^*.

5.2.6 Payment rules

Given the commitment x^* and dispatch q^*, the market participants are compensated or charged as appropriate according to a payment rule. Many payment rules operate by determining a price for electric power or possibly also for commitment of generators. The typical pricing mechanism

is represented by another optimization problem solved by the SO:

$$\begin{array}{lll} \min_{q_i,x_i,q_c} & \sum_i \left(c_i x_i + a_i q_i + \frac{1}{2} b_i q_i^2\right) + a_c q_c + \frac{1}{2} b_c q_c^2 & \\ \text{s.t.} & \sum_i q_i + q_c = q_d^0 & (\perp p) \\ & x_i = x_i^* & (\perp r_i) \forall i \\ & 0 \leq q_i \leq \overline{q_i} x_i \forall i & \\ & q_c \text{ free} & \end{array} \tag{5.4}$$

yielding the same commitment x^* and dispatch q^* and also an energy price p^* and a commitment price r^* as Lagrange multipliers.

The dispatch and prices can be combined in a number of ways to create interestimg payment rules, some of which are quite common in wholesale electric power markets. In general the dispatch quantity q_i can be paid as bid by

$$a_i q_i + \frac{1}{2} b_i q_i^2$$

or as priced by

$$p q_i$$

The competitive dispatch can similarly be paid as bid or as priced. The commitment x_i can be paid as bid by

$$c_i x_i$$

or as priced by

$$r_i x_i$$

We identify four specific payment rules as sufficiently interesting and representative of existing rules in practice:

1. Price power only (PPO). All power generation is compensated at the price p, and commitment is not explicitly compensated.

2. As-bid (AB). Power generation and commitments are compensated as bid. The prices are not used.

3. Price power with uplift (PPU). Suppliers are compensated at the greater of as-bid costs and priced power generation. The commitment prices are not used.

4. Price power and commitment (PPC). Power generation is paid at the price p, and commitment of strategic supplier i is paid at the price r_i.

Rule (PPU) is the most representative of current practice in many power markets. Rule (PPO) is in effect in a few locations, and if GENCO costs were in fact convex, this rule would achieve the ideal of the Cournot equilibrium. Rule (AB) is often discussed but not used. Rule (PPC) has recently been proposed (O'Neill et al.) as a way of extending the efficiency of pricing with convex costs to the realistic case of nonconvex costs.

In each rule, we can then evaluate the payoff u_i to supplier i as the net profit. Since demand is random, so is the dispatch and the prices and therefore the payoffs, when evaluating the payoffs, we take the expected value.

The payoff under rule (PPO) is:

$$u_i^{\mathrm{PPO}} = pq_i - \left(a_i^0 q_i + \frac{1}{2} b_i^0 q_i^2 + c_i^0 x_i\right) \tag{5.5}$$

Under rule (AB) we have:

$$u_i^{\mathrm{AB}} = \left(a_i q_i + \frac{1}{2} b_i q_i^2 + c_i x_i\right) - \left(a_i^0 q_i + \frac{1}{2} b_i^0 q_i^2 + c_i^0 x_i\right) \tag{5.6}$$

Rule (PPU) yields:

$$u_i^{\mathrm{PPU}} = \max\left\{a_i q_i + \frac{1}{2} b_i q_i^2 + c_i x_i, pq_i\right\} - \left(a_i^0 q_i + \frac{1}{2} b_i^0 q_i^2 + c_i^0 x_i\right) \tag{5.7}$$

And with (PPC) we have:

$$u_i^{\mathrm{PPC}} = pq_i + r_i x_i - \left(a_i^0 q_i + \frac{1}{2} b_i^0 q_i^2 + c_i^0 x_i\right) \tag{5.8}$$

Since we are most interested in the total cost of generation as a proxy for inefficiency of the equilibrium outcome under any given payment rule, we note here that the total cost of generation is

$$TC = \sum_i \left(c_i^0 x_i + a_i^0 q_i + \frac{1}{2} b_i^0 q_i^2\right) + a_c q_c + \frac{1}{2} b_c q_c^2 \tag{5.9}$$

Note that this is different from the objective of the SO dispatch problem (5.3) as it uses the true cost parameters of the strategic suppliers, rather than their bids. Note also that this total cost of generation will differ from one compensation rule to the next not because the formula is any different, but because the different rules lead to different equilibrium bidding strategies on the parts of the strategic suppliers and thus different dispatches and commitments.

So far we have not discussed how demand should be charged. Typically, this is done using the power generation price p, so that the total charge paid by demand is pq_d. In this case it is important to determine whether the SO will generate enough revenue from demand to cover the payment to suppliers. This revenue adequacy question can also be addressed in our framework, but we have not done so at this time. In practice consumers ultimately pay all the costs incurred by the SO, possibly through a surcharge, but it is still worth considering whether the power price is adequate.

5.2.7 Special scenarios: Cost structures and bid structures

In addition to the four payment rules we have outlined, we will consider certain special scenarios determined by particular choices of cost structure, bid structure, and payment rule. First to recover the competitive equilibrium we may require all suppliers to bid their true cost functions. This is accomplished by setting $\underline{a}_i, \overline{a}_i = a_i^0$, $\underline{b}_i, \overline{b}_i = b_i^0$, $\underline{c}_i, \overline{c}_i = c_i^0$. Under this competitive equilibrium, payments to suppliers may still be made by any of the four payment rules we consider, but the choice of payment rule will only affect the profits to individual suppliers, not the dispatch and total cost of generation. Comparing this competitive equilibrium to the strategic equilibrium under the different payment rules, we can evaluate the contribution of market power to the efficiency differences between the payment rules. We will refer to this scenario as CE.

Second the PPO rule is sometimes found in combination with a prohibition on nonconvex bids, which we may model by setting $\underline{c}_i, \overline{c}_i = 0$. We will refer to this combination of bid structure and payment rule as PPO/CB.

Third to recover the Cournot equilibrium we may use the PPO rule and and prohibit nonconvex bids and futher assume all suppliers have convex costs by setting $c_i^0 = 0$. This Cournot scenario does not correspond to a rule that a regulatory authority can make, as it requires the condition that

all suppliers do not actually have nonconvex costs, and this may or may not hold, independent of a regulatory decision. Rather the Cournot model is useful in comparison with PPO/CB to understand the effect of nonconvex costs under market rules that do not consider nonconvexity at all.

5.2.8 Technical issues in evaluating dispatch and payment

With a given set of supply bids and a realization of demand, there is a technical issue that may arise in evaluating the power dispatch q and the commitment x and the prices p, r by solving the problems (5.3) and (5.4). Namely, these optimization problems may have nonunique solutions.

To begin with, there may be several different commitment vectors x^* for which (x^*, q^*) achieves the minimum cost in (5.3) for some q^*. It is not obvious how this should be resolved. We consider all such optimal commitment vectors x^*, and for each of them solve the pricing problem (5.4) to obtain the power dispatch and prices and ultimately the payoffs. We then evaluate the payoffs under this choice of bids and realization of demand as the average of the payoffs evaluated from all such optimal commitment vectors.

This way of resolving the nonuniqueness of the optimal commitment is defensible from the perspective of the actual operation of wholesale electric power markets. Ties can happen and the must be broken somehow. The SO must take some precaution to ensure that no market participant is systematically disadvantaged by the particular tiebreaking method, if only for fear of legal action. One reasonable method that meets this fairness requirement is to randomize the order in which the commitment variables listed in the algorithm solving (5.3). In the long run this method would behave like our abstract method.

Now, for each commitment vector x^*, there may be several different power dispatch vectors q^* so that (x^*, q^*) is optimal for (5.3). We prohibit this possibility by the requirement that all marginal cost bids are strictly increasing, i.e. $\underline{q}_i, q_c > 0$. Finally we note that the unboundedness and elasticity of the competitive fringe supply ensures that the prices are uniquely defined.

5.2.9 Stochastic demand

For several reasons we choose to model demand as a fixed random quantity q_d that is realized after the bids are received but whose distribution is known in advance by all market participants. First of course it seems more realistic that demand should be somehow uncertain. All participants in the power market have some ability to predict demand and to plan their own actions accordingly, but ultimately it is as random as the weather.

Second if the demand is deterministic, then the multiple bid parameters that we allow may be redundant, making the Nash equilibrium nonunique, This is known to occur in the Cournot model, making the bids of a_i and b_i redundant, and stochastic demand is a known remedy to this. In our more complicated model it is not clear which cost structures, bid structures and payment rules might suffer from such redundancy, so we simply avoid this difficulty be introducing stochastic demand from the outset.

Third we use stochastic demand as a mechanism to promote continuity of the payoff functions. With deterministic demand, we would expect to see payoffs jump when, for example, the demand is just enough to be served by exactly one GENCO, and two GENCOs are tied for the cheapest. Stochastic demand is intended partially to smooth these jumps. Though the existence theory of mixed equilibria in games with nondiscrete strategy spaces and nonconcave payoffs does provide for some discontinuity in payoffs, we do not see a way to check the sufficient condition in our model, either analytically or numerically. Therefore we add stochasticity to our game to promote continuity of payoffs. We can check numerically the influence of the scale of smoothing, as quantified by the standard deviation $\hat{q}_d$ of demand, on the continuity of payoffs.

In our model we approximate the normal demand distribution by a finitely supported distribution with a fixed number n_d of equiprobable atoms at equally spaced quantiles. All players now maximize the expected value of their profit under stochastic demand.

5.2.10 Necessity of mixed equilibrium

Evidently the Nash game model that we propose here has infinite strategy spaces S_i, and indeed they are convex compact polyhedra in $\mathbb{R}^3$. In our model the feature of unit commitment appears

certain generate discontinuity in the payoffs, but let us set this aside for the moment. If the payoffs u_i were continuous and quasiconcave in the strategies s_i, then we would be guaranteed a pure equilibrium, and a there are a number of different algorithms we might consider to find a pure equilibrium, potentially much more efficiently than enumerating a discrete mesh of strategies.

First, if we could formulate the derivatives of the payoffs algebraically, then we could model the pure equilibrium problem as a complementarity problem in a modeling system such as GAMS and solved with PATH. But we do not see any way to formulate the payoffs algebraically, let alone their derivatives.

In principle we could evaluate the payoffs and their derivatives at given points using finite difference approximations and attempt to follow a trajectory in the gradient field to a pure equilibrium. But even concavity of the payoffs does not guarantee that such trajectories do not cycle.

So even with continuity and concavity the only sure method of finding a pure equilibrium in this model is to discretize the strategy space and check each point to see if it is a pure equilibrium relative to the other discrete strategies thus enumerated. At this point we might as well expand our search to mixed equilibrium, which is guaranteed to exist even with nonconcavity and some discontinuity of payoffs.

5.2.11 Discrete approximation of strategy spaces

The infinite strategy spaces

$$S_i = [\underline{a}_i, \overline{a}_i] \times [\underline{b}_i, \overline{b}_i] \times [\underline{c}_i, \overline{c}_i]$$

are approximated by approximating each of the three dimensions individually and taking the Cartesion product of these one-dimensional finite approximations. That is, $[\underline{a}_i, \overline{a}_i]$ is approximated by

$$\{a_{i1}, \ldots, a_{in_a}\}$$

for some integer $n_a \geq 1$. The discretization points are the midpoints of n_a subintervals of equal length partitioning the interval. That is,

$$a_{ij} = \underline{a}_i + (j - 1/2)(\overline{a}_i - \underline{a}_i)/n_a$$

The b and c parameter spaces are similarly discretized into n_b points and n_c points. Similarly the quantiles chosen to approximate the demand distribution are the midpoints of n_d subintervals of length $1/n_d$ partitioning the interval $[0, 1]$.

5.3 Computational details

5.4 General description and parameters of experiments

We have run a number of experiments to explore what this model can tell us. For several reasons we have focused on a very simple case with two strategic suppliers. First this case gives the clearest contrast with the competitive equilibirum other than monopoly, which is not really an equilibrium problem at all. Thus this case serves best to highlight the possible strategic pitfalls of the various payment rules.

Second the number of strategy profiles s for which the payoffs $u(s)$ must be evaluated becomes intractably large for even a few players. We prefer at this stage of research to spend our computational effort on finer discretization of the strategy spaces rather than more players.

Third with three or more players, the complementarity formulation of a mixed Nash equilibrium problem is nonlinear, and thus the algorithm that we are using to solve CP formulation is not guaranteed to terminate at a solution. We have observed this difficulty in practice.

All our experiments are conducted in the symmetric setting, That is, the two players i_1 and i_2 are identical in terms of their cost parameters and their bid parameters and the discretizations of their strategy spaces. Therefore we formulate and solve the smaller, symmetric CP formulation, obtaining a symmetric equilibrium. The economic interpretation of this symmetric equilibrium is clearer than that of an asymmetric equilibrium of an asymmetric game. Nevertheless our computer code is written so as to be easily used to run experiments on asymmetric models in future work.

5.4.1 Computer implementation

This discretized Nash equilibrium model of strategic bidding in electric power markets with unit commitment and various payment rules has been implemented in a sequence of computer codes written in GAMS.

The user sets the cost parameters of the strategic and competitive suppliers, the expected demand and its standard deviation, the bounds on the bid spaces, and the number of discretization points of the bid parameters and of the demand distribution. For each combination of bids and realized demand and commitment vector we solve the dispatch and pricing problem (5.4) and collect the results, consisting of the perceived cost of generation, the dispatch q, the prices p and r, the profits to the strategic suppliers, and the true cost of generation. This problem is solved in GAMS using the nonlinear programming solver CONOPT. When passing from one such combination to the next, we do not terminate and restart the solver, but rather we pass the new combination to the solver and extract the new results. Altering an optimization model and obtaining a new solution while the solver is kept running is enabled by the GUSS tool in GAMS. This technique decreases that amount of time required by our model by an order of magnitude.

Furthermore, for these combinations of bid, demand and commitment, each giving a separate optimization problem to solve, we have found that we can combine a number of such separate optimization problems into a single batch. Concretely, if two optimization problems,

$$\min_{t \in T} f(t)$$

and

$$\min_{y \in Y} g(y)$$

are to be solved then we can solve both simultaneously by solving the Cartesian product

$$\min_{(t,y) \in T \times Y} f(t) + g(y)$$

This batching procedure requires more memory than solving the problems one at a time, but reduces the computation time. In our model we have found the best overall performance with batch sizes around 100.

Then for each combination of bids and demand realization, the commitments with minimum apparent cost are selected, and the corresponding dispatches, true costs, and profits are averaged. Finally, for each combination of bids, the dispatches, true costs, and profits are averaged over demand realizations. The resulting average profits define the payoffs $u_i(s_1, s_2)$ for each pair of strategies s_1 and s_2.

With the payoffs in hand we solve the discrete Nash equilibrium model (4.6) using the complementarity solver PATH in GAMS. The resulting equilibrium is a pair of probability distributions over the discrete bid spaces of the two strategic suppliers, and from this we compute the expected values of the dispatch, strategic profits, and true cost of generation under equilibrium.

5.4.2 Evaluating criteria for existence of equilibrium

In our method we consider a sequence of increasingly fine discrete approximations of the strategy space, obtaining an equilibrium in each approximate game. We may observe convergence of this sequence of equilibria, but that is no guarantee that the limit is an equilibrium of the original game on an infinite strategy space. To verify that the limit is an equilibrium, we have two sufficient conditions, one simple, the other more complicated. The simple condition is that the payoffs u_i be continuous. The more complicated condition involves the structure of the set of points of discontinuity of u_i. We are not yet able to evaluate the more complicated condition, but we can at least partially evaluate the simple condition.

In our method we can obtain direct numerical information on the continuity of u_i since at each iteration of the method, we evaluate u_i at a number of points. To see how this is done, consider a function f of a real scalar variable t, evaluated at points t_k with $t_{k+1} > t_k$ and $f_k = f(t_k)$. A finite difference approximation f'_k of f at t_k is given by

$$f'_k = \frac{f_{k+1} - f_k}{x_{k+1} - x_k} \tag{5.10}$$

Then the maximum

$$\overline{f}' = \max_k |f'_k| \tag{5.11}$$

converges to a Lipschitz constant for f as the coarseness of the mesh of sample points converges to 0, if f is Lipschitz continuous. On evaluating the payoffs u_i at each sampled point in the strategy space, we apply this method in each of the coordinate dimensions and to each player i, and take the maximum L as a lower bound for a Lipschitz constant for u. Clearly the behavior of L does not completely characterize continuity of u. For example, L only accounts for the coordinate directions in strategy space. But if L appears to be bounded as the coarseness of the discrete approximation of the strategy space converges to 0, then we take this as a good indication that u is continuous.

To demonstrate the necessity of our mixed equilibrium approach, recall that one of the sufficient conditions in the principal existence theorem for pure equilibrium is that the payoffs u_i be quasiconcave in the strategy variable s_i of player i. We cannot easily evaluate quasiconcavity of a function from local information (i.e. derivatives or finite difference approximations) but we can evaluate the stronger condition of concavity, at least in coordinate directions. For this consider again a function f taking values f_k at points $t_k \in \mathbb{R}$ with first order finite differences f'_k. Now compute the second order finite differences

$$f''_k = \frac{f'_{k+1} - f'_k}{x_{k+1} - x_k} \tag{5.12}$$

If the maximum

$$\hat{f}'' = \max_k \max\{f''_k, 0\} \tag{5.13}$$

of the positive parts is positive, then f is not concave. As with the discontinuity measure L, we apply this idea to the payoffs u_i in each dimension of s_i, taking the maximum H as an indication of nonconcavity.

5.5 Results

We have conducted a number of experiments using our discrete approximation mixed equilibrium model of strategic bidding in electric power markets. Generally we vary some parameter of interest in a range and solve the equilibrium problem for all payment rules and market structures for each value of the parameter, recording results such as the expected total cost TC of generation in equilibrium, the expected dispatch q_i of the strategic suppliers, the expected strategic payoffs

u_i, the discontinuity measure L of the payoffs, and the payoff nonconcavity measure H. We now describe the most interesting of these experiments and results.

5.5.1 Varying demand with expensive competitive fringe

In this first experiment, we vary the expected demand q_d and examine the expected total cost TC at equilibrium in order to understand how the different payment rules and market structures com[are with respect to the most important criterion of overall system cost. The competitive fringe is parametrized to be fairly expensive, relative to the strategic suppliers, so as to highlight the role of market power.

Each strategic supplier is characterized by $\overline{q}_i = 1$, $c_i^0 = 1$, $a_i^0 = 1$, $b_i^0 = 1$, $\underline{c}_i = 0$, $\overline{c}_i = 5$, $\underline{a}_i = 0, \overline{a}_i = 10, \underline{b}_i = 1, \overline{b}_i = 1$. Thus the strategic suppliers are required to bid the cost parameter b_i truthfully but may vary the parameters a_i and b_i. This keeps the computational burden manageable.

The competitive fringe is characterized by $a_c = 0$ and $b_c = 10$. The numbers of discretization points for the strategic variables are $n_{a,i} = 6$, $n_{b,i} = 1$, $n_{c,i} = 6$, and the number of discrete demand points is $n_d = 8$. Demand itself has standard deviation $\hat{q}_d = 0.1$, and we vary the mean demand $\overline{q}_d$ from a minimum of 0 to a maximum of 4, or twice the total strategic capacity.

(put this in a separate subsection?) First we note that the equilibrium strategies found in this experiment at various levels of expected demand and different payment rules include both pure equilibria and mixed equilibria. For example, with $q_d = 1.75$, the equilibrium bid for PPO is $(a_i, b_i, c_i) = (9.16, 1, 0.83)$, a pure strategy. In this equilibrium, both strategic suppliers overbid their marginal costs but bid their unit commitment costs more or less honestly. This fits with our intuitive view of the PPO rule as requiring suppliers to overbid to recover commitment costs that are not explicitly paid. On the other hand, at the same expected demand, PPC, PPU, and AB all yielded the mixed strategy placing equal weight on the three pure bids $(4.16, 1, 5.83), (2.5, 1, 7.5), (0.83, 1, 9.16)$. These three points are collinear, suggesting that the equilibrium is a line segment in the limit as the discretization coarseness converges to 0. Each point in the support of this mixed strategy represents a significant overstatement of the true costs of the strategic firms, so clearly they possess market power. It appears that the strategic firms

randomize between overstating their marginal costs and overstating their commitment costs. The reason for randomization may be that the optimal reaction to overstating marginal costs is for the other firm to overstate commitment costs, and vice versa. Thus this example demonstrates the need to consider mixed equilibrium in the presence of unit commitment bidding.

Now let us see what is the expected total cost TC of generation with equilibrium bids under these different payment rules. Consider the plot of TC against $\overline{q}_d$ in figure (5.1). Under all four rules, TC shows a broadly increasing trend as $\overline{q}_d$ increases. This is in line with our intuition from competitive equilibrium, in which higher demand is more expensive to serve. Looking closer we see that all four rules have a single local maximum and then a local minimum and then continue increasing. This interior peak occurs at $\overline{q}_d = 2.5$ under the three rules that explicitly compensate unit commitment, but it occurs at $\overline{q}_d = 3$ for PPO. So the natural question is, why does this peak occur?

Looking at a plot of the expected generation q_i by each strategic supplier against $\overline{q}_d$ in figure (5.2), we see that all four rules show q_i increasing from 0 to 1 as $\overline{q}_d$ increases from 0 to 4. All four rules behave similar to a step function, with q_i stuck at 0, then increasing rapidly to about 0.5, and remaining there for an interval, and then increasing rapidly to 1. Evidently, in this interval where q_i is about 0.5, one firm is committed and the other firm is not committed, with equal probability, and it requires significantly higher demand to see both firms committed with relatively high probability. Finally, it appears that in all four rules, this interval where q_i stays at 0.5 terminates at just the same value of $\overline{q}_d$ as where the peak in TC occurs. So the interpretation is that, as $\overline{q}_d$ increases, but not so much as to require the commitment of both firms, it becomes increasingly expensive to serve this higher demand with only one firm and the competitive fringe, and once it is worthwhile to commit both firms, the total costs drop as both firms are dispatched and the very expensive competitive fringe does not need to supply so much.

From the standpoint of the total cost, the most striking observation is that PPO performs quite poorly. It appears that the SO fails to see the benefit of committing two firms until demand is so high that the residual demand served by the competitive fringe is quite costly. And the reason why

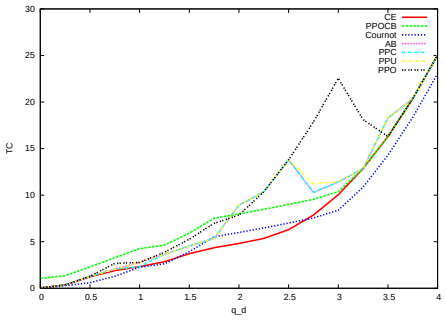

Figure 5.1 Expected total cost

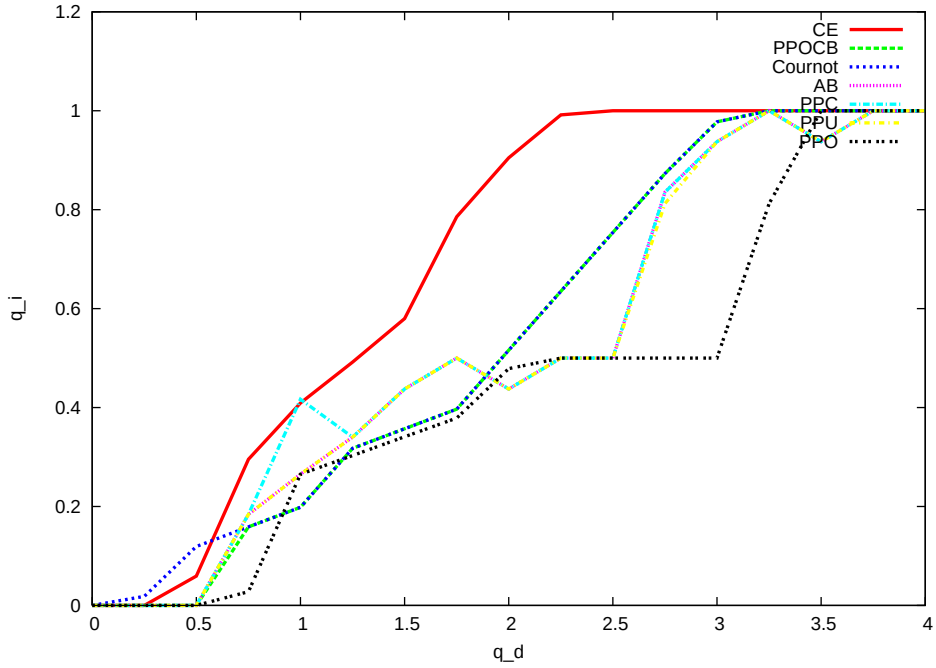

Figure 5.2 Expected strategic generation

the SO does not see two firms as optimal is that both firms have been induced to overstate their costs in order to compensate for the lack of explicit payment for commitment.

Finally, from a plot of the expected payoff u_i to each strategic firm in figure (5.3), we see that the suppliers are essentially unaffected by either the sudden transition from committing one supplier to committing both or by the lack of explicit compensation for commitment in the PPO rule. They are insulated by market power.

One tentative conclusion suggested by this experiment is that is that the three rules that do explicitly compensate unit commitment appear to give essentially equivalent outcomes from the standpoint of economic efficiency, and they all do better than PPO because PPO seems to lead to overbidding and under commitment.

5.5.2 Varying demand with inexpensive competitive fringe

Running the same experiment, but with a less expensive competitive fringe defined by $b_c = 2$, we expect to see the influence of market power diminish, relative to the first experiment. The total cost of generationn is plotted in figure 5.4. Taking the difference in TC between any of the four payment rules and the competitive equilibrium (CE) as an indicator of market power, under scenarios of cheap and costly competitive supply, we see some market power, but it is much more pronounced under costly supply, especially in rule PPO.

5.5.3 Verifying convergence with finer strategy discretization

We now describe an experiment that we have conducted to verify the convergence of our method as the coarseness of the discretization of the bid space converges to 0. The problem parameters are as in the first experiment, with expected demand $\overline{q}_d$ fixed arbitrarily at 1.618. The bid discretization parameters are given by $n_c = 1$ and values of $n_a = n_b$ ranging from 1 to 14. For each value of $n_a = n_b$, we solve the discrete equilibrium model for all the payment rules and market scenarios. We then plot the resulting expected values of the main outputs in equilibrium, the total cost of generation TC (figure 5.5), the strategic dispatch q_i (figure 5.6), and the strategic profit u_i (figure 5.7). Certainly u_i is converging rather rapidly. And q_i and TC are oscillating with

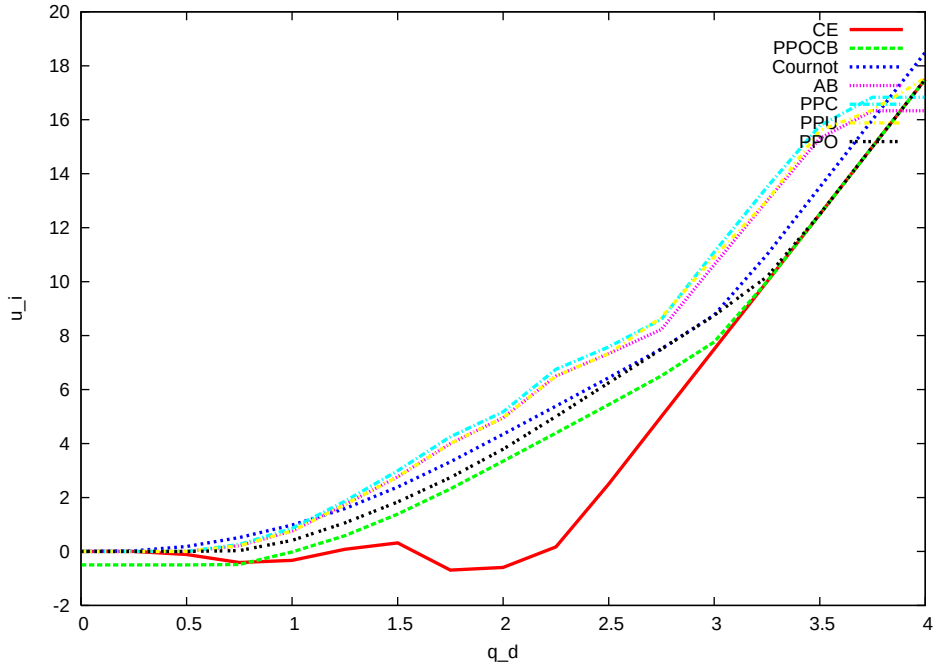

Figure 5.3 Expected strategic profit

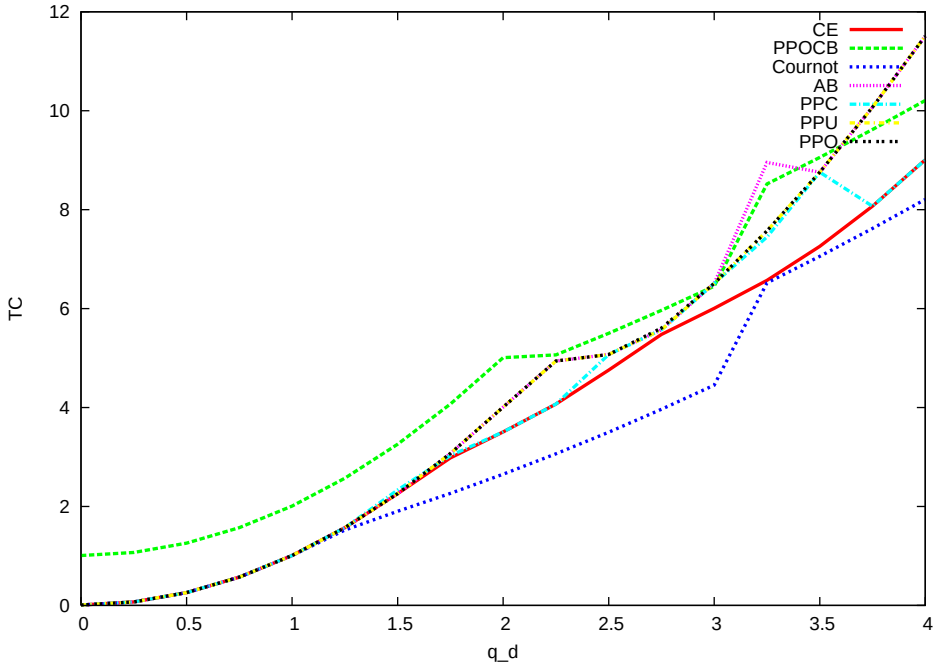

Figure 5.4 Expected total cost of generation, $b_c = 2$

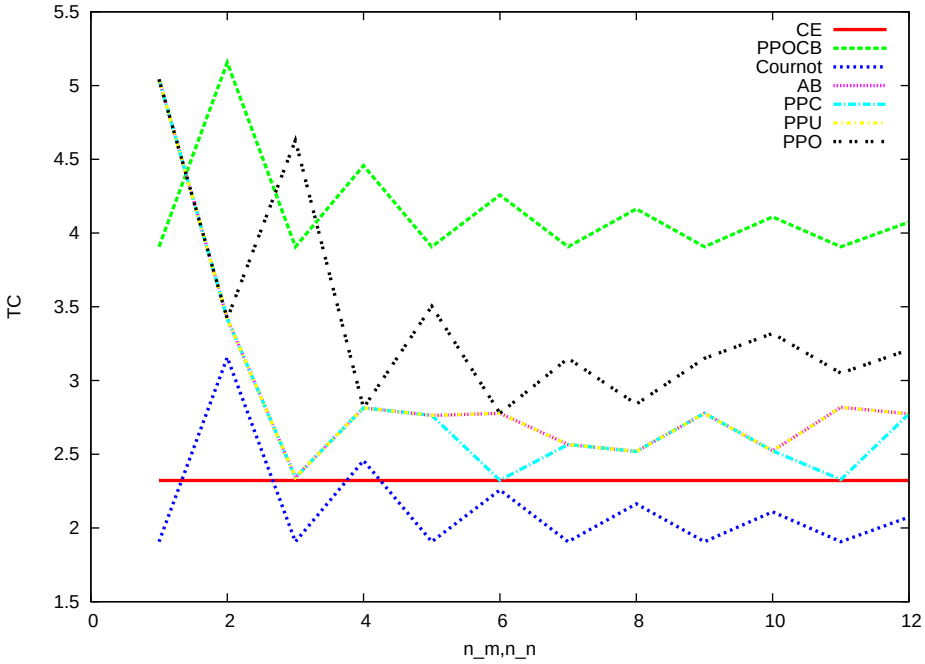

Figure 5.5 Total cost under mesh refinement

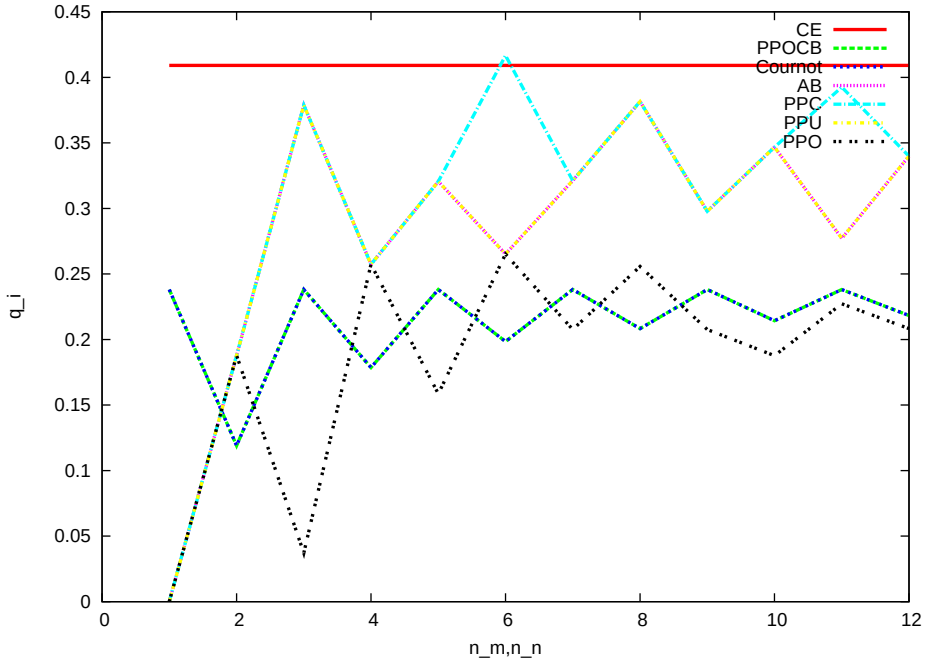

Figure 5.6 Strategic dispatch under mesh refinement

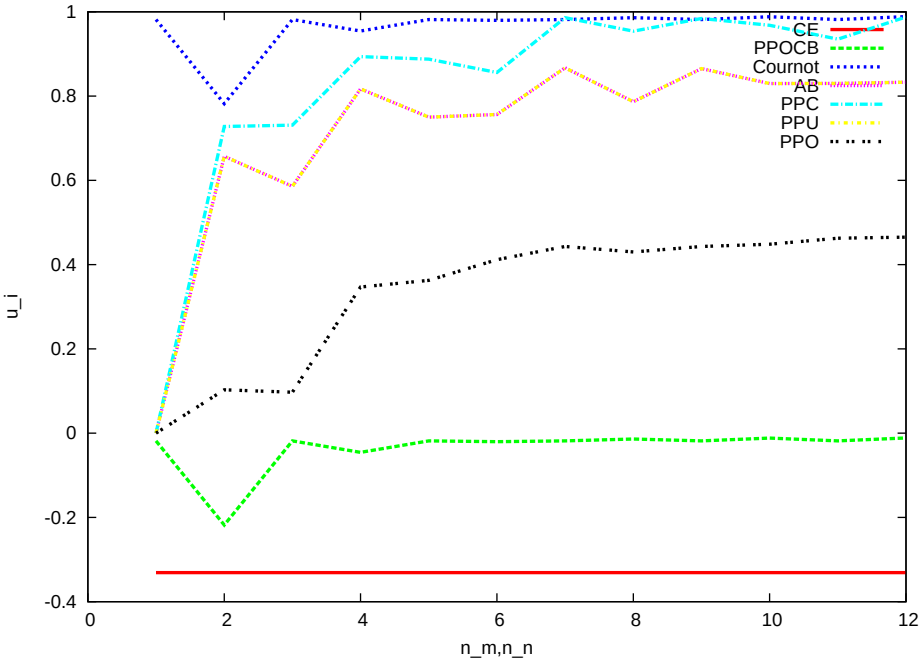

Figure 5.7 Strategic profit under mesh refinement

decreasing amplitude and in general converging somewhat slowly.

We can also evaluate the measure L of discontinuity as the strategy mesh is defined. Figure 5.8 shows no sign that L remains bounded as the mesh gets finer, suggesting that the payoffs are discontinuous, so that we actually do not have a guarantee that the limiting distribution of strategies is an equilibrium. Further work is needed to evaluate the more complicated sufficient condition for mixed equilibrium based on the structure of the set of discontinuities of the payoffs.

And finally we can evaluate the measure H of nonconcavity. Figure 5.9 shows appreciable nonconcavity in the payoffs under all payment rules and special market conditions except for competitive equilibrium (CE). This demonstrates further the need to consider mixed equilibrium by precluding our only means to a theoretical guarantee that a pure equilibrium exists.

5.5.4 Refining the bid space in only one dimension

Returning to figure 5.8 we notice that the different rules show different behaviors of L as the bid spaces are refined. In particular the three special scenarios requiring convex bids, namely CE, PPOCB, and Cournot, appear to show slower growth in L, and CE even has $L = 0$. Conjecturing that these rules might actually have continuous payoffs, we need to investigate the behavior of L with much finer discretization. Refining two bid dimension at once, as we have done so far, would require substantial computation, so we focus in this experiment on refinement of the discretization of the a_i-spaces, representing the b_i- and c_i-spaces by just one point. A plot of L as n_a ranges from 1 to 64 in figure 5.10 shows that these rules do indeed have continuous payoffs as L appears to level off.

5.5.5 Refining the discrete approximation of stochastic demand

In this experiment we vary the number n_d of atoms used to approximate the continuous distribution of random demand q_d. The parameters of the experiment remain as in the first experiment, with expected demand $\overline{q}_d = 2.75$.

First consider a plot of the total cost TC in figure 5.11 As n_d increases, the values of TC for the different market rules stabilize. Interestingly, the relative values of TC for the different rules

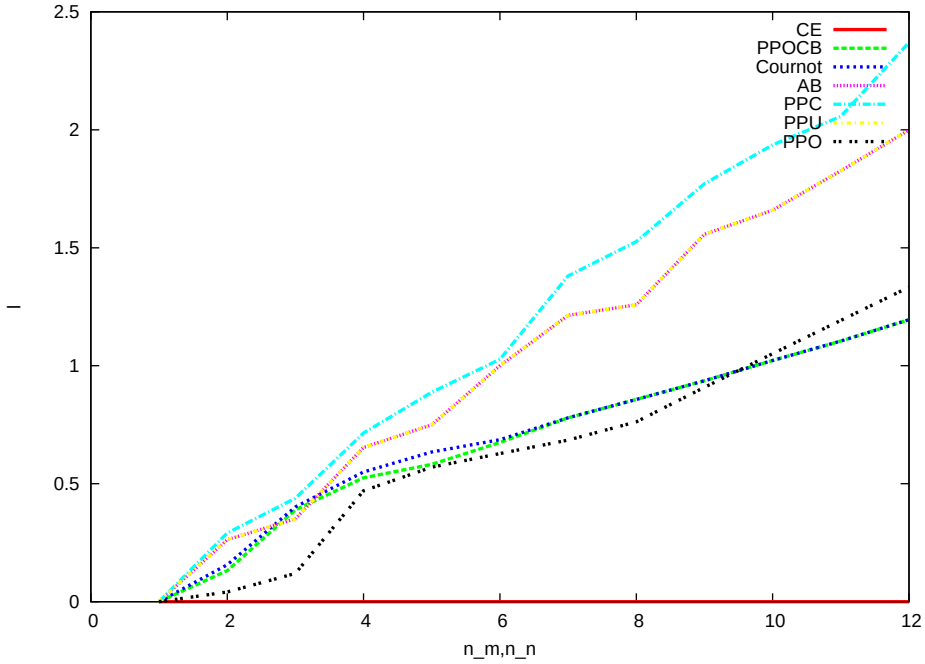

Figure 5.8 Payoff discontinuity under mesh refinement

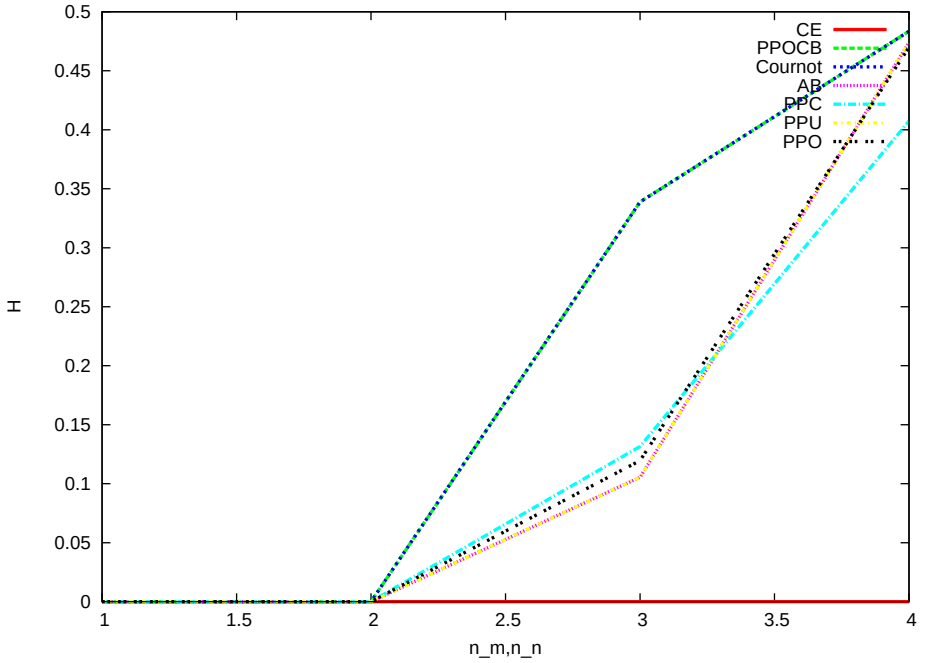

Figure 5.9 Payoff nonconcavity under mesh refinement

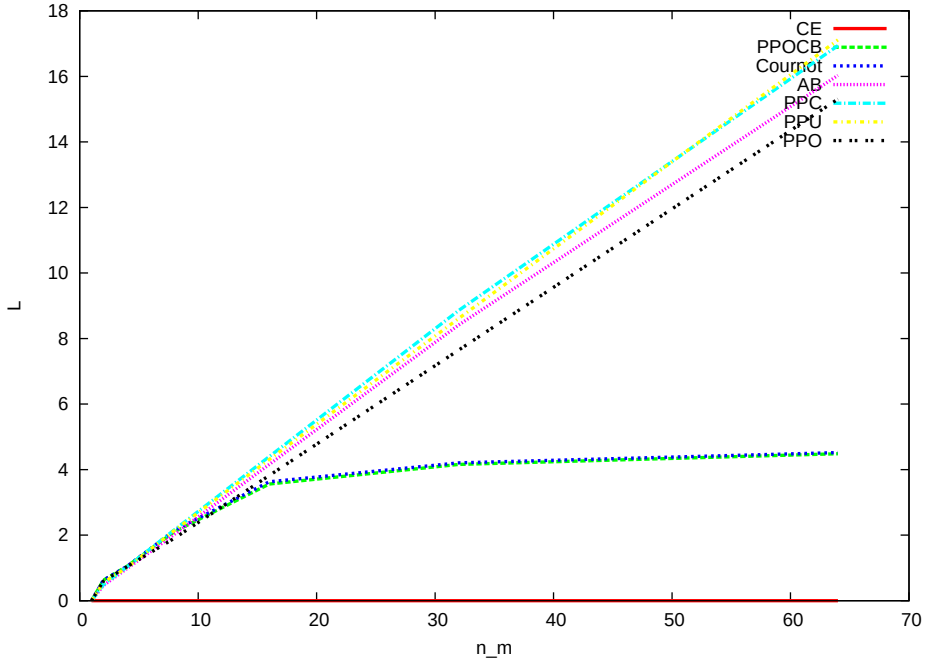

Figure 5.10 Payoff discontinuity under mesh refinement in a_i only

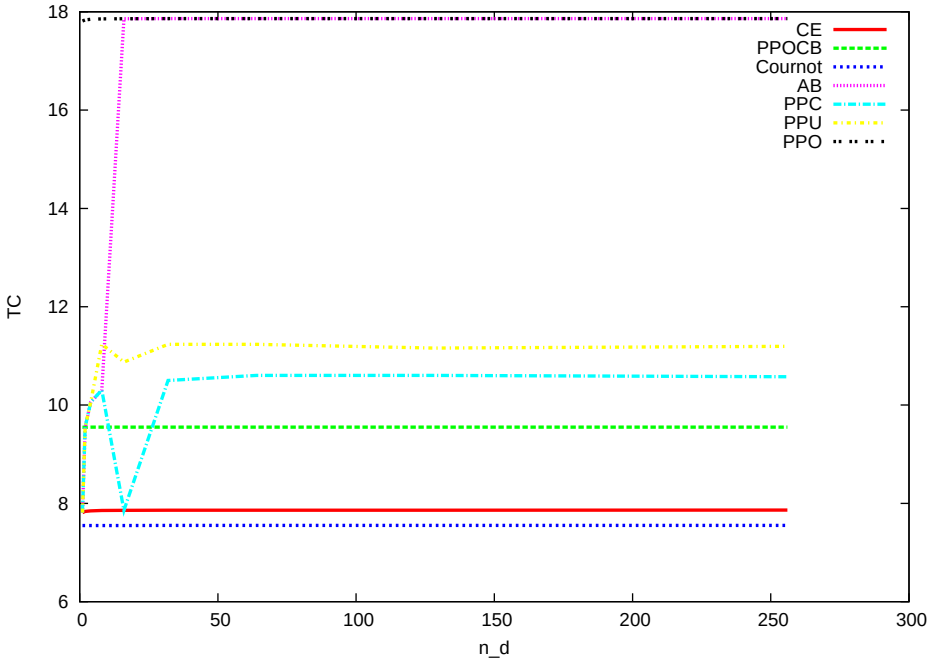

Figure 5.11 Total cost under refinement of demand distribution

are qualitatively different at larger values of n_d than were used in the first experiment, enabling us to see differences among rules that were indistinguishable in that first experiment. In further work we will use larger values of n_d. The other results q_i, u_i, L, and H showed similar behavior, appearing to converge with values of $n_d \geq 32$ or so.

One other interesting observation that we can make from this experiment concerns the limiting values of L and H as n_d increases. Looking at plots of L and H in figures 5.12 and 5.13 we see that neither one converges to 0, suggesting that finer approximation of the distribution of demand is not a remedy for the essential discontinuity and nonconcavity of this model.

5.5.6 Varying expected demand with fine demand distribution

Having learned that we need a fairly large number of demand scenarios in order to ensure accuracy and even tell the difference between some payment rules, we repeat the first experiment with $n_d = 33$. The resulting total cost TC, as a function of expected demand $\overline{q}_d$, is plotted in figure 5.14. There is little difference from the cost with $n_d = 8$, with the exception that the intermediate peak in TC for PPU is smaller and occurs at a lower value of $\overline{q}_d$ than those for PPC and AB, and as before these are smaller and earlier than the peak for PPO.

Also, considering a plot of the expected strategic dispatch q_i in figure 5.15, we see that PPU induces both strategic suppliers to produce for lower values of $\overline{q}_d$ than do PPC or AB, thus preventing costly use of the competitive fringe. So we are now able to distinguish PPU from PPC and AB, and we may now have greater confidence in claiming that PPU restrains market power more effectively than AB and PPC, though these are still better than PPO.

It still appears, however, that PPOCB yields even lower total cost than PPU. PPOCB is an appealing rule and market structure, partly because it absolves the SO of solving the technically difficult unit commitment problem. However, for low demand, PPOCB yields negative profits to the strategic suppliers, as we can see in a plot in figure 5.16 of u_i. Since even market power and noncompetitive behavior do not allow all market participants to recover their costs, PPOCB might be considered unfair to suppliers and might yield long term underinvestment in capacity.

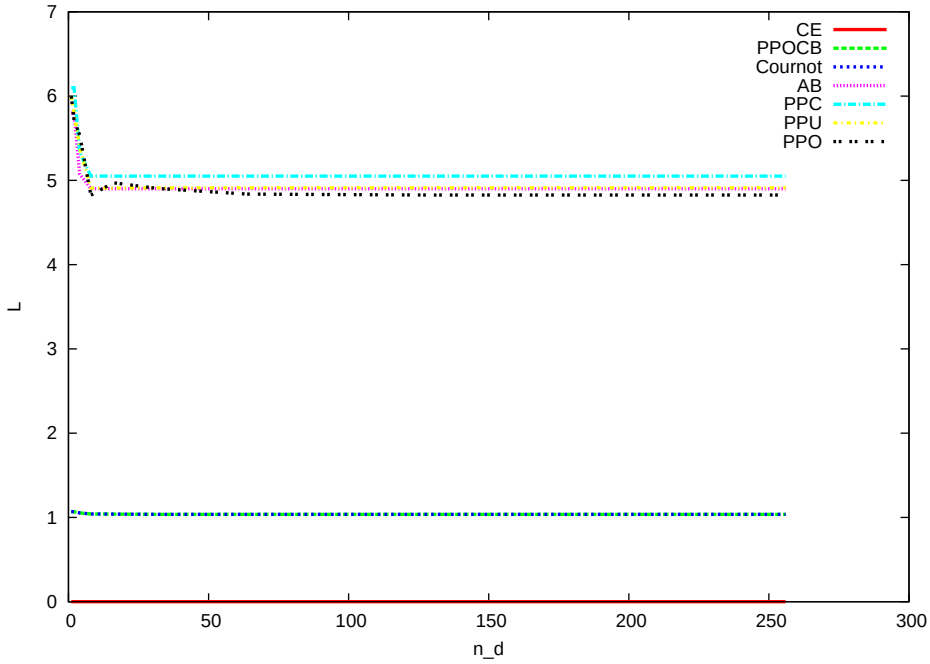

Figure 5.12 Discontinuity under refinement of demand distribution

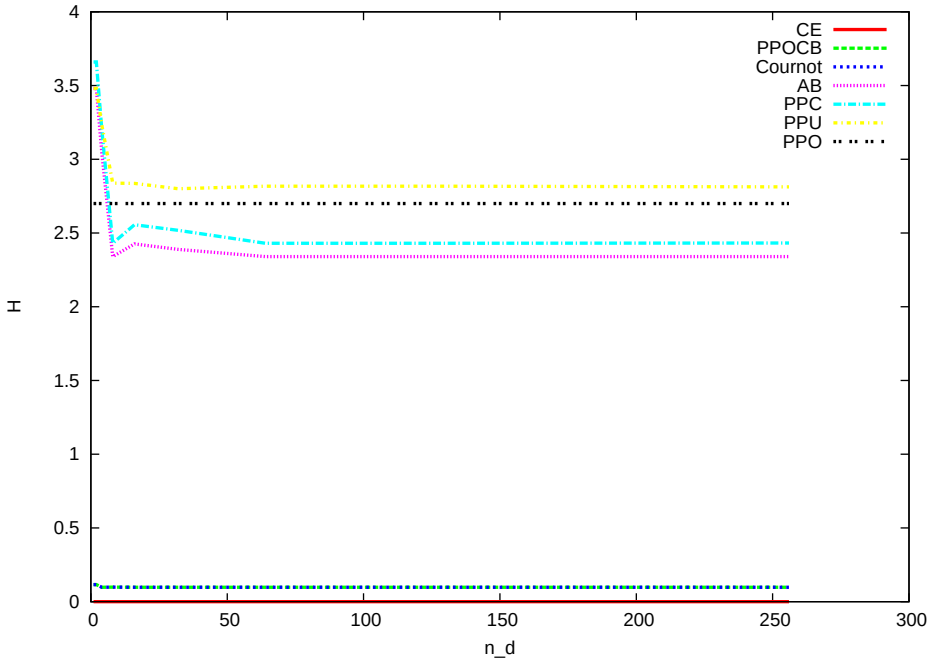

Figure 5.13 Nonconcavity under refinement of demand distribution

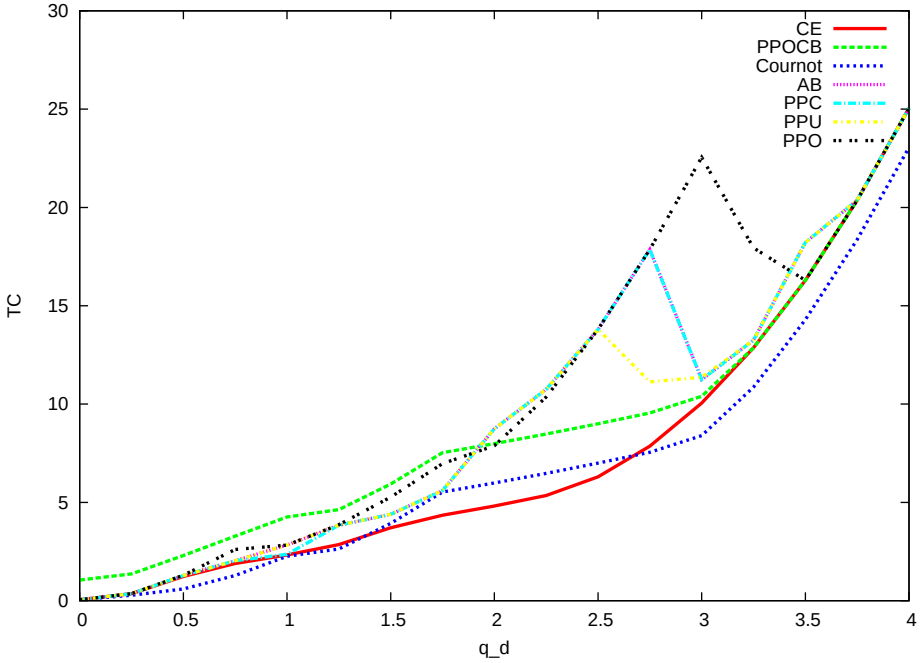

Figure 5.14 Total cost with varying demand, $n_d = 33$

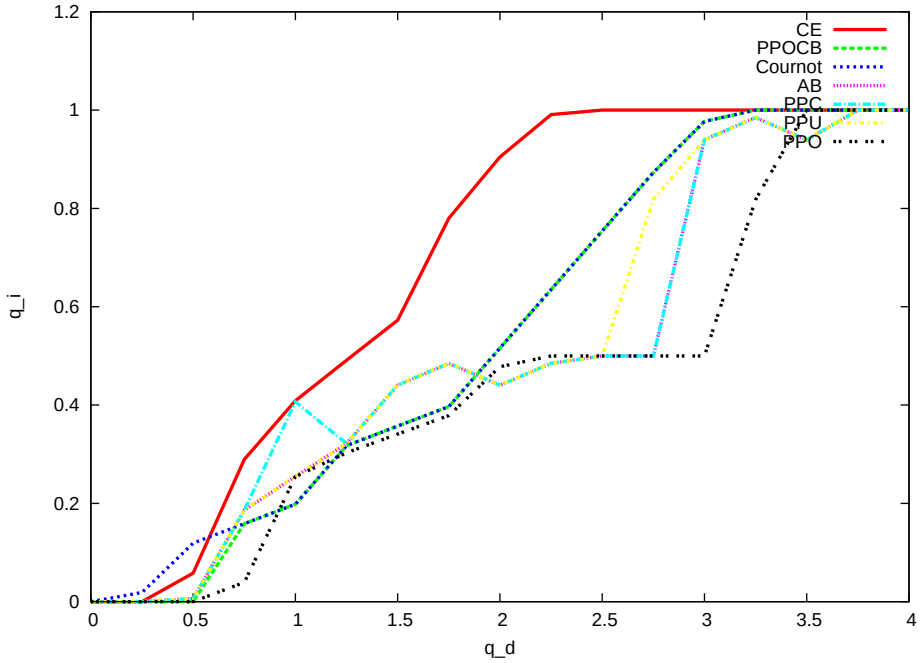

Figure 5.15 Strategic dispatch with varying demand, $n_d = 33$

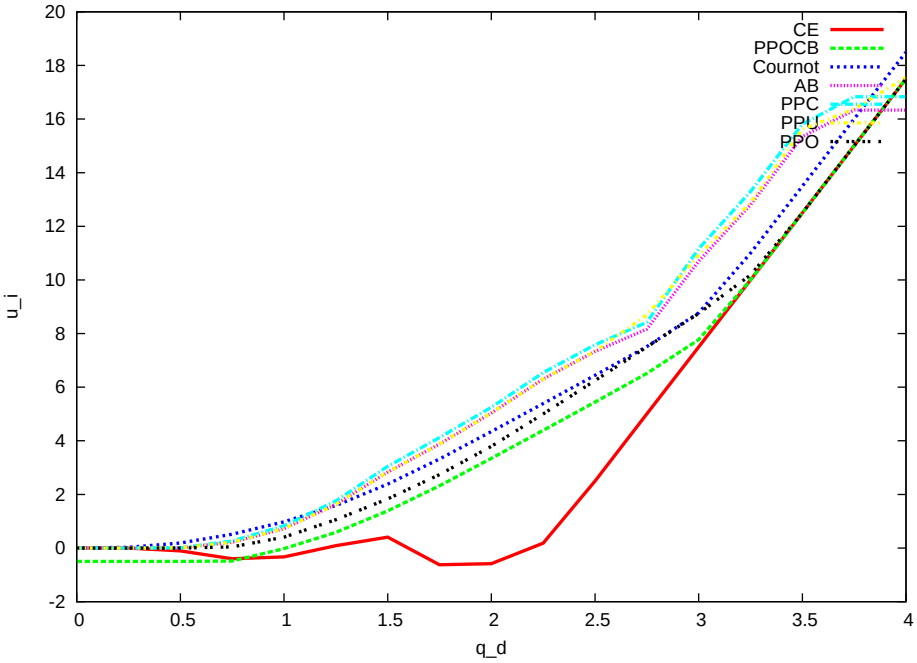

Figure 5.16 Strategic payoff with varying demand, $n_d = 33$

Finally, though CE and the Cournot model both show lower total costs than PPU and PPOCB, they should not be seen as realistic policy alternatives, as they simply assume strategic behavior and nonconvexity, respectively, do not exist. Thus we conclude by recommending both PPU and PPOCB as the best market rules and structures in the presence of nonconvex costs, from the viewpoint of restraining market power and maintaining capacity investment incentives.

5.6 Analysis and conclusions

The purpose of this model is to deliver insight into the economic efficiency that can be expected from several payment rules and market structures present or contemplated in wholesale electric power markets. Frequently, these rules are analyzed under the assumption of competitive behavior, in which all participants report their true cost structure to the system operator, and then the SO is able to dispatch generation at minimum total cost. In that setting we need only ensure that the bidding structure allows participants to report their costs with sufficient accuracy. Some analyses seek a pure Nash equilibrium in bids to determine the susceptibility of the different rules to market power by comparing the true net cost of generation under this equilibrium to that under competitive behavior. To find a pure equilibrium reliably, and indeed to ensure that one exists, one must make convexity and continuity assumptions that preclude essentially any kind of unit commitment, among other market structures that one might want to model.

In our work, we use mixed Nash equilibrium to enable the inclusion of unit commitment in the bidding structure, the dispatch algorithm, and the payment rules. We are able to give a reasoned comparison of market structures and payment rules that takes explicit account of unit commitment. We conclude that the best structures are PPU (pricing power with an uplift to costs as bid) and PPOCB (pricing power with a requirement of convex bids). In particular we are able to show that the recently proposed rule PPC (pricing both power and commitment) has greater potential for net losses from noncompetitive behavior than PPU and PPOCB with no compensating benefit.

Because of the combinatorial nature of our method, we are only able to model small instances of the equlibrium problem that we consider - we are limited to few strategic suppliers, few cost parameters, and a coarse representation of the bid space. This is the curse of dimensionality (and

nonconvexity). So our method is not presently applicable to grid-scale models of an electric power market.

Our model does however have two important applications. First, the small scale allows us to gain intuition into how unit commitment might affect market power under different market rules. In this application, we can say what kind of strategic behavior and noncompetitive outcomes system operators ought to look out for under different rules, though we cannot say that the overall costs of unit commitment-induced market power as predicted by our model will be as great in practice. Second, such a small scale model is numerically realistic in cases where transmission congestion isolates a small part of a power grid containing, say, two large power plants owned by different companies.

One may draw an analogy between our model and the familiar partial equilibrium model for a market in a single good, composed of a supply curve and a demand curve. The partial equilibrium model gives us intuition into the predictions of a computable general equilibrium model giving a numerically realistic picture of an economy of many goods and production and consumption sectors. There is no numerically realistic model of market power induced by unit commitment in power market with many individual power plants and generation companies, but our model shows what that kind of market power should look like.

Chapter 6

Implementation of a Large-Scale Optimal Power Flow Solver Based on Semidefinite Programming

6.1 Introduction

The optimal power flow (OPF) problem seeks decision variable values that yield an optimal operating point for an electric power system in terms of a specified objective function, subject to both network constraints (i.e., the power flow equations, which model the relationship between voltages and power injections) and engineering constraints (e.g., limits on voltage magnitudes, active and reactive power generations, and flows on transmission lines and transformers). Total generation cost is typically chosen as the objective function.

The OPF problem is nonconvex due to the nonlinear power flow equations [31]. Nonconvexity of the OPF problem has made solution techniques an ongoing research topic since the problem was first introduced by Carpentier in 1962 [9]. Many OPF solution techniques have been proposed, including successive quadratic programs, Lagrangian relaxation, genetic algorithms, particle swarm optimization, and interior point methods [40, 56].

Recently, significant research attention has focused on the application of semidefinite programming to the OPF problem [5, 29]. Through the use of a rank relaxation, the OPF problem is reformulated as a convex semidefinite program. If the relaxed problem satisfies the rank constraint (i.e., the semidefinite program has zero duality gap), the globally optimal solution to the original OPF problem can be determined in polynomial time. No prior OPF solution method offers a guarantee of finding the global solution; semidefinite programming approaches thus have a substantial advantage over traditional solution techniques. Note, however, that the rank constraint is

not always satisfied, which means that semidefinite relaxations do not give physically meaningful solutions for all realistic power system models [32]. Recent research has investigated the conditions under which the rank constraint is satisfied; to date, sufficient conditions for rank constraint satisfaction include requirements on power injection limits and either radial networks (typical of distribution system models) or appropriate placement of controllable phase shifting transformers [55, 8, 43, 30]. Additional research includes the use of semidefinite programming to create voltage stability margins in the power flow problem [34].

This paper first focuses on modeling aspects that must be addressed in order to apply the semidefinite program to realistic power system models. The first issue addressed is multiple generators at the same bus. By equating bus power injections with power generation, existing formulations only allow a single generator to exist at a bus. We use the concept of equal marginal generation cost to produce a formulation allowing for multiple generators at the same bus, each with separate cost functions and generation limits.

A method for incorporating flow limits on parallel lines is then presented. Existing formulations limit the flow between two buses, which cannot properly account for parallel lines with different electrical properties and flow limits. In contrast, the proposed method limits the flow on each individual line and can therefore account for parallel lines. Lines in this formulation can have off-nominal voltage ratios and non-zero phase shifts.

This paper next advances research in the computational tractability of applying semidefinite programming to large power system models. Semidefinite programming formulations of the OPF problem constrain a $2n \times 2n$ symmetric matrix to be positive semidefinite, where n is the number buses in the system. The semidefinite program size thus grows as the square of the number of buses, which makes solution of the OPF problem by semidefinite programming computationally intractable for large systems. Recent work using matrix completion [15, 36, 26] reduces the computational burden inherent in solving large systems by taking advantage of the sparse matrix structure created by realistic power system models. Sojoudi and Lavaei [43] and Jabr [24] present formulations that decompose the single large $2n \times 2n$ positive semidefinite matrix constraint into

positive semidefinite constraints on many smaller matrices. If the matrices from these decompositions satisfy a rank constraint, the $2n \times 2n$ matrix also satisfys the rank constraint and the optimal solution can be obtained. Sojoudi and Lavaei's decomposition is based on a cycle basis of the network. Jabr's decomposition is based on the maximal cliques of a chordal extension of the network.

We provide several enhancements to the existing decompositions. Specifically, we present a heuristic algorithm for combining some of the many small matrices resulting from the decomposition. Since linking constraints are required between terms of the decomposed matrices that refer to the same term in the $2n \times 2n$ matrix, it is not always advantageous to create the smallest possible matrices. Combining matrices eliminates some of these linking constraints, which can result in significant computational speed increases. We justify the claim that the proposed algorithm can substantially increase computational speed using both theoretical arguments and several test cases.

A further enhancement presented in this paper is a technique for recovering the optimal voltage profile from the decomposed matrices. None of the existing literature describes a method for actually obtaining the optimal voltage profile from a solution to a decomposed formulation.

Although we focus on Jabr's decomposition [24] due to the voluminous supporting literature on matrix completion with chordal extensions (e.g. [15, 36, 26]), both of these enhancements could be applied to Sojoudi and Lavaei's decomposition [43] as well.

We finally describe a modification to Jabr's decomposition that allows for application to general power systems. Jabr's decomposition uses a Cholesky factorization of the absolute value of the imaginary part of the bus admittance matrix to form a chordal extension of the network. However, this matrix may not be positive definite (for instance, in networks with sufficiently large shunt capacitances), thus preventing calculation of a Cholesky factorization. We describe an alternative matrix that is always positive definite and gives an equivalent chordal extension, thus enabling Jabr's decomposition for general networks.

The paper is organized as follows. Section 6.2 provides both the classical formulation of the OPF problem and a proposed semidefinite programming formulation that incorporates multiple generators at the same bus and parallel lines, including lines with off-nominal voltage ratios and

non-zero phase-shifts. Section 6.3 first gives an overview of Jabr's matrix completion decomposition and then presents three advances in decompositions for large-scale system models: an algorithm that improves computation speed by combining matrices, a technique for recovering the optimal voltage profile from a solution to a decomposed formulation, and a modification to Jabr's method that extends its applicability to general power system networks.

6.2 The OPF Problem and Modeling Issues

We first present the OPF problem as it is classically formulated. Specifically, this formulation is in terms of rectangular voltage coordinates, active and reactive power generation, and apparent power line-flow limits. Each bus may have multiple generators, and parallel lines are allowed. This classical OPF formulation is generally nonconvex. We then describe a semidefinite programming formulation of the OPF problem adopted from [29] that handles the modeling issues of multiple generators at the same bus and parallel lines.

6.2.1 Classical OPF Formulation

Consider an n-bus power system, where $\mathcal{N} = \{1, 2, \ldots, n\}$ represents the set of all buses. $\mathcal{G}$ represents the set of all generators and $\mathcal{G}_i$ represents the set of all generators at bus i (if no generators exist at bus i, then $\mathcal{G}_i$ is the empty set). Let $P_{Gg} + jQ_{Gg}$ represent the active and reactive power output of generator $g \in \mathcal{G}$. Let $P_{Di} + jQ_{Di}$ represent the active and reactive load demand at each bus $i \in \mathcal{N}$. Let $V_i = V_{di} + jV_{qi}$ represent the voltage phasors in rectangular coordinates at each bus $i \in \mathcal{N}$. Superscripts "max" and "min" denote specified upper and lower limits. Let $\mathbf{Y}$ denote the network admittance matrix.

$\mathcal{L}$ represents the set of all lines, with line $k \in \mathcal{L}$ having terminals at buses l_k and m_k, with parallel lines allowed (i.e. more than one line between the same terminals). Let S_k represent the apparent power flow on the line $k \in \mathcal{L}$.

We consider a quadratic objective function associated with each generator $g \in \mathcal{G}$, typically representing a dollar/hour variable operating cost. (Recent research [43] has extended the semidefinite programming formulation to any convex cost function.)

The classical OPF problem can then be written as

$$\min \sum_{g\in\mathcal{G}} c_{g2}P_{Gg}^2 + c_{g1}P_{Gg} + c_{g0} \tag{6.1a}$$

$$\text{subject to}$$

$$P_{Gg}^{\min} \le P_{Gg} \le P_{Gg}^{\max} \qquad \forall g \in \mathcal{G} \tag{6.1b}$$

$$Q_{Gg}^{\min} \le Q_{Gg} \le Q_{Gg}^{\max} \qquad \forall g \in \mathcal{G} \tag{6.1c}$$

$$\left(V_i^{\min}\right)^2 \le V_{di}^2 + V_{qi}^2 \le \left(V_i^{\max}\right)^2 \qquad \forall i \in \mathcal{N} \tag{6.1d}$$

$$|S_k| \le S_k^{\max} \qquad \forall k \in \mathcal{L} \tag{6.1e}$$

$$\begin{aligned}\sum_{g\in\mathcal{G}_i}(P_{Gg}) - P_{Di} =& V_{di}\sum_{h=1}^{n}(G_{ih}V_{dh} - B_{ih}V_{qh}) \\ &+ V_{qi}\sum_{h=1}^{n}(B_{ih}V_{dh} + G_{ih}V_{qh}) \qquad \forall i \in \mathcal{N}\end{aligned} \tag{6.1f}$$

$$\begin{aligned}\sum_{k\in\mathcal{G}_i}(Q_{Gg}) - Q_{Di} =& V_{di}\sum_{h=1}^{n}(-B_{ih}V_{dh} - G_{ih}V_{qh}) \\ &+ V_{qi}\sum_{h=1}^{n}(G_{ih}V_{dh} - B_{ih}V_{qh}) \qquad \forall i \in \mathcal{N}\end{aligned} \tag{6.1g}$$

Note that this formulation limits the apparent power flow measured at each end of a given line, recognizing that active and reactive line losses can cause these quantities to differ.

6.2.2 Semidefinite Programming OPF Formulation

This section first describes the OPF formulation in dual form, including the capability to incorporate parallel lines and multiple generators at the same bus. Let e_i denote the i^{th} standard basis vector in $\mathbb{R}^n$. Define the matrix $Y_i = e_i e_i^T \mathbf{Y}$, where the superscript T indicates the transpose operator.

Matrices employed in the bus power injections and voltage magnitude constraints are

$$\mathbf{Y}_i = \frac{1}{2}\begin{bmatrix} \mathrm{Re}\left(Y_i + Y_i^T\right) & \mathrm{Im}\left(Y_i^T - Y_i\right) \\ \mathrm{Im}\left(Y_i - Y_i^T\right) & \mathrm{Re}\left(Y_i\ + Y_i^T\right) \end{bmatrix} \tag{6.2}$$

$$\bar{\mathbf{Y}}_i = -\frac{1}{2}\begin{bmatrix} \mathrm{Im}\left(Y_i + Y_i^T\right) & \mathrm{Re}\left(Y_i - Y_i^T\right) \\ \mathrm{Re}\left(Y_i^T - Y_i\right) & \mathrm{Im}\left(Y_i + Y_i^T\right) \end{bmatrix} \tag{6.3}$$

$$\mathbf{M}_i = \begin{bmatrix} e_i e_i^T & \mathbf{0} \\ \mathbf{0} & e_i e_i^T \end{bmatrix} \tag{6.4}$$

A "line" in this formulation includes both transmission lines and transformers, where transformers may include both a phase shift and an off-nominal voltage ratio. That is, line k is modeled as a Π circuit (with series admittance $g_k + jb_k$ and shunt capacitances $\frac{b_{sh,\,k}}{2}$) in series with an ideal transformer (with turns ratio $1 : \tau_k e^{j\theta_k}$) in the same manner as in [56]. Note that a small minimum resistance is enforced on all lines in accordance with [29]. Define f_i as the i^{th} standard basis vector in $\mathbb{R}^{2n}$. Matrices employed in the line-flow constraints are then

$$\begin{aligned}
\mathbf{Z}_{k_l} &= \frac{g_k}{\tau_k^2}\left(f_{l_k}f_{l_k}^T + f_{l_k+n}f_{l_k+n}^T\right) \\
&\quad - c_l\left(f_{l_k}f_{m_k}^T + f_{m_k}f_{l_k}^T + f_{l_k+n}f_{m_k+n}^T + f_{m_k+n}f_{l_k+n}^T\right) \\
&\quad + s_l\left(f_{l_k}f_{m_k+n}^T + f_{m_k+n}f_{l_k}^T - f_{l_k+n}f_{m_k}^T - f_{m_k}f_{l_k+n}^T\right) \qquad (6.5) \\
\mathbf{Z}_{k_m} &= g_k\left(f_{m_k}f_{m_k}^T + f_{m_k+n}f_{m_k+n}^T\right) \\
&\quad - c_m\left(f_{l_k}f_{m_k}^T + f_{m_k}f_{l_k}^T + f_{l_k+n}f_{m_k+n}^T + f_{m_k+n}f_{l_k+n}^T\right) \\
&\quad + s_m\left(f_{l_k+n}f_{m_k}^T + f_{m_k}f_{l_k+n}^T - f_{l_k}f_{m_k+n}^T - f_{m_k+n}f_{l_k}^T\right) \qquad (6.6) \\
\bar{\mathbf{Z}}_{k_l} &= -\left(\frac{2b_k + b_{sh,k}}{2\tau_k^2}\right)\left(f_{l_k}f_{l_k}^T + f_{l_k+n}f_{l_k+n}^T\right) \\
&\quad + c_l\left(f_{l_k}f_{m_k+n}^T + f_{m_k+n}f_{l_k}^T - f_{l_k+n}f_{m_k}^T - f_{m_k}f_{l_k+n}^T\right) \\
&\quad + s_l\left(f_{l_k}f_{m_k}^T + f_{m_k}f_{l_k}^T + f_{l_k+n}f_{m_k+n}^T + f_{m_k+n}f_{l_k+n}^T\right) \qquad (6.7) \\
\bar{\mathbf{Z}}_{k_m} &= -\left(b_k + \frac{b_{sh,k}}{2}\right)\left(f_{m_k}f_{m_k}^T + f_{m_k+n}f_{m_k+n}^T\right) \\
&\quad + c_m\left(f_{l_k+n}f_{m_k}^T + f_{m_k}f_{l_k+n}^T - f_{l_k}f_{m_k+n}^T - f_{m_k+n}f_{l_k}^T\right) \\
&\quad + s_m\left(f_{l_k}f_{m_k}^T + f_{m_k}f_{l_k}^T + f_{l_k+n}f_{m_k+n}^T + f_{m_k+n}f_{l_k+n}^T\right) \qquad (6.8)
\end{aligned}$$

where, for notational convenience,

$$c_l = \left(g_k\cos\left(\theta_k\right) + b_k\cos\left(\theta_k + \frac{\pi}{2}\right)\right)/\left(2\tau_k\right) \qquad (6.9)$$

$$c_m = \left(g_k\cos\left(-\theta_k\right) + b_k\cos\left(-\theta_k + \frac{\pi}{2}\right)\right)/\left(2\tau_k\right) \qquad (6.10)$$

$$s_l = \left(g_k\sin\left(\theta_k\right) + b_k\sin\left(\theta_k + \frac{\pi}{2}\right)\right)/\left(2\tau_k\right) \qquad (6.11)$$

$$s_m = \left(g_k\sin\left(-\theta_k\right) + b_k\sin\left(-\theta_k + \frac{\pi}{2}\right)\right)/\left(2\tau_k\right) \qquad (6.12)$$

Define vectors of Lagrange multipliers associated with lower inequality bounds on active power, reactive power, and voltage magnitude as $\underline{\psi}_k$, $\underline{\gamma}_i$, and $\underline{\mu}_i$, and those associated with upper bounds as $\bar{\psi}_k$, $\bar{\gamma}_i$, and $\bar{\mu}_i$, respectively.

Define a scalar variable λ_i as the aggregated Lagrange multiplier (i.e. the locational marginal price (LMP)) of active power at each bus i. Note that λ_i is not constrained to be non-negative.

Define two 3×3 symmetric matrices per line to represent generalized Lagrange multipliers for the line-flow limits measured from each line terminal: $\mathbf{H}_{k_l}$ and $\mathbf{H}_{k_m}$, with superscript cd indicating the (c, d) element of the corresponding matrix.

Define 2×2 symmetric matrices to represent the generalized Lagrange multipliers for the quadratic cost function associated with each generator: $\mathbf{R}_g$, with $\mathbf{R}_g^{cd}$ the (c, d) element of $\mathbf{R}_g$.

Finally, define a scalar real-valued function ρ and matrix-valued function $\mathbf{A}$.

$$\begin{aligned}\rho = \sum_{i \in \mathcal{N}} \Big\{ & \lambda_i P_{Di} + \underline{\gamma}_i Q_i^{\min} - \bar{\gamma}_i Q_i^{\max} + \underline{\mu}_i \left(V_i^{\min}\right)^2 - \bar{\mu}_i \left(V_i^{\max}\right)^2 \\ & + \sum_{g \in \mathcal{G}_i} \left(\underline{\psi}_g P_{Gg}^{\min} - \bar{\psi}_g P_{Gg}^{\max} + c_{g0} - \mathbf{R}_g^{12} \right) \Big\} \\ & - \sum_{k \in \mathcal{L}} \left\{ \left(S_k^{\max}\right)^2 \left(\mathbf{H}_{k_l}^{11} + \mathbf{H}_{k_m}^{11} \right) + \mathbf{H}_{k_l}^{22} + \mathbf{H}_{k_m}^{22} + \mathbf{H}_{k_l}^{33} + \mathbf{H}_{k_m}^{33} \right\}\end{aligned} \tag{6.13}$$

$$\begin{aligned}\mathbf{A} = & \sum_{i \in \mathcal{N}} \left\{ \lambda_i \mathbf{Y}_i + \left(\bar{\gamma}_i - \underline{\gamma}_i \right) \bar{\mathbf{Y}}_i + \left(\bar{\mu}_i - \underline{\mu}_i \right) \mathbf{M}_i \right\} \\ & + 2 \sum_{k \in \mathcal{L}} \left\{ \mathbf{H}_{k_l}^{12} \mathbf{Z}_{k_l} + \mathbf{H}_{k_m}^{12} \mathbf{Z}_{k_m} + \mathbf{H}_{k_l}^{13} \bar{\mathbf{Z}}_{k_l} + \mathbf{H}_{k_m}^{13} \bar{\mathbf{Z}}_{k_m} \right\}\end{aligned} \tag{6.14}$$

where

$$Q_i^{\max} = -Q_{Di} + \sum_{g \in \mathcal{G}_i} Q_{Gg}^{\max} \tag{6.15}$$

$$Q_i^{\min} = -Q_{Di} + \sum_{g \in \mathcal{G}_i} Q_{Gg}^{\min} \tag{6.16}$$

The semidefinite programing formulation of the dual OPF problem may then be written as

$$\max \quad \rho \tag{6.17a}$$

subject to

$$\mathbf{A} \succeq 0 \tag{6.17b}$$

$$\mathbf{H}_{k_l} \succeq 0, \quad \mathbf{H}_{k_m} \succeq 0 \qquad \forall k \in \mathcal{L} \tag{6.17c}$$

$$\mathbf{R}_g \succeq 0, \quad \mathbf{R}_g^{11} = 1 \qquad \forall g \in \mathcal{G} \tag{6.17d}$$

$$\left\{\lambda_i = c_{g1} + 2\sqrt{c_{g2}}\mathbf{R}_g^{12} + \bar{\psi}_g - \underline{\psi}_g \quad \forall g \in \mathcal{G}_i\right\} \quad \forall i \in \mathcal{N} \tag{6.17e}$$

$$\underline{\psi}_g \geq 0, \bar{\psi}_g \geq 0, \ \underline{\gamma}_i \geq 0, \ \bar{\gamma}_i \geq 0, \ \underline{\mu}_i \geq 0, \ \bar{\mu}_i \geq 0 \tag{6.17f}$$

where $\succeq 0$ indicates the corresponding matrix is positive semidefinite.

6.2.3 Discussion

Several aspects of the semidefinite programming formulation deserve special attention. We focus on those aspects that differ from previous formulations (e.g., [29]) due to the proposed formulation's allowing of multiple generators at the same bus and the possibility of off-nominal transformer voltage ratios and non-zero phase shifts.

The formulation given in (6.17) includes the possibility of multiple generators at the same bus. As shown in (6.17e), all generators at the same bus i must have the same aggregate active power Lagrange multiplier λ_i. This is related to the principle of equal marginal costs in the economic dispatch problem [17]. Since generator reactive power injections do not appear in the cost function of (6.1), reactive power Lagrange multipliers are only needed for each bus rather than for each generator at each bus. This is seen in (6.15) and (6.16), which determine the allowed range of bus i reactive power injection.

The formulation (6.17) also includes the possibility of parallel lines (i.e., multiple lines with the same terminal buses) and the ability to represent transformers with both off-nominal voltage ratios and non-zero phase-shifts. Previous formulations limited the total power flow between two buses in order to limit line flows, precluding the ability to separately limit line flows on parallel

lines, and solely used a Π-model, which does not incorporate off-nominal voltage ratios and non-zero phase-shifts. The additional modeling flexibility in the formulation in (6.17) comes at the price of additional complexity. Incorporating parallel lines removes the ability to form the line-flow matrices directly from the bus admittance matrix, instead requiring the more complicated expressions in (6.5), (6.6), (6.7), and (6.8). Incorporating off-nominal voltage ratios and non-zero phase-shifts breaks the symmetry of the Π-model such that different line-flow matrices are required for each line terminal (i.e., $\mathbf{Z}_{k_l}$ in (6.5) and $\bar{\mathbf{Z}}_{k_l}$ in (6.7) for active and reactive power flows measured from the sending terminal and $\mathbf{Z}_{k_m}$ in (6.6) and $\bar{\mathbf{Z}}_{k_m}$ in (6.8) for the receiving terminal).

For large system models, numerical difficulties in the semidefinite programming solver may prevent convergence to acceptable precision. We have found several practical techniques that reduce numerical difficulties with large systems. First, ignore engineering limits that will clearly not be binding at the solution. Many system models specify large values for limits that are intended to be unlimited, particularly for reactive power generation and line-flow limits. We do not incorporate terms corresponding to very large limits. Similarly, some generators have a linear cost function (i.e., $c_{g2} = 0$). The corresponding $\mathbf{R}_g$ matrix can be eliminated from the formulation.

Numerical difficulties often occur when the system model has very "tight" limits. For instance, the active power generation of a synchronous condenser is constrained to be zero. A second technique for reducing numerical difficulties is to use equality constraints rather than inequality constraints to model these limits. When the active power output of a generator is constrained to a very small range, fix the generator at the midpoint of this range and directly add the associated generation cost to the objective function.

6.3 Advances in Matrix Completion Decompositions

In this section, we describe several advances in the decomposition techniques used to reduce the computational burden of large semidefinite formulations of the OPF problem. First, we review the maximal clique decomposition introduced by Jabr [24]. Next, we present a decomposition algorithm that significantly reduces the required computation time of Jabr's method. We then describe a technique for obtaining the optimal voltage profile from the decomposed matrices. Although we

focus on Jabr's decomposition [24], these advances can be applied to Sojoudi and Lavaei's decomposition [43] as well. Finally, we present a modification to Jabr's decomposition that extends his method to general networks rather than only networks with admittance matrices that satisfy a definiteness requirement.

6.3.1 Overview of Jabr's Decomposition

Jabr's decomposition uses the matrix completion theorem [15]. Several graph theoretic definitions are required for understanding of this theorem. A "clique" is defined as a subset of the graph vertices where each vertex in the clique is connected to all other vertices in the clique. A "maximal clique" is a clique that is not a proper subset of another clique. A graph is "chordal" if each cycle of length four or more nodes has a chord, which is an edge connecting two nodes that are not adjacent in the cycle. See [24, 50] for more details.

The matrix completion theorem can now be stated. Let $\bar{\mathbf{A}}$ be a partial (i.e., not all entries of $\bar{\mathbf{A}}$ have known values) symmetric matrix with associated undirected graph. The matrix $\bar{\mathbf{A}}$ can be completed to a positive semidefinite matrix (i.e., the unknown entries of $\bar{\mathbf{A}}$ can be chosen such that $\bar{\mathbf{A}} \succeq 0$) if and only if the submatrices associated with each of the maximal cliques of the graph associated with $\bar{\mathbf{A}}$ are all positive semidefinite.

The matrix completion theorem allows replacing the single large $2n \times 2n$ positive semidefinite constraint (6.17b) by many smaller matrices that are each constrained to be positive semidefinite. This significantly reduces the problem size for large, sparse power networks.

Jabr [24] notes two important aspects of this decomposition that are relevant to our advances. First, since the maximal cliques can have non-empty intersection (i.e., contain some of the same buses), different matrices may contain terms that refer to the same term in the $2n \times 2n$ matrix. Therefore, linking constraints are required to force equality between terms that are shared between maximal cliques. To specify these linking constraints, Jabr recommends forming a "clique tree": a maximum weight spanning tree of a graph with nodes corresponding to the maximal cliques and the edge weights between each node pair given by the number of shared buses in each clique pair. Using a maximal weight spanning tree of this graph, which can be calculated using Prim's

algorithm [18], reduces the number of linking constraints required: equality constraints are only enforced between the appropriate terms in maximal cliques that are adjacent in the maximal weight spanning tree.

Second, the graphs corresponding to realistic power networks are not chordal. A chordal extension of the graph is thus required in order to use the matrix completion theorem. A chordal extension adds edges between non-physically connected nodes (i.e., edges in the chordal extension of the graph may exist between buses that are not connected by a line in the power system) to obtain a chordal graph. Jabr recommends obtaining a chordal extension using a Cholesky decomposition of the absolute value of the imaginary part of the network's admittance matrix. To minimize the total number of edges, Jabr recommends using a Cholesky decomposition with minimum fill-in obtained by a minimum degree ordering of the row/column indices [3].

6.3.2 Matrix Combination Algorithm

We first describe a modification to Jabr's decomposition that results in a significant computational speed improvement. This modification accounts for the trade-off between the size of maximal cliques and the number of linking constraints. Smaller maximal cliques generally reduce the total size of the positive semidefinite constrained matrices. The overlap between maximal cliques, as determined by the clique tree approach, establishes the number of linking constraints.

The size and computational burden of the optimization problem is determined by both the size of the positive semidefinite matrix constraints and the number of linking constraints. Since many common semidefinite program solvers, such as SeDuMi [47], CSDP [7], SDPA [54], and SDPT3 [49], use primal–dual methods that solve both the primal and dual problems simultaneously and since a primal constraint corresponds to a dual variable, an approximation of the "size" of the semidefinite program can be made by adding the total number of variables required to form the matrices with the number of linking constraints.

Jabr's decomposition uses a Cholesky decomposition with minimum fill-in to obtain small maximal cliques, thus obtaining a heuristic for minimizing the number of variables in the positive semidefinite matrix constraints. This approach does not account for the computational burden

imposed by the linking constraints. The literature provides theoretical support for the concept of reducing computational burden by combining matrices (see section 4 of [36]).

We next describe our matrix combination heuristic. Let L be a parameter specifying the maximum number of matrices. Consider a semidefinite program formed from the chordal extension of a power system network as in Jabr's decomposition, with maximal clique i containing d_i buses. Since the matrices corresponding to the maximal cliques are symmetric and contain both real and imaginary voltage components, matrix i (corresponding to maximal clique i) has $d_i\,(2d_i + 1)$ variables. If maximal cliques i and k, adjacent in the clique tree, share s_{ik} buses, then $s_{ik}\,(2s_{ik} + 1)$ linking constraints are required between the corresponding matrices. For each pair of adjacent maximal cliques in the clique tree, determine the change in the optimization problem "size" Δ_{ik} if the cliques i and k were combined, as given by

$$\begin{aligned}\Delta_{ik} = {} & d_{ik}\,(2d_{ik} + 1) - d_i\,(2d_i + 1) \\ & - d_k\,(2d_k + 1) - s_{ik}\,(2s_{ik} + 1)\end{aligned} \tag{6.18}$$

where $d_{ik} = d_i + d_k - s_{ik}$ is the number of buses in the combined clique.

While the number of matrices is greater than L, combine a pair of adjacent maximal cliques with smallest Δ_{ik}. Then recalculate the value of Δ_{ik} for all maximal cliques adjacent to the newly combined clique. Repeat until the number of matrices is equal to L. Constrain the resulting set of matrices to be positive semidefinite in the OPF formulation (6.17).

We test this heuristic using two system models: the IEEE 300-bus system [2], which has quadratic generator cost functions; and a 3012-bus model of the Polish system for evening peak demand in winter 2007-2008 [56], which has parallel lines, line-flow limits, buses with multiple generators, and linear generator cost functions. These systems were chosen since their increasing orders of magnitude in number of buses demonstrate how the heuristic scales with system size. Matrix combination techniques do not result in a notable speed improvement for small systems; no matrix combination approach reduced the computational time for the IEEE 30-bus system [2].

The formulation in (6.17) was implemented using YALMIP version 3 [33], SeDuMi version 1.3 [47], and MATLAB R2011a. A computer with an 64-bit Intel i7-2600 Quad Core CPU at 3.40 GHz with 16 GB of RAM was used to run the formulation. A tolerance of 1×10^{-9} for SeDuMi's "eps" was used in the calculation of these results.

Figures 6.1 and 6.2 show how the solver time (i.e., the time used by the semidefinite programming solver (SeDuMi)), varies with the parameter L. Note that these figures do not include the set up time required to initialize the formulation. However, particularly for large systems, the set up time is a small fraction (typically around 15% to 20%) of the solver time. Also note that solver times for the 3012-bus system are not available for $L < XXX$ due to lack of computational capability.

The solver times for Jabr's decomposition as described in [24] are the rightmost points (no matrix combinations) in Figures 6.1 and 6.2. As L decreases from the rightmost point, the solver times decrease by, at most, approximately a factor of 2.5 for the 300-bus system and a factor of XX for the 3012-bus system as compared to the solver time without combining matrices. The plots thus show that matrix combining can result in significant improvements in solver time. However, as L continues to decrease, the speed improvements from removing linking constraints are overcome by the additional variables required for the larger matrices (in the extreme, returning to a single $2n \times 2n$ matrix). Thus, the solver times exhibit a steep increase for small L.

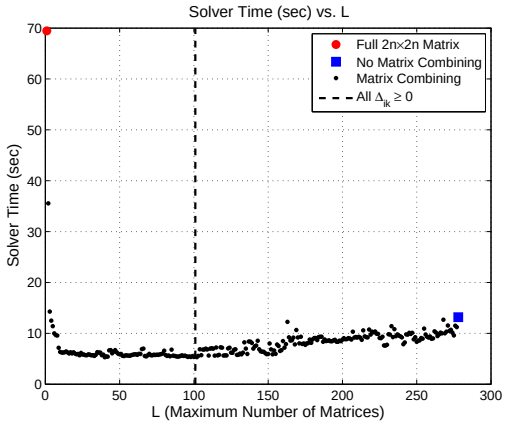

Figure 6.1 Solver time vs. L for IEEE 300-Bus System

Figure 6.2 Solver time vs. L for 3012-Bus System

Rather than combining matrices until below a specified parameter value, we also tried combining matrices until no pair of adjacent maximal cliques had a negative value of Δ_{ik} (i.e., stopping combining matrices once the heuristic indicated no further advantage to doing so). In our numerical experience, however, this approach did not always identify a set of matrices that minimized the solver time. In Figures 6.1 and 6.2, the number of matrices for which no remaining adjacent pairs of maximal cliques had negative Δ_{ik} is identified by the vertical dashed line. In Fig. 6.1, the vertical dashed line is very near the minimum solver time. However, for the 3012-bus system in Fig. 6.2, the dashed line does not occur near a minimum solver time; faster solver times were obtained for smaller values of L. This reinforces the fact that our measure of semidefinite program size is a heuristic approximation of the computational burden.

Based on these results, choosing L equal to approximately XXXX% of the initial number of matrices appears to give near minimum solver times. (Expressing L as a percentage of the original number of matrices allows for easy comparison between systems.)

Table 6.1 summarizes these results by providing the solver times for each system / decomposition pair along with a "speed up factor" (SUF) of the improvement of the matrix combination approach with $L = XX\%$ of the original number of matrices as compared to not combining matrices. Note that computational limitations precluded obtaining results from the full $2n \times 2n$ matrix for the 3012-bus system.

System	$2n \times 2n$	No Combining	Combining ($L = XX\%$)	SUF
30-bus	0	0	0	0
300-bus	69.453	13.182	5.366	2.46
3012-bus	–	0	0	0

Table 6.1 Solver Times (sec) for Various Algorithms

6.3.3 Obtaining the Optimal Voltage Profile

When using any decomposition algorithm, the solver returns a solution consisting of a set of positive semidefinite matrices. Existing literature indicates that if all the matrices have nullspaces with appropriate rank, the optimal voltage profile can be recovered [29, 24]. (For formulations that separate real and imaginary voltage components, like (6.17), the nullspace of all matrices must have rank less than or equal to two.) However, existing literature does not describe *how* to actually recover the optimal voltage profile. In this section, we describe a technique for obtaining the optimal voltage profile.

An overview of this technique follows. First obtain vectors in the nullspaces of each positive semidefinite constrained matrix. These vectors, when rearranged such that they correspond to complex "phasor" voltages, can each be multiplied by a different complex scalar and remain in the relevant nullspace. Obtaining the optimal voltage profile requires that the values of these complex scalars be chosen such that terms that refer to the same voltage are consistent and that the voltage angle at the reference bus is 0°. A linear calculation of the nullspace of an appropriately specified matrix gives a vector of such scalar values that satisfy these requirements. With valid values for these scalars, a vector is constructed from the scalars and the nullspace vectors that is itself a real scalar multiple of the optimal voltage profile. Using knowledge of a single voltage magnitude from a binding voltage limit, the resulting vector is scaled such that the known voltage magnitude value is obtained. The rescaled vector is then the optimal voltage profile.

We next present the mathematical details of this technique. Assume we have an optimal solution to (6.17) consisting of a set of d positive semidefinite matrices $\bar{\mathbf{A}}_i$ with $\operatorname{rank}\left(\operatorname{null}\left(\bar{\mathbf{A}}_i\right)\right) \leq 2 \; \forall i \in \{1, \ldots, d\}$. Let u_i be an eigenvector in the nullspace of $\bar{\mathbf{A}}_i$. Alternatively, both the primal $\bar{\mathbf{W}}_i$ and dual $\bar{\mathbf{A}}_i$ matrices are available if a primal–dual solver is used ($\bar{\mathbf{W}}_i$ is the positive semidefinite constrained matrix in the primal formulation and is the generalized Lagrange multiplier of the $\bar{\mathbf{A}}_i$ matrix in the dual formulation (6.17b); see [29] for further details). If, due to numeric problems in the solver, the primal matrices have a solution with better rank characteristics (i.e., the $\bar{\mathbf{W}}_i$ matrices are numerically closer to having rank less than or equal to two than the nullspace of

the $\bar{\mathbf{A}}_i$ matrices), the optimal voltage profile can be recovered by setting u_i equal to an eigenvector corresponding to a non-zero eigenvalue of $\bar{\mathbf{W}}_i$.

Let r_i be the number of buses in the maximal clique corresponding to matrix i. Convert each eigenvector from representing real and imaginary voltage components to complex "phasor" form: $\underline{u}_i = u_{di} + j u_{qi}$, where $u_{di} = u_{i,\,1:r_i}$, $u_{qi} = u_{i,\,r_i+1:2r_i}$, and subscript $1 : r_i$ indicates the first through r_i^{th} elements of the corresponding vector.

Each nullspace vector $\underline{u}_i$ can be multiplied by a complex scalar $\alpha_i + j\beta_i$. Thus, elements of $\underline{u}_i$ are linearly related to the complex voltages at the buses associated with the maximal clique from which the corresponding matrix was formed. Let $\underline{u}_{i,\,l}$ (i.e., the l^{th} element of the $\underline{u}_i$ vector) and $\underline{u}_{k,\,m}$ correspond to the same bus for some $i \neq k$. The optimal voltage profile is obtained when values of $\alpha_i + j\beta_i$ and $\alpha_k + j\beta_k$ are determined such that the voltage at the bus is consistent (that is, $(\alpha_i + j\beta_i)\,\underline{u}_{i,\,l} = (\alpha_k + j\beta_k)\,\underline{u}_{k,\,m}$) for all appropriate vector element pairs. In other words, obtaining the optimal voltage profile requires determining values for $\alpha_i + j\beta_i$ that create agreement between all terms representing the same voltage from the nullspace vectors of different matrices.

Obtaining such values of $\alpha_i + j\beta_i$ can be done by performing a linear calculation. Define a matrix $\mathbf{C}$ that is comprised of the elements of the u_i vectors in a particular arrangement. The matrix $\mathbf{C}$ has number of rows equal to $2d$ (twice the number of $\bar{\mathbf{A}}_i$ matrices) plus one, and number of columns equal to $2d$. The elements of the u_i vectors are arranged such that a vector in the nullspace of $\mathbf{C}^T\mathbf{C}$ is a real scalar multiple of a vector of the requisite values of α and β. The matrix $\mathbf{C}$ is specified as follows.

The first row of $\mathbf{C}$ sets the reference bus voltage to have zero angle. Let the m^{th} component of u_{ds} and u_{qs} correspond to the real and imaginary parts of the slack bus voltage. Then angle reference is enforced using the constraint

$$\beta_s u_{ds,\,m} + \alpha_s u_{qs,\,m} = 0 \tag{6.19}$$

This constraint is implemented in the first row of $\mathbf{C}$ by setting the $(1, s)$ entry of $\mathbf{C}$ to the m^{th} component of u_{qs} and the $(1, s + d)$ entry of $\mathbf{C}$ to the m^{th} component of u_{ds}.

The remaining rows of $\mathbf{C}$ are used to enforce consistency between pairs of terms from separate matrices that correspond to the same voltage component. Since both the real and imaginary components of the voltages must be consistent, two constraints, and thus two rows of the $\mathbf{C}$ matrix, are required for each such pair of terms. Let the l^{th} component of $\underline{u}_i$ and the m^{th} component of $\underline{u}_k$ correspond to the same voltage. Enforcing consistency requires

$$\alpha_i u_{di,\,l} - \beta_i u_{qi,\,l} - \alpha_k u_{dk,\,m} + \beta_k u_{dk,\,m} = 0 \tag{6.20}$$

$$\beta_i u_{di,\,l} + \alpha_i u_{qi,\,l} - \beta_k u_{dk,\,m} - \alpha_k u_{qk,\,m} = 0 \tag{6.21}$$

Constraint (6.20), which enforces consistency on the real voltage component, is implemented by setting the i, k, $(i+d)$, and $(k+d)$ entries of next row of the $\mathbf{C}$ matrix to $u_{di,\,l}$, $-u_{dk,\,m}$, $-u_{qi,\,l}$, and $u_{dk,\,m}$, respectively. Similarly, constraint (6.21), which enforces consistency on the imaginary voltage component, is implemented by setting the i, k, $(i+d)$, and $(k+d)$ entries of following row of the $\mathbf{C}$ matrix to $u_{qi,\,l}$, $-u_{qk,\,m}$, $u_{di,\,l}$, and $-u_{dk,\,m}$, respectively.

If all $\bar{\mathbf{A}}_i$ matrices of the solution have nullspaces with rank less than or equal to two, then $\mathbf{C}^T\mathbf{C}$ has a singe zero eigenvalue. (For a solution where some of the $\bar{\mathbf{A}}$ matrices have nullspaces with rank greater than two, the corresponding $\mathbf{C}^T\mathbf{C}$ matrix does not have a zero eigenvalue, thus indicating that a consistent voltage profile cannot be extracted from the solution.) Let η be an eigenvector corresponding to a zero eigenvalue of $\mathbf{C}^T\mathbf{C}$. Then η specifies a vector of α_i and β_i values that satisfy the angle reference constraint (6.19) and voltage consistency constraints (6.20) and (6.21). Specifically, the vector of α_i values is given by $\eta_{1:d}$ and the vector of β values is given by $\eta_{d+1:2d}$.

With known values of α_i and β_i, a vector U of length $2n$ that is a real scalar multiple of the optimal voltage profile is constructed by properly arranging the α_i, β_i and $\underline{u}_i$ values. For instance, if $\underline{u}_{i,\,k}$ corresponds to the voltage at bus m, then $U_m = \mathrm{Re}\left((\alpha_i + j\beta_i)\,\underline{u}_{i,\,k}\right)$ and $U_{m+n} = \mathrm{Im}\left((\alpha_i + j\beta_i)\,\underline{u}_{i,\,k}\right)$.

Since η has one degree of freedom in its length, the optimal voltage profile is a scalar multiple χ of U. To determine the value of χ, one additional piece of information is required: the voltage

magnitude at any bus. As argued in [29], at least one voltage magnitude constraint in the OPF problem is binding at a solution. A binding voltage magnitude constraint is identified by a non-zero value of the corresponding Lagrange multiplier (either $\underline{\mu}_i$ or $\bar{\mu}_i$ in (6.17)). After identifying a binding voltage magnitude constraint, the voltage magnitude at the bus is the corresponding voltage limit. Let $\bar{V}_k$ be the value of a binding voltage magnitude limit at bus k. The value of χ is chosen to obtain this voltage magnitude:

$$\chi = \sqrt{\frac{\bar{V}_k^2}{U_k^2 + U_{k+n}^2}} \tag{6.22}$$

The optimal voltage profile is then

$$V^{opt} = \chi \left(U_{1:n} + jU_{n+1:2n} \right) \tag{6.23}$$

6.3.4 Extending Jabr's Decomposition to All Systems

The first step in Jabr's decomposition is to form a chordal extension of the network using a Cholesky decomposition of the absolute value of the imaginary part of the bus admittance matrix associated with the network (i.e., $\mathrm{chol}\left(\left|\mathrm{Im}\left(\mathbf{Y}\right)\right|\right)$). A Cholesky decomposition can only be performed on a positive definite matrix. Since not all power system networks have admittance matrices that satisfy $\left|\mathrm{Im}\left(\mathbf{Y}\right)\right| \succ 0$ (e.g., networks with sufficiently large shunt capacitances, such as the 9-bus system in MATPOWER [56]), Jabr's method cannot be applied to all networks.

Jabr's method only uses the *sparsity pattern* (i.e., location of the non-zero elements) of the Cholesky decomposition. Thus, an alternative, positive definite matrix whose Cholesky decomposition exhibits the same sparsity pattern would extend Jabr's decomposition to general power systems. In this section, we present a such an alternative matrix.

Let $\mathbf{D}$ represent the incidence matrix associated with the network (i.e., each row of $\mathbf{D}$ corresponds to a line and has two non-zero elements: $+1$ in the column corresponding to the line's "from" bus and -1 in the column corresponding to the line's "to" bus). The matrix $\mathbf{E}$ in (6.24) has a Cholesky decomposition with the same sparsity pattern as $\left|\mathrm{Im}\left(\mathbf{Y}\right)\right|$.

$$\mathbf{E} = \mathbf{D}^T\mathbf{D} + \mathbf{I}_{n\times n} \tag{6.24}$$

where $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix.

Since $\mathbf{D}^T\mathbf{D}$ has a Laplacian structure, it is positive semidefinite. Adding an identity matrix increases all eigenvalues by one, and thus $\mathbf{E}$ is positive definite.

The bus admittance matrix $\mathbf{Y}$ has generalized Laplacian structure, with weightings from the line susceptances, plus diagonal terms corresponding to shunt admittances. The $\mathbf{E}$ matrix's similar construction implies that its Cholesky decomposition has the same sparsity pattern as the Cholesky decomposition of $|\mathrm{Im}\,(\mathbf{Y})|$. Using the Cholesky decomposition of $\mathbf{E}$ therefore extends Jabr's method to general power networks.

6.4 Conclusion and Future Work

This paper has addressed two categories of practical issues associated with implementing a large-scale optimal power flow solver based on semidefinite programming: modeling issues associated with realistic power system models and using network sparsity to reduce computational time via matrix completion decompositions. Specific modeling issues addressed include multiple generators at the same bus and parallel lines.

The paper includes three advances in computational aspects using matrix completion decompositions. First, a proposed matrix combination algorithm considers the impact of "linking constraints" between terms in certain decomposed matrices that refer to the same term in the original $2n \times 2n$ matrix. Since the linking constraints associated with two matrices are not necessary if the matrices are combined to form one matrix, the computational burden of the problem may decrease if matrices are combined. We propose an algorithm for matrix combination that takes a single parameter: the maximum allowed number of matrices. Although a relatively wide-range of choices for this parameter significantly reduce the solver time as compared to not combining matrices, numerical results indicate that a choice of this parameter at XX% of the number of matrices without using matrix combination typically results in near minimum solver time. Calculations for large systems shows the efficacy of the matrix combination approach: the IEEE 300-bus system shows

a factor of approximately 2.5 decrease in solver time and a 3012-bus model of the Polish system shows a factor of XXX decrease in solver time over not combining matrices.

The next advance in matrix decomposition approaches is a method for constructing the optimal voltage profile from a solution consisting of decomposed matrices. Although existing literature discusses the use of matrix decompositions [24, 43], it does not give a method for obtaining the optimal voltage profile.

Finally, Jabr's decomposition [24] is extended to general power system networks. Jabr's decomposition uses a Cholesky factorization of the absolute value of the imaginary part of the bus admittance matrix. Since a Cholesky factorization cannot be calculated for matrices that are not positive definite, this approach cannot be used for some networks (e.g., networks with large shunt capacitances). Jabr's decomposition only uses the sparsity pattern of the result of the Cholesky decomposition. We propose an alternative matrix guaranteed to be positive definite with the same sparsity pattern that therefore extends Jabr's method to general power system networks.

Future work on this topic includes investigation of alternative load models. Currently, the formulation only includes the capability for constant power and constant impedance load models. Another prevalent load model is constant current, which is not trivially incorporated into the formulation. Investigation of whether a constant current model can be included in a semidefinite programming-based OPF solver is thus future work.

Additional future work includes refinement and public release of the code used to obtain the results for this paper. We intend to release an extension to the research software MATPOWER [56] that allows users to easily specify a semidefinite program solver for the OPF problem. This will speed research progress by negating the need for every researcher to create their own semideifinite programming implementation and will quickly distribute the advances detailed in this paper.

Chapter 7

An Extended Bidding Structure and Economic Dispatch Model

7.1 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 [52]. 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 (before the market clearing price is known) so as to place meaningful price-quantity bids on the market [27]. 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 [42]. 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.

7.1.1 FERC's Ruling on Demand Response

In its recent Order No. 745 [13], 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 (and implementations) 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 [13, 41, 21]. 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 [13, 10, 11, 19, 20]. 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 [52, 13, 12, 6]. 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 (e.g. day-ahead) market. Instead, DR is treated as a sale of the "consuming right" from certain consumers (DR provider) to other consumers (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 [22]. The work in Chapter 2 instantiates 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 [16, 44] and their citing documents.

7.1.2 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. [4] 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. [46] 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. [48] 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 (which is particularly inspiring to our current work). However, those demand bids were modeled by integer variables and constraints, thus the dispatch mechanism fell short of good economic properties. [39] 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 [45] for a detailed discourse on market design. Following a brief note on the nomenclature in Section 7.1.3, Section 7.2 proposes our characterization of different demand types and their respective cost-minimizing or surplus-maximizing problems. Based on this, Section 7.3 develops the new bidding structure and the corresponding central dispatch model, accompanied by the proof of its incentive compatibility. Section 7.4 implements the model for an experiment and presents the experiment results, and Section 7.5 draws some conclusions and summarizes the points.

7.1.3 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.

7.2 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 her wholesale market representative, e.g. load serving entity (LSE), are unable to estimate accurately, see, e.g.,[27]. 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 7.1 illustrates a structural overview of our work.

[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.

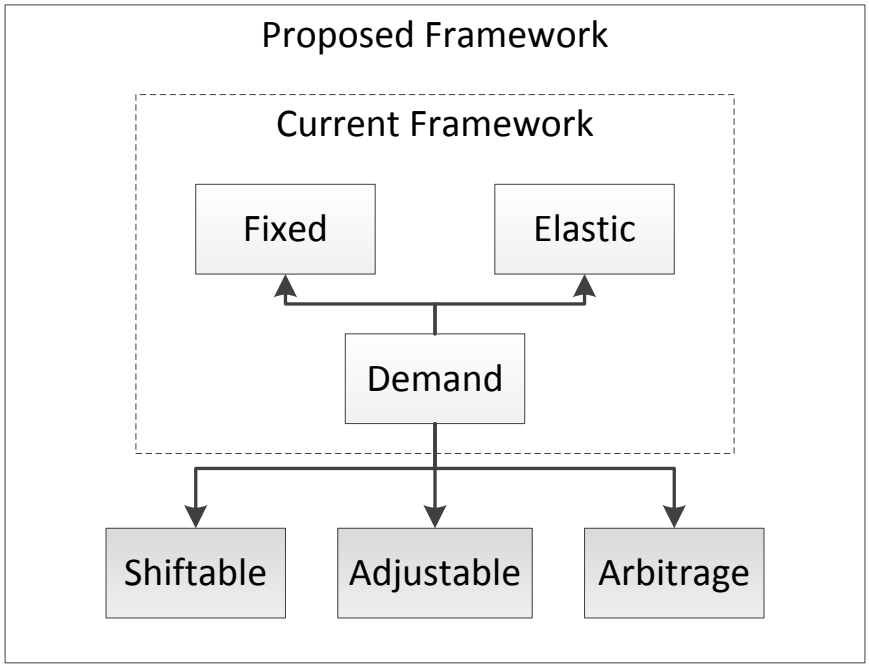

Figure 7.1 Framework for demand-side participation

7.2.1 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 [37]. 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.

7.2.2 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):

$$\max_{d_k} \quad \sum_t [V_{k,t}(d_{k,t}) - p_t d_{k,t}] \tag{7.1}$$

$$\text{s.t.} \quad \underline{d}_{k,t} \leq d_{k,t} \leq \bar{d}_{k,t}, \ \forall t \tag{7.2}$$

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

7.2.3 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 $d^{\text{ta}}_{k,t}$, 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 [27, 45]. An adjustable demander minimizes the cost of consumption by solving (ADJ)(p):

$$\min_{r_k^+, r_k^-} \quad \sum_t [p_t(d_{k,t}^{\text{ta}} + r_{k,t}^+ - r_{k,t}^-) + D_{k,t}(r_{k,t}^+, r_{k,t}^-)] \tag{7.3}$$

$$\text{s.t.} \quad 0 \leq r_{k,t}^+ \leq \bar{r}_{k,t}^+, \ \forall t \tag{7.4}$$

$$0 \leq r_{k,t}^- \leq \bar{r}_{k,t}^-, \ \forall t \tag{7.5}$$

Realistically, the parameter $\bar{r}_{k,t}^-$ is upper bounded by $d_{k,t}^{\text{ta}}$. 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:

$$D_{k,t}(r_{k,t}^+, r_{k,t}^-) = \alpha_{k,t}^+ (r_{k,t}^+)^2 + \beta_{k,t}^+ |r_{k,t}^+| + \alpha_{k,t}^- (r_{k,t}^-)^2 + \beta_{k,t}^- |r_{k,t}^-| \tag{7.6}$$

where α and β are parameters.

7.2.4 Shiftable Demand

Shiftable demand requires a total amount of electricity to be delivered within a given time range, and 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 $d_{k,m}^{\text{tr}}$ amount to be delivered within the time range $T_{k,m} \subset T$. A shiftable demander minimizes her consumption cost by solving (SHI)(p):

$$\min_{d_k^{\text{sh}}} \quad \sum_t p_t d_{k,t}^{\text{sh}} \tag{7.7}$$

$$\text{s.t.} \quad \sum_{t \in T_{k,m}} d_{k,t}^{\text{sh}} = d_{k,m}^{\text{tr}}, \ \forall m, T_{k,m} \tag{7.8}$$

$$\underline{d}_{k,t}^{\text{sh}} \leq d_{k,t}^{\text{sh}} \leq \bar{d}_{k,t}^{\text{sh}}, \ \forall t \tag{7.9}$$

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.

7.2.5 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 [51]. 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):

$$\max_{b_k, s_k, h_k} \quad \sum_t p_t (s_{k,t} - b_{k,t}) \tag{7.10}$$

$$\text{s.t.} \quad h_{k,t} = h_{k,t-1} + b_{k,t} e_k - s_{k,t}, \qquad \forall t \tag{7.11}$$

$$h_{k,1} = h_{k,|T|} \tag{7.12}$$

$$0 \leq b_{k,t} \leq \bar{b}_k, \qquad \forall t \tag{7.13}$$

$$0 \leq s_{k,t} \leq \bar{s}_k, \qquad \forall t \tag{7.14}$$

$$0 \leq h_{k,t} \leq \bar{h}_k, \qquad \forall t \tag{7.15}$$

In the defining equation (7.11) 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 [23], which needs more constraints to express. For expositional purpose, we make e_k a constant bidding parameter. Constraint (7.12) 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.

7.3 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.

7.3.1 Central Model and its Properties

Table 7.1 lists the parameters and variables in the model, with subscripts omitted for clarity. The parameters represent the bids submitted to the system operator.

Table 7.1 Bidding Parameters and Decision Variables

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

In a distributed decision-making paradigm, given the market clearing prices p_t, demanders solve their respective behaviorial models presented in the last section. On a similar basis, generator k responds to the price p by solving (GEN)(p):

$$\max_{g_k} \quad \sum_t [p_t g_{k,t} - C_{k,t}(g_{k,t})] \tag{7.16}$$

$$\text{s.t.} \quad \underline{g}_{k,t} \le g_{k,t} \le \bar{g}_{k,t}, \ \forall t \tag{7.17}$$

$$g_{k,t} - g_{k,t-1} \le R_k^{\mathrm{U}}, \ \forall t \tag{7.18}$$

$$g_{k,t-1} - g_{k,t} \le R_k^{\mathrm{D}}, \ \forall t \tag{7.19}$$

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

$$\sum_k (g_{k,t} - d_{k,t} - d_{k,t}^{\mathrm{sh}} - r_{k,t}^+ + r_{k,t}^- + s_{k,t} - b_{k,t}) = \sum_k (d_{k,t}^{\mathrm{fx}} + d_{k,t}^{\mathrm{ta}}), \qquad \forall t \tag{7.20}$$

by adjusting the prices p_t.

We postulate a central dispatch model, as follows.

(Central Model):

$$\begin{aligned} \min_{\substack{g,d,d^{\text{sh}},r^+\\ r^-,b,s,h}} \quad & \sum_{k,t}[C_{k,t}(g_{k,t}) - V_{k,t}(d_{k,t}) + D_{k,t}(r_{k,t})] \\ \text{s.t.} \quad & \text{(7.2), (7.8), (7.9), (7.4), (7.5)} \\ & \text{(7.11)-(7.15), (7.17)-(7.20)} \end{aligned}$$

The price p_t is set as the optimal Lagrangian multiplier (or dual variable) of the corresponding constraint in (7.20). Note that the model minimizes the total social cost (negative of the social welfare), hence it is economically efficient.

Theorem 7.1 Given a set of bidding parameters, suppose that $\hat{x} := (\hat{g}, \hat{d}, \hat{d}^{\text{sh}}, \hat{r}^+, \hat{r}^-, \hat{b}, \hat{s}, \hat{h})$ solves the Central Model and $\hat{p}$ is the optimal Lagrangian multiplier of the constraint (7.20). Then $\hat{g}$ solves (GEN)($\hat{p}$), $\hat{d}$ solves (ELA)($\hat{p}$), $\hat{d}^{\text{sh}}$ 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 (7.20), of the Central Model:

$$\begin{aligned} \max_{p}\min_{x} \quad & \sum_{k,t}[C_{k,t}(g_{k,t}) - V_{k,t}(d_{k,t}) + D_{k,t}(r_{k,t})] \\ & + \sum_t p_t[\sum_k (g_{k,t} - d_{k,t} - d^{\text{sh}}_{k,t} - r^+_{k,t} + r^-_{k,t} \\ & \qquad + s_{k,t} - b_{k,t} - d^{\text{fx}}_{k,t} - d^{\text{ta}}_{k,t})] \\ \text{s.t.} \quad & \text{(7.2), (7.8), (7.9), (7.4), (7.5), (7.11)-(7.15), (7.17)-(7.19)} \end{aligned}$$

Consequently, $\hat{x}$ solves

$$\begin{aligned} \min_{x} \quad & \sum_{k,t}[C_{k,t}(g_{k,t}) - V_{k,t}(d_{k,t}) + D_{k,t}(r_{k,t})] \\ & + \sum_t \hat{p}_t[\sum_k (g_{k,t} - d_{k,t} - d^{\text{sh}}_{k,t} - r^+_{k,t} + r^-_{k,t} \\ & \qquad + s_{k,t} - b_{k,t} - d^{\text{fx}}_{k,t} - d^{\text{ta}}_{k,t})] \\ \text{s.t.} \quad & \text{(7.2), (7.8), (7.9), (7.4), (7.5), (7.11)-(7.15), (7.17)-(7.19)} \end{aligned}$$

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. Q.E.D.

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 [45]. This leads to the conclusion that the extended Central Model is incentive compatible.

7.3.2 Abstraction

While the specific formats proposed above focus on the demand side, the structure can be applied to both sides of the market (e.g., 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

$$\max_{x_k \in X_k} \quad f_k(x_k) - x_k^\top p \tag{7.21}$$

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

$$\max_{x} \quad \sum_k f_k(x_k) \tag{7.22}$$

$$\text{s.t} \quad \sum_k x_k = 0 \ (\perp p) \tag{7.23}$$

$$x_k \in X_k, \ \forall k \tag{7.24}$$

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 (from the analysis in previous sections) 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.

7.3.3 Two Additional Merits

There are two related points that we need to clarify:

7.3.3.1 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:

$$z_{k,l,t} - B_{k,l}(\delta_{l,t} - \delta_{k,t}) = 0,\ \forall (k,l) \in \mathcal{A}, t \tag{7.25}$$

$$g_{k,t} - d_{k,t} - d^{\text{sh}}_{k,t} - r^{+}_{k,t} + r^{-}_{k,t} + s_{k,t} - b_{k,t} - \sum_{l:(k,l)\in\mathcal{A}} z_{k,l,t} = d^{\text{fx}}_{k,t} + d^{\text{ta}}_{k,t},\ \forall k, t \tag{7.26}$$

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

7.3.3.2 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. $d^{\text{fx}}, \bar{d}, \bar{d}^{\text{sh}}, d^{\text{ta}}, \bar{b}$, 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 [25, 38], which is part of the reason why unit commitment and economic dispatch are usually practiced as two decision processes rather than one.

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

7.4.1 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 [28]. For the demand data, we sum up the fixed demand bids

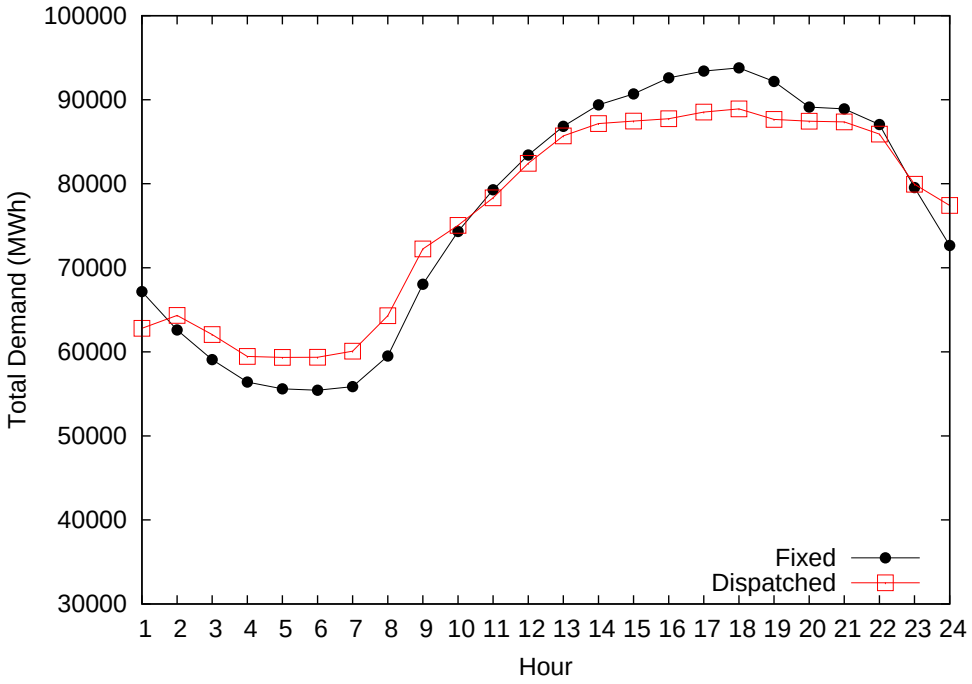

Figure 7.2 Day-ahead demand profile of FERC dataset 4012

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 7.2 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, and various unit commitment requirements and costs. A unit commitment process similar to the one documented in [28] 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.

[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.

7.4.1.1 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 7.3 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 minimization problem, as follows.

$$V_t(d_t) = \sum_{o \in \mathcal{O}} p^{\text{db}}_{t,o} d^{\text{db}}_{t,o} \tag{7.27}$$

$$d_t = \sum_{o \in \mathcal{O}} d^{\text{db}}_{t,o} \tag{7.28}$$

$$d^{\text{db}}_{t,o} \leq \bar{d}^{\text{db}}_{t,o}, \ \forall o \in \mathcal{O} \tag{7.29}$$

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

7.4.1.2 Adjustable Demand

We assume 1% of each hour's base-case demand becomes the target level p^{ta}_t of the adjustable demand. The deviation function $D_t(r^+_t, r^-_t)$ is taken in the form of (7.6), 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^{ta}_t while $\bar{r}^-_t$ is set to $\sum_t p^{\text{ta}}_t$.

7.4.1.3 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^{tr}_m for range m to be 1% of the sum of the hourly base-case demand in the range.

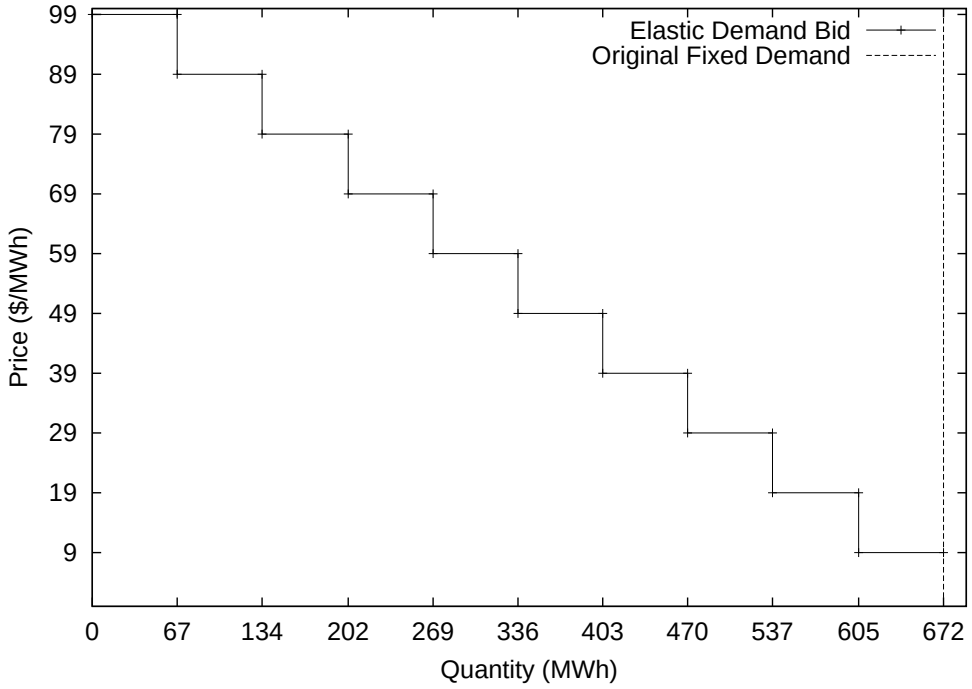

Figure 7.3 Elastic demand bid for hour 1

7.4.1.4 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.

7.4.2 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 7.2 as the "Dispatched" curve.

7.4.2.1 Effect on the LMP

Figure 7.4 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, and elastic demand is the least effective, in terms of dampening the price fluctuation.

7.4.2.2 Effect on the Social Welfare

Table 7.2 lists the cost (negative of the social welfare) results. The first column indicates the hypothesized market composition, the second column is the cost from the current bidding design, i.e. treating all demand as fixed, the third column is the optimal cost from our proposed bidding

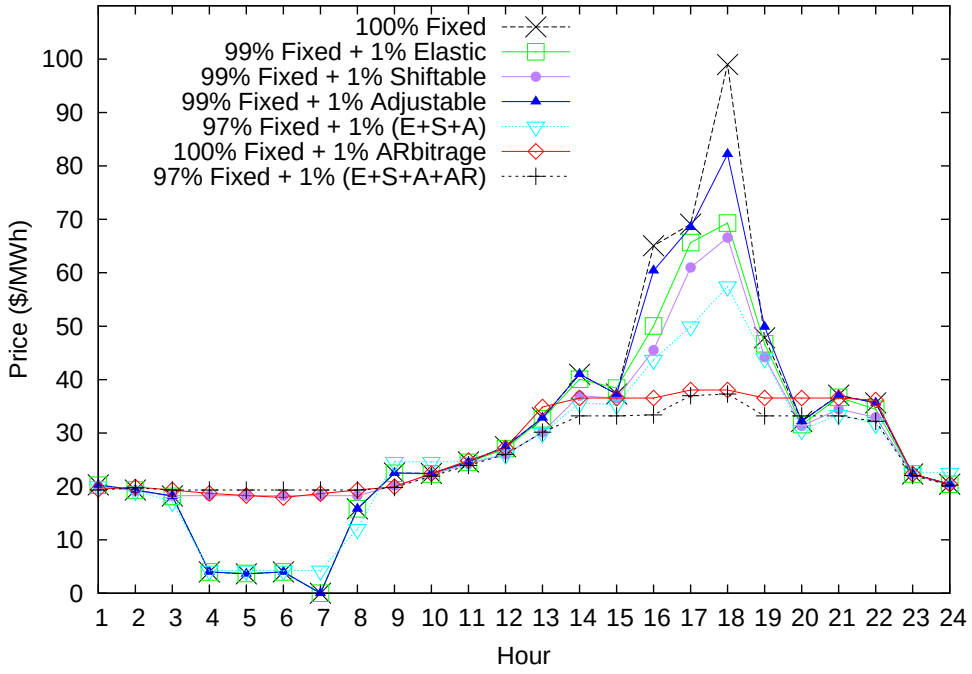

Figure 7.4 Effect of extended bidding on LMP

Table 7.2 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%

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.

7.4.3 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, i.e. 1%. This is fathomable, as an arbitrageur's buy/sell schedule is driven solely by the temporal price differences, and 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 7.5, 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, e.g., the first 0.2% increment of the arbitrage level contributed about half of the peak price reduction. This observation prompts a question: what is the "optimal" percentage of storage on the market? Figure 7.6 below provides some useful information to address this question.

In Figure 7.6, 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, i.e. 0.65, 0.75 and 0.85.

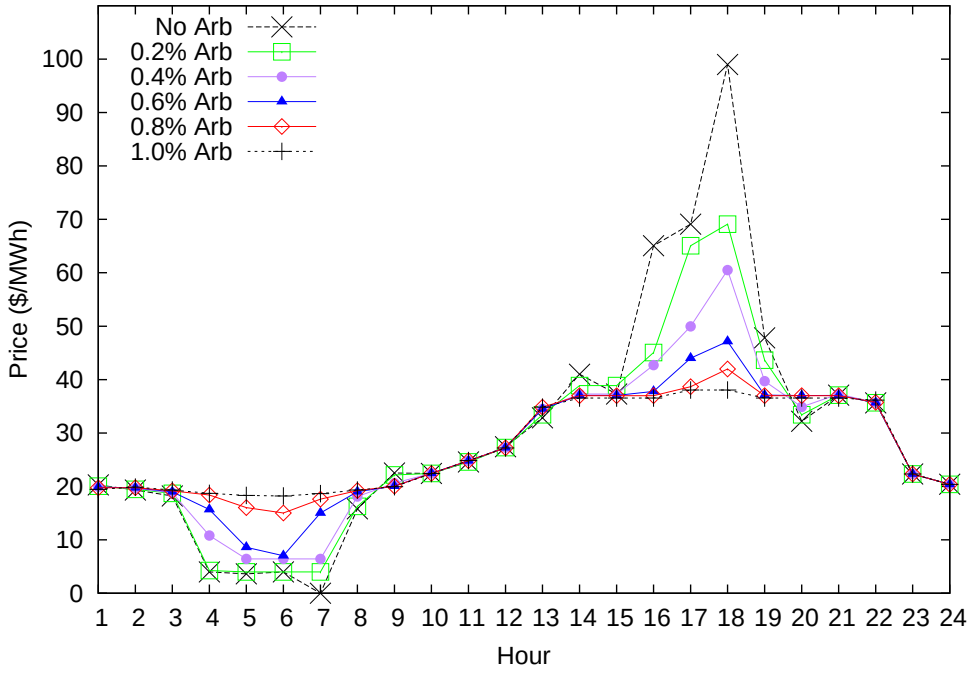

Figure 7.5 LMP for different arbitrage levels

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.

7.5 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 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.

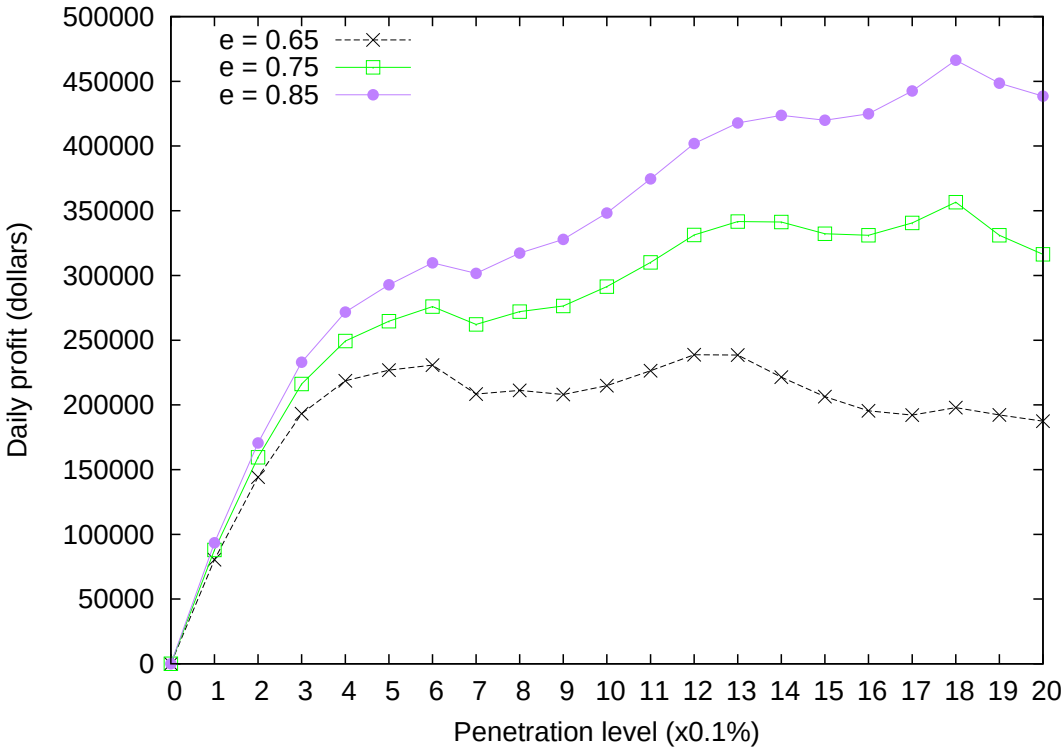

Figure 7.6 Profit of arbitrage for different penetration and efficiency

LIST OF REFERENCES

[1]

[2] Power Systems Test Case Archive.

[3] P.R. Amestoy, T.A. Davis, and I.S. Duff. Algorithm 837: AMD, An Approximate Minimum Degree Ordering Algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 30(3):381–388, 2004.

[4] José Manuel Arroyo and Antonio J. Conejo. Multiperiod auction for a pool-based electricity market. *IEEE Transactions on Power Systems*, 17(4):1225–1231, November 2002.

[5] X. Bai, H. Wei, K. Fujisawa, and Y. Wang. Semidefinite Programming for Optimal Power Flow Problems. *International Journal of Electrical Power & Energy Systems*, 30(6-7):383–392, 2008.

[6] Richard N. Boisvert, Peter A. Cappers, and Bernie Neenan. The benefits of customer participation in wholesale electricity markets. *The Electricity Journal*, 15(3):41 – 51, 2002.

[7] B. Borchers and J.G. Young. Implementation of a Primal–Dual Method for SDP on a Shared Memory Parallel Architecture. *Computational Optimization and Applications*, 37(3):355–369, 2007.

[8] S. Bose, D.F. Gayme, S. Low, and K.M. Chandy. Optimal Power Flow Over Tree Networks. In *49th Annual Allerton Conference on Communication, Control, and Computing, 2011*, Sept. 28-30 2011.

[9] J. Carpentier. Contribution to the Economic Dispatch Problem. *Bull. Soc. Franc. Elect*, 8(3):431–447, 1962.

[10] Hung-po Chao. Demand management in restructured wholesale electricity markets, May 2010.

[11] Hung-po Chao. Demand response in wholesale electricity markets: the choice of customer baseline. *Journal of Regulatory Economics*, 39:68–88, 2011.

[12] FERC. Wholesale competition in regions with organized electric markets, October 2008. http://www.ferc.gov/whats-new/comm-meet/2008/101608/E-1.pdf.

[13] FERC. Demand response compensation in organized wholesale energy markets, March 2011. http://www.ferc.gov/EventCalendar/Files/20110315105757-RM10-17-000.pdf.

[14] Drew Fudenberg and Jean Tirole. *Game Theory*, volume 1 of *MIT Press Books*. The MIT Press, 07 1991.

[15] M. Fukuda, M. Kojima, K. Murota, K. Nakata, et al. Exploiting Sparsity in Semidefinite Programming via Matrix Completion I: General Framework. *SIAM Journal on Optimization*, 11(3):647–674, 2001.

[16] J.K. Galbraith. *American Capitalism: The Concept of Countervailing Power*. Classics In Economics Series. Transaction Pub, 1980.

[17] J.D. Glover, M.S. Sarma, and T.J. Overbye. *Power System Analysis and Design*. Thompson Learning, 2008.

[18] J.L. Gross and J. Yellen. *Graph Theory and its Applications*. CRC press, 2006.

[19] William W. Hogan. Providing incentives for efficient demand response, October 2009. FERC Docket EL09-68-000.

[20] William W. Hogan. Demand response pricing in organized wholesale markets, 2010a. FERC Docket RM10-17-000.

[21] William W. Hogan. Economists' brief on ferc order 745 regarding demand response compensation, 2012.

[22] ISO New England. Market Rule 1 Appendix E, November 2012. http://www.iso-ne.com/regulatory/tariff/sect_3/.

[23] F. Novachek J. Himelic. Sodium sulfur battery energy storage and its potential to enable further integration of wind (wind-to-battery project), December 2011.

[24] R. A. Jabr. Exploiting Sparsity in SDP Relaxations of the OPF Problem. *IEEE Transactions on Power Systems*, PP(99):1, 2011.

[25] Raymond B. Johnson, Shmuel S. Oren, and Alva J. Svoboda. Equity and efficiency of unit commitment in competitive electricity markets. *Utilities Policy*, 6(1):9 – 19, 1997.

[26] S. Kim, M. Kojima, M. Mevissen, and M. Yamashita. Exploiting Sparsity in Linear and Nonlinear Matrix Inequalities via Positive Semidefinite Matrix Completion. *Mathematical Programming*, 129(1):33–68, 2011.

[27] Daniel S. Kirschen. Demand-side view of electricity markets. *IEEE Transactions on Power Systems*, 18(2):520–527, May 2003.

[28] Eric Krall, Michael Higgins, and Richard P. O'Neill. Rto unit commitment test system, July 2012.

[29] J. Lavaei and S.H. Low. Zero Duality Gap in Optimal Power Flow Problem. *IEEE Transactions on Power Systems*, 27(1):92–107, Feb. 2012.

[30] J. Lavaei, D. Tse, and B. Zhang. Geometry of Power Flows in Tree Networks. In *To appear in 2012 IEEE Power & Energy Society General Meeting*, July 22-27 2012.

[31] B.C. Lesieutre and I.A. Hiskens. Convexity of the Set of Feasible Injections and Revenue Adequacy in FTR Markets. *IEEE Transactions on Power Systems*, 20(4):1790 – 1798, November 2005.

[32] Bernard C. Lesieutre, Daniel K. Molzahn, Alex R. Borden, and Christopher L. DeMarco. Examining the Limits of the Application of Semidefinite Programming to Power Flow Problems. In *49th Annual Allerton Conference on Communication, Control, and Computing, 2011*, Sept. 28-30 2011.

[33] J. Lofberg. YALMIP: A Toolbox for Modeling and Optimization in MATLAB. In *IEEE International Symposium on Computer Aided Control Systems Design, 2004*, pages 284–289. IEEE, 2004.

[34] D.K. Molzahn, B.C. Lesieutre, and C.L. DeMarco. A Sufficient Condition for Power Flow Insolvability with Applications to Voltage Stability Margins. *Submitted to IEEE Transactions on Power Systems*, 2012.

[35] Holzer J.T. Lesieutre B.C. Molzahn, D.K. and C.L. DeMarco.

[36] K. Nakata, K. Fujisawa, M. Fukuda, M. Kojima, and K. Murota. Exploiting Sparsity in Semidefinite Programming via Matrix Completion II: Implementation and Numerical Results. *Mathematical Programming*, 95(2):303–327, 2003.

[37] Sam Newell and Attila Hajos. Demand response in the midwest iso - an evaluation of wholesale market design, January 2010.

[38] Richard P. O'Neill, Paul M. Sotkiewicz, Benjamin F. Hobbs, Michael H. Rothkopf, and William R. Stewart. Efficient market-clearing prices in markets with nonconvexities. *European Journal of Operational Research*, 164(1):269–285, 2005.

[39] D. Papadaskalopoulos, P. Mancarella, and G. Strbac. Decentralized, agent-mediated participation of flexible thermal loads in electricity markets. In *Intelligent System Application to Power Systems (ISAP), 2011 16th International Conference on*, pages 1–6, 2011.

[40] Zhifeng Qiu, G. Deconinck, and R. Belmans. A Literature Survey of Optimal Power Flow Problems in the Electricity Market Context. In *Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES*, pages 1–6, March 2009.

[41] Larry E. Ruff. Economic principles of demand response in electricity, October 2002.

[42] Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, and Roger E. Bohn. *Spot Pricing of Electricity*. Kluwer international series in engineering and computer science: Power electronics & power systems. Kluwer Academic, 1988.

[43] S. Sojoudi and J. Lavaei. Physics of Power Networks Makes Hard Optimization Problems Easy To Solve. In *To appear in 2012 IEEE Power & Energy Society General Meeting*, July 22-27 2012.

[44] George J. Stigler. The economist plays with blocs. *The American Economic Review*, 44(2):pp. 7–14, 1954.

[45] S. Stoft. *Power System Economics: Designing Markets for Electricity*. IEEE Press. IEEE Press, 2002.

[46] G. Strbac and D. Kirschen. Assessing the competitiveness of demand-side bidding. *Power Systems, IEEE Transactions on*, 14(1):120–125, 1999.

[47] J.F. Sturm. Using SeDuMi 1.02, A MATLAB Toolbox for Optimization Over Symmetric Cones. *Optimization Methods and Software*, 11(1):625–653, 1999.

[48] Chua-Liang Su and Daniel Kirschen. Quantifying the effect of demand response on electricity markets. *IEEE Transactions on Power Systems*, 24:1199 – 1207, 2009.

[49] R.H. Tütüncü, K.C. Toh, and M.J. Todd. Solving Semidefinite-Quadratic-Linear Programs using SDPT3. *Mathematical Programming*, 95(2):189–217, 2003.

[50] G. Valiente. *Algorithms on Trees and Graphs*. Springer Verlag, 2002.

[51] Rahul Walawalkar, Jay Apt, and Rick Mancini. Economics of electric energy storage for energy arbitrage and regulation in New York. *Energy Policy*, 35(4):2558 – 2568, 2007.

[52] Hon. Jon Wellinghoff and David L. Morenoff. Recognizing the importance of demand response: The second half of the wholesale electric market equation. *Energy Law Journal*, 28(2), 2007.

[53] L. A. Wolsey. *Integer programming*. Wiley-Interscience, New York, NY, USA, 1998.

[54] M. Yamashita, K. Fujisawa, M. Fukuda, K. Kobayashi, K. Nakata, and M. Nakata. Latest Developments in the SDPA Family for Solving Large-Scale SDPs. *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 687–713, 2012.

[55] B. Zhang and D. Tse. Geometry of Feasible Injection Region of Power Networks. In *49th Annual Allerton Conference on Communication, Control, and Computing, 2011*, Sept. 28-30 2011.

[56] R.D. Zimmerman, C.E. Murillo-Sánchez, and R.J. Thomas. MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education. *IEEE Transactions on Power Systems*, (99):1–8, 2011.