Investment under Dynamic Competition in Wholesale Electricity Markets

Joseph A. Cullen∗  Stanley S. Reynolds†

January 29, 2015

Abstract

A model of capacity investment in electricity market competition is developed and analyzed that incorporates two important features of generation technology: minimum generation rates and generation unit start-up costs. These technology features may have important implications for competitive outcomes, but can present difficulties for market analysis since they yield production non-convexities for firms. The model allows non-convexities at the firm level, while implying a convex aggregate production set, via an assumption that firms are ‘small’ relative to market size. A correspondence between competitive market equilibrium and the solution to a planner’s problem is established for both short-run (exogenously fixed generation capacities) and long-run (endogenous capacities) dynamic models. The model is solved to understand the impact of environmental regulation on short-run profits and long-run investment when firms compete under dynamic constraints.

∗Washington University
†University of Arizona, P.O. Box 210108, Tucson, AZ 85721, reynolds@eller.arizona.edu
1 Introduction

Many economic analyses of the performance of wholesale electricity markets utilize static models of perfect competition. This type of analysis is based on the grid-stack for generation units, which orders units from those with lowest marginal cost (MC) to those with highest MC and uses unit generation capacities to create a short run supply curve for the market. Mansur [2008] refers to this approach as the competitive benchmark analysis. This approach has been the basis for assessments of exercise of market power by generation suppliers in numerous studies such as Borenstein, Bushnell, Wolak 2002 and Joskow and Kahn [2002]. Long-run versions of this approach have been the foundation for assessments of regulatory policy changes, such as the introduction of real-time pricing (RTP) for retail customers (see Borenstein and Holland [2005], Borenstein [2005]).

A potential limitation of competitive benchmark analysis is its neglect of some technological features of electricity generation technology. For instance, Harvey and Hogan [2001] argue that tests of market power in wholesale electricity markets should recognize that generation units are subject to minimum operating rate constraints and must incur start-up costs each time they begin operating.

It is well understood that the grid-stack model of generation supply abstracts from numerous features of real-world electricity generation and distribution technology. Features that are potentially important for market performance include transmission and distribution system constraints, minimum and maximum output rate constraints for generation units, generation unit ramping constraints that limit changes in the rate of generation, and generation unit start-up costs. Harvey and Hogan [2001] are critical of the competitive benchmark approach used in Borenstein et al. [2002] and Joskow and Kahn, arguing that the apparent large price-cost margins estimated in these papers could well be due to unmodeled technological constraints. Mansur [2008] explores this issue further in an analysis of mark-ups in PJM. He estimates a reduced form model of system marginal costs that takes into account technological constraints that are ignored by the competitive benchmark approach. He finds that this reduced form approach yields significantly smaller estimates of price-cost mark-ups.
than does the competitive benchmark approach.¹ Also, there is a large literature in power systems engineering that examines how a utility or electricity system should be operated and managed, taking these technological constraints into account. A common approach involves a so-called unit commitment model, which allows optimization methods to be applied to a system with multiple generation units [see Bouffard et al. [2005] and Hobbs et al. [2002]]. However, incorporating rich detail about technology comes at a cost - unit commitment models typically have relatively short time horizons and are limited in how stochastic elements are modeled.

In this paper we propose a dynamic model of a wholesale electricity market that incorporates two key features of electricity generation technology: minimum operating rate constraints and generation unit start-up costs. The model allows for multiple types of generation technologies, each with its own marginal generation cost, minimum generation rate, and start-up cost. The only exogenous uncertainty in the model is the stochastic process for demand shocks.² The assumptions regarding generation introduce non-convexities into production technology, which in turn complicates market analysis and may lead to non-existence of competitive equilibrium. In order to pursue our objective of a competitive market analysis that incorporates these technology features, we introduce a formulation in which non-convexities are permitted at the firm (or, generation unit) level, but for which the aggregate production technology is convex. We assume that individual firms are small relative to the size of the market; specifically, firms are assumed to be of measure zero in the formal model. In this regard we follow the approach to dynamic competitive market analysis in Jovanovic [1982], Hopenhayn [1990], and Hopenhayn [1992]. Their models allow for non-convex production technology at the firm level, but yield a convex aggregate production technology. While the ‘zero measure’ assumption is clearly inconsistent with the presence of large suppliers in some

¹Our formulation is somewhat simpler than those of Hopenhayn [1990, 1992], which allow for both aggregate uncertainty and firm-specific shocks. An alternative that would be more in line with his models would be to include stochastic firm-specific start-up costs for ‘off’ units, and stochastic firm-specific fixed costs for ‘on’ units. Indeed, this is the approach taken by Cullen [2010] in his structural estimation of generator start-up costs based on ERCOT data. However, including firm-specific shocks is not necessary in order to derive interesting implications for a dynamic model of an electricity market with non-convexities.

²
wholesale electricity markets, we will attempt to convince the reader that this formulation provides a useful competitive benchmark that captures empirically relevant non-convexities.

This work is most closely related to Cullen [2010] which estimates a dynamic structural model of generation costs. Importantly, he uses estimated generation and start-up costs to numerically simulate a dynamic competitive equilibrium, which he uses for counterfactual policy analysis. However, there is no guarantee that equilibrium exists in his model or that the computation approach used will converge to an equilibrium. We improve on this work by showing that the short-run competitive equilibrium allocations are solutions to a social planner’s problem, and vice versa. This result provides an avenue for showing existence of competitive equilibrium, since we are able to show existence of a solution to the planner’s problem. Equally importantly, this correspondence between equilibrium and planner’s solution allows us to use the planner’s problem as a platform for computation and numerical analysis.

We extend the literature by showing that correspondence between equilibrium and planner’s solution extends to the long run scenario where firms invest endogenously in generating capital given their expectations about competition in the future. This allows us to address of important policy and market design questions as they relate the long-run equilibrium investment choices of firms. The rest of the paper proceeds as follows. In section 2 we characterize the model of competition. In section 3, we develop the connection between market equilibrium and the solution to the planner’s problem. In section 4, the model is extended to allow generation capacities to be endogenous. In section 5, we compute the solution to a calibrated version of the model to motivate further policy analysis. Section ?? concludes.

2 A Model of Wholesale Electricity Competition

The model encompasses a sequence of operating decisions regarding generation units; which units to start up, which units to shut down, and generation rates for ‘on’ units. There is an infinite sequence of time periods, with per-period discount factor, $\delta \in (0, 1)$. For the
applications in Section 5 the period is one hour, which yields a discount factor very close to one. In Sections 2, 3 and ?? we assume there is an exogenously fixed vector of generation capacities. This formulation captures short-run market dynamics for fixed capacities. In Section 4 we extend the model to allow for entry and investment in new generation capacity. This extended formulation captures long-run market dynamics.

2.1 Demand and Consumer Welfare

Wholesale demand varies across time periods according to the value of a demand shock (or, shift) variable, $\theta$, which is assumed to follow a Markov process. There is an inverse market demand function, $P(Q, \theta)$, that is continuous and (weakly) decreasing in total output $Q$. Gross benefit is defined as,

$$B(Q, \theta) \equiv \int_0^Q P(z, \theta)dz. \quad (1)$$

The demand shock variable belongs to a finite (possibly very large) set of values: $\theta \in \{1, 2, ..., \bar{\theta}\} \equiv \Theta$. Transition probabilities are given by, $\text{Prob}[\theta_{t+1} = j \mid \theta_t = i] = \rho(i, j)$. The specification that demand shocks belong to a finite set and follow a Markov process simplifies the analysis of dynamic competition, while still allowing for realistic empirical application. $P(0, \theta)$ is assumed to have a finite upper bound for all $\theta \in \Theta$.

Wholesale electricity market demand is derived from downstream retail electricity demand. For some applications a (possibly large) fraction of retail customers make purchases at a fixed retail price. If some retail customers buy at a fixed retail price while others are subject to time-varying prices tied to the wholesale price (i.e., real time pricing) then the wholesale demand function will be a weighted average of retail demand at the fixed retail price and retail demand at a price linked to the wholesale price [see Borenstein and Holland, 2005]. In this case the function $B(Q, \theta)$ defined in (1) as the integral over the inverse demand is still a vital part of our analysis, but it should not be interpreted as a gross benefit function.
2.2 Generation

There are \( J \) different generation technologies. We assume that all generation units for any technology \( j \in \{1, ..., J\} \) are identical in terms of costs and capacity. We use the following notation:

\[
\begin{align*}
  k_j &= \text{total amount of type } j \text{ capacity} \\
  c_j &= \text{marginal cost of generation for type } j \\
  f_j &= \text{investment cost per unit of type } j \text{ capacity} \\
  s_j &= \text{start-up cost per unit of type } j \text{ capacity} \\
  m_j &= \text{minimum generation rate per unit of type } j \text{ capacity}; m_j \in (0, 1)
\end{align*}
\]

Marginal cost of generation is assumed to be constant for each technology type. Generation units that are turned on are restricted to operate between the minimum and maximum generation rate for their technology; type \( j \) generators must operate at a rate in the interval \([m_j, 1]\) per unit of capacity that is turned on. So there are four exogenous parameters - \( c_j, f_j, s_j \) and \( m_j \) - for each technology \( j \in J \). Total type-\( j \) capacity \( k_j \) is exogenous in Sections 2 and 3, but is made endogenous in Section 4.

This formulation captures two features of electricity generation technologies that are often abstracted away from in economic models of electricity markets: unit start-up costs and minimum generation rates.\(^3\) These features introduce non-convexities into production technology, which in turn complicates market analysis and may lead to non-existence of competitive equilibrium. In order to pursue our objective of a competitive market analysis that incorporates these technology features, we introduce a formulation in which non-convexities are permitted at the firm (or, generation unit) level, but for which the aggregate production technology is convex. We assume that individual firms are small relative to the size of the market; specifically, firms are assumed to be of measure zero in the formal model.

The formulation above is intended to provide a model of dispatchable technologies such as fossil fuel generation units and nuclear power plants. The model will be extended in Section ?? to allow for a random component in actual or potential generation. This kind of

\(^3\)On the other hand, we do abstract from some other technology features, such as ramping constraints that limit the rate at which generator output may be adjusted over time.
randomness is required in order to adequately model renewable energy generation, including hydroelectric generation.

2.3 Market, Feasibility, & Equilibrium

The market is comprised of a large number of small firms who operate as price takers. Each firm is identified with one unit of capacity of a particular type of generation. The mass of type-\(j\) technology firms is \(k_j\). The production technology for a firm can be described quite simply. At the start of period \(t\) a firm’s capacity is either ‘off’ or ‘on’; \(\omega_t = 0\) indicates ‘off’ and \(\omega_t = 1\) indicates ‘on’. If ‘on’ then the firm chooses a generation rate between the min and max rates for its generation type; a type-\(j\) firm’s output in \(t\) is \(\gamma_t \in [m_j \omega_t, \omega_t]\). In period \(t\), the firm also chooses its operating state next period denoted as \(\omega_{t+1}\).\(^4\)

\[
\begin{align*}
\omega_t &\in \{0, 1\} \\
\gamma_t &\in [m_j \omega_t, \omega_t]
\end{align*}
\] (2)

The following notation is used to describe the aggregate production technology. A vector \(x\) indicates the amount of each type of capacity that is ‘on’ at the beginning of the period; \(x_{jt}\) is equal to the mass of type-\(j\) firms that have \(\omega_t = 1\). The vector \(q\) is the amount of generation from the \(J\) types of generators, where, \(q_{jt} \in [m_j x_{jt}, x_{jt}]\) for \(j \in \{1, ..., J\}\). The choice for capacity that is on next period is given by \(x_{jt+1}\), where \(x_{jt+1} \in [0, k_j]\) for \(j \in \{1, ..., J\}\); \(x_{jt+1}\) is the mass of type-\(j\) firms that choose \(\omega_{t+1} = 1\). The aggregate production technology has two parts: one that specifies constraints within each period and a second that describes generator transitions across periods. For the first part, define the constraint set:

\[T(x_t, k) \equiv \{(q_t, x_{t+1}) : 0 \leq x_{jt} \leq k_j, m_j x_{jt} \leq q_{jt} \leq x_{jt}; j = 1, ..., J\}\] (3)

\(T(x, k)\) specifies how aggregate vectors of outputs and operating decisions for a period are constrained by \(x\), the vector of ‘on’ capacities at the start of the period and the exogenous

\(^4\)Note that an implication of feasible production technology is that if a firm shuts down its generation unit then this unit will not produce for at least two periods.
vector of capacities, \( k \).

An allocation is defined by a sequence for \((x_t, q_t, x_{t+1})\). In general, the values for this sequence of vectors will depend on realizations of the demand shock process, since as we will show, firms’ decisions will depend on realizations of this process. Because of this, it is useful to describe an allocation as a stochastic process. It will be convenient to denote a history of realizations of demand shocks through time period \( t \) as, \( \theta^t \), and the set of all possible histories through \( t \) as, \( \Theta^t \).

**Definition:** A feasible allocation from \( x_0 \) is a stochastic process \( \{x_t, q_t, x_{t+1}\}_{t=0}^{\infty} \) that

1. is measurable with respect to the set of possible histories of demand shocks,
2. satisfies \((q_t, x_{t+1}) \in T(x_t, k)\) for each realization of the process, and
3. satisfies \( (\text{??}) \) for each realization of the process.

The set of feasible allocations from any \( x_0 \) is convex, since set \( T(x, k) \) is convex for each \( x \) and transition equation \( (\text{??}) \) is linear. Note that the set of feasible allocations for the market is convex, even though the production possibilities set for an individual firm given by (2) is not convex. Measurability with respect to demand shock histories essentially means that there is an outcome for period \( t \) vector \((x_t, q_t, x_{t+1})\) corresponding to each possible demand shock history, \( \theta^t \).

In each period \( t \) firms are assumed to observe the current price \( p_t \) and the history of demand shocks through \( t \), \( \theta^t \). Firms are assumed to have rational expectations regarding future prices. Equilibrium prices for a given period depend on the history of demand shocks through that period. A price process \( \{p_t\} \) is a stochastic process that is measurable with respect to the set of all possible histories of demand shocks. Given a price process \( \{p_t\} \), the value of a type-\( j \) firm in period \( t \) is given by:

\[
v_{jt}(\omega_t, \theta^t) = \sup_{\{\gamma, \alpha, \beta\}} \mathbb{E}_t \left[ \sum_{\tau=0}^{\infty} \delta^\tau \left[ \omega_{t+\tau} (\alpha_{t+\tau} \gamma_{t+\tau} (p_{t+\tau} - c_j)) - (1 - \omega_{t+\tau}) s_j \beta_{t+\tau} \right] \right]
\]

The suprenum in (4) is with respect to stochastic processes for decisions \((\gamma_t, \alpha_t, \beta_t)\) that are
measurable with respect to the set of possible demand shock histories and for which each realization of the process satisfies (2). A policy for decisions regarding shut-downs, output, and start-ups is profit maximizing for a firm with own initial state \( \omega_t \) if it attains (4).

DEFINITION: An allocation \( \{x_t, q_t, y_t, z_t\}_{t=0}^{\infty} \) together with a price process \( \{p^*_t\} \) is a market equilibrium if:

(i) The allocation is feasible from \( x_0 \),

(ii) The allocation is consistent with profit maximizing policies for all firms, and

(iii) \( p^*_t = P(\sum_{j=1}^{J} q_{jt}, \theta_t) \) for all \( t \geq 0 \).

Condition (ii) states that all firms adopt policies that attain (4) and that when these policies are aggregated, the resulting allocation is the competitive equilibrium allocation. Condition (iii) is a standard market clearing condition.

### 3 Dynamic Market Equilibrium

In this section we develop the connection between market equilibrium and the solution to the planner’s problem. We show that an allocation is a market equilibrium if and only if it is a solution to the planner’s problem. We then show that a solution to the planner’s problem exists, and that the resulting allocation and price process constitute a market equilibrium. The equivalence of a competitive market equilibrium and a social optimum is, of course, a fairly standard type of result and parallels results for dynamic market equilibrium models in Lucas and Prescott [1971], Jovanovic [1982], and Hopenhayn [1990, 1992]. However, our formulation is somewhat different than those of these papers, and demonstrating this connection requires proof.\(^5\) The equivalence of market equilibrium allocations and solutions

\(^5\)The model in Lucas and Prescott [1971] has a single type of representative firm, in contrast to the model in this paper which has several types corresponding to different technologies and firm-specific ‘on/off’ states. The model in Jovanovic [1982] has heterogeneous firms, but no aggregate shocks. The models in Hopenhayn [1990, 1992] allow for aggregate shocks and heterogeneous firms via a distribution of firm-specific states. The model of electricity competition in this paper is similar to the class of models covered in Hopenhayn [1990] and the proof of Proposition 3.1 is similar to proofs in that paper. However, Hopenhayn [1990] imposes a continuity condition on transitions across firm-specific states; this condition does not hold in the model of the present paper.
to the planner’s problem is important because it provides a way to prove existence of market equilibrium and because it allows us to use the planner’s problem as a vehicle for computation.

Before turning to equilibrium results we describe the planner’s objective. The planner has access to a vector $k$ of total generation capacities for the $J$ technologies and makes operating decisions in each period after observing $(x, \theta) \in X(k) \times \Theta$; $(x, \theta)$ serves as a state vector for the planner. The operating decisions are embodied in vector, $(q, y, z)$, where $q$ specifies production rates, $y$ specifies how much of available generation in the period is left ‘on’, and $z$ specifies unit start-ups. The single period payoff, $H$, for the planner is total surplus for the period, which is equal to gross benefit less generation cost and start-up cost.

$$H(q, z, \theta) = B(\sum_j q_j, \theta) - \sum_j c_j q_j - \sum_j s_j z_j \tag{5}$$

$H$ is bounded and is concave and differentiable in $(q, z)$ for each $\theta \in \Theta$. Concavity follows from concavity of $B$ in total output and linearity of costs in outputs and start-up capacities. That $H$ is bounded follows from our assumption that $P(0, \theta)$ has a finite upper bound for all $\theta$ and from the capacity constraints that bound outputs for each technology.

**Proposition 3.1.** An allocation $\{a_t\} = \{q_t, y_t, z_t, x_t\}$ and price process $\{p_t\}$ constitute a market equilibrium from $(x_0, \theta_0)$ iff the allocation solves the planner’s problem of maximizing discounted expected total surplus from $(x_0, \theta_0)$.

Proofs are in the appendix. The *if* part of the proof of Proposition 3.1 is constructed by first showing that any welfare maximizing allocation, along with the associated market clearing price process, maximizes aggregate market profits of firms, taking the price process as exogenous. The second step is to show that there is an assignment of operating policies to individual firms such that aggregate market profit maximization implies maximization of individual firms’ profits. The *only if* part of the proof uses concavity of the planner’s single period return $H$ and convexity of the set of feasible allocations, to show that no alternative feasible allocation yields higher payoff to the planner than a market equilibrium allocation.

Let’s examine the planner’s problem more closely. The planner makes operating decisions
to maximize expected total surplus, where the single period return is \( H \) in equation (5). This can be described by a stationary stochastic dynamic programming problem with the following Bellman equation,

\[
W(x, k, \theta) = \max_{(q, g, z) \in T(x, k)} \{ H(q, z, \theta) + \delta E[W(y + z, k, \theta') | \theta'] \}
\]  

(6)

where \( \theta' \) is the next period demand shock. The capacity vector \( k \) is included in the value function \( W \) to facilitate the analysis of the long run model in Section 4.

**Proposition 3.2.** A market equilibrium exists from any \((x_0, \theta_0) \in X(k) \times \Theta\).

The proof of Proposition 3.2 proceeds by showing that a solution to the planner’s problem exists. An optimal policy for a planner’s solution generates a feasible allocation and a price process which, by Proposition 3.1 constitute a market equilibrium.

**Proposition 3.3.** There is a unique equilibrium price process from any \((x_0, \theta_0) \in X(k) \times \Theta\).

The proof is almost identical to that of Theorem 2 in Hopenhayn [1990], and therefore omitted.\(^6\) The idea of the proof is that if there were two distinct equilibrium price processes then there must be two distinct equilibrium allocations. Social welfare is equal at these two equilibrium allocations, since equilibrium allocations maximize the planner’s objective. But if equilibrium price processes are distinct then the marginal social value of output differs across the two allocations for some histories of demand shocks. This would imply that a convex combination of the two allocations, which is feasible by convexity of the aggregate technology, would yield strictly higher social welfare.

Note that the allocations that generate the unique equilibrium price process need not be unique. It could be that two different allocations with differing amounts of ‘on’ capacities of particular types support the same prices. Moreover, even for a specific allocation that supports the price process, there may exist alternative assignments of ‘on/off’ capacities for

\(^6\)The almost-everywhere qualifier used to state the result in Hopenhayn [1990] is unnecessary here since we use a finite, discrete set of aggregate shocks.
individual firms of a particular type that are consistent with the same total amount of ‘on’
capacity for that type.

Renewable electricity generation from sources such as wind turbines and solar PV panels
varies over time due to both cyclical and random variation in weather conditions. The
potential amount of electricity generation from a hydroelectric facility is also subject to
random variation in weather conditions. The specification of generation from subsection 2.2
must be extended to allow for a random component if we are to adequately model renewables.

Under some special conditions, the model developed above can be reinterpreted as a
model that includes renewable generation. For instance, suppose that renewable generation
follows a Markov process and that all potential renewable generation is always dispatched.
In addition, suppose that ordinary wholesale demand is subject to an additive demand shock
that follows a Markov process. Wholesale market clearing requires that quantity demanded
equals renewable generation plus total non-renewable generation, \( Q \). Under these conditions,
inverse demand is a function of \( Q \) and of a random variable equal to the demand shock minus
renewable generation.

4 Entry and Investment

The model is extended to allow generation capacities to be endogenous. Entry and capacity
investment are synonymous in this formulation; entry of type- \( j \) firms of measure \( \hat{k}_j \) implies
type- \( j \) capacity investment equal to \( \hat{k}_j \). Stage one of the model is an initial entry/investment
stage in which firms make one-time decisions about whether or not to enter. Entry decisions
collectively determine the vector of generation capacities available in the market. Stage two
encompasses an infinite horizon sequence of operating decisions. This second stage operates
exactly as the model of Sections 2 and 3, with a vector \( k \) of capacities inherited from stage
one. This extended model can be viewed as a model of long run competition in which firms
have an incentive to enter as long as the expected profit associated with entry is positive.\(^7\)

\(^7\)The models of dynamic competition in Jovanovic [1982] and Hopenhayn [1990,1992] also allow for entry
(and exit) of firms. There are firm-specific differences in efficiency across firms in these models and both
entry and exit occur over time in equilibrium. Thus, these models allow for the significant amounts of entry
At the beginning of stage one there is a positive measure of potential entrants for type-$j$ generation, for $j \in J$. Each potential entrant chooses whether to enter or not. Not entering yields a payoff of zero. Entry requires paying investment cost $f_j$ for a single unit of capacity, and yields a return equal to the value associated with participation in stage two. Stage one value $\tilde{v}_j$ for a type-$j$ entrant may be expressed as the maximum of its no-entry and entry payoff:

$$\tilde{v}_j = \max\{0, -f_j + \tilde{\delta}\mathbb{E}[v_{j0}(0, \theta^0)]\}.$$ (7)

The discount factor $\tilde{\delta} \in (0, 1)$ reflects a time lag between stage one, when investment cost is incurred, and stage two, when generation unit operations commence. We allow this discount factor to differ from the discount factor used in stage two, since the investment time lag is likely to much longer, say 1/2 year, than the period length in stage two, e.g., one hour. The expectation is taken over some distribution of initial demand shocks. The value $v_{j0}(0, \theta^0)$ on the RHS of (7) is the value beginning with the initial period in stage two, for a generation unit that is ‘off’; this value is defined in equation (4).

The specification of the aggregate production technology is modified to account for endogenous capacities. We assume an upper bound $\bar{k}_j > 0$ for total capacity for each technology $j$; $\bar{k}$ is the vector of these upper bounds. $\bar{k}_j$ serves as the measure of type-$j$ potential entrants.

We modify the definition of a feasible allocation as follows:

**DEFINITION:** A *feasible allocation* for the long run model is a stochastic process $\{x_t, q_t, y_t, z_t, k\}_{t=0}^{\infty}$ that

(i) satisfies $k \in K \equiv [0, \bar{k}_1] \times [0, \bar{k}_2] \times \ldots \times [0, \bar{k}_J]$ and $x_0 = 0$,

(ii) is measurable with respect to the set of possible histories of demand shocks,

(iii) satisfies $(q_t, y_t, z_t) \in T(x_t, k)$ for each realization of the process, and

(iv) satisfies (??) for each realization of the process.

... and exit that are observed over time in many industries. By contrast, wholesale electricity markets tend to have relatively stable generation technologies and very long-lived capacity investments, so we would expect less entry/exit ‘churn’ in these markets, compared to many other industries. For this reason we opt for a formulation with a one-time entry opportunity, and no exit option.
The set of feasible allocations is convex, since sets $K$ and $T(x,k)$ are convex and transition equation \((??)\) is linear.

The definition of market equilibrium is also modified to account for endogenous entry.

**DEFINITION:** An allocation \(\{x_t, q_t, y_t, z_t, k_t\}_{t=0}^{\infty}\) together with a price process \(\{p_t^*\}\) is a *long run market equilibrium* if:

\(i\) The allocation is feasible,

\(ii\) The allocation is consistent with profit maximizing policies for all firms, and

\(iii\) \(p_t^* = P(\sum_{j=1}^{J} q_{jt}, \theta_t)\) for all \(t \geq 0\).

A policy for decisions regarding entry, start-ups, output, and shut-downs is *profit maximizing* for a firm if it attains (7). Note that this extends the notion of profit maximization from Section 2 to include the entry decision.

Now the main results of this section may be stated.

**Proposition 4.1.** An allocation \(\{a_t\} = \{q_t, y_t, z_t, x_t, k_t\}\) and price process \(\{p_t\}\) constitute a long run market equilibrium iff the allocation solves the planner’s problem of maximizing discounted expected total surplus.

**Proposition 4.2.** A long run market equilibrium exists.

As in the proof of the corresponding Proposition 3.2, the proof is based on showing that a solution to the planner’s problem exists, and then using Proposition 4.1 to establish the result. The objective function for stage one of the planner’s problem is:

\[
\tilde{W}(k) = -\sum_j f_j k_j + \tilde{\delta}E[W(0,k,\theta_0)]
\]

The value function $W(x, k, \theta)$ from (6) is shown to be concave in $k$ for all $(x, \theta)$ in the proof of Proposition 4.2. This implies that the planner’s objective function $\tilde{W}$ is concave, and therefore continuous, in $k$. Since the planner selects $k$ from compact set $K$, there exists a vector of capacities that maximizes $\tilde{W}$. 

13
Corollary 4.3. If the total capacity constraint is not binding in long run equilibrium then firms earn zero expected profit.

The Corollary is a direct result of the definition of payoffs for a potential entrant in (7) and the definition of long run market equilibrium. If $k_j < \bar{k}_j$ in equilibrium then there is a positive measure of type-$j$ potential entrants who do not enter. If type-$j$ entrants earn positive profits, then non-entering type-$j$ firms are not maximizing profit since they could earn greater profit by entering.

5 Compuations and Applications

The model is applied to the Texas ERCOT market in order to illustrate characteristics of competitive equilibrium outcomes. These are preliminary numerical simulations of the model, and several simplifying assumptions are made in order to reduce computation time and complexity. There are just two generation technologies ($J = 2$); a baseload technology and a mid-merit technology. We assume that baseload capacity is always ‘on’, so that all start-up and shut down decisions pertain only to the mid-merit technology. This simplifies computation considerably since, conditional on the demand shock, the planner’s value function is a function of a single variable, rather than a vector. Furthermore, we assume that there is enough capacity to meet maximum demand at price equal to mid-merit marginal generation cost. The assumed parameters for generation technologies are in Table One.

<table>
<thead>
<tr>
<th>Param.</th>
<th>Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>$30$/MWh</td>
<td>baseload marginal cost</td>
</tr>
<tr>
<td>$m_1$</td>
<td>0.5</td>
<td>baseload minimum generation rate</td>
</tr>
<tr>
<td>$k_1$</td>
<td>32,000 MW</td>
<td>baseload capacity</td>
</tr>
<tr>
<td>$c_2$</td>
<td>$60$/MWh</td>
<td>mid-merit marginal cost</td>
</tr>
<tr>
<td>$m_2$</td>
<td>0.7</td>
<td>mid-merit minimum generation rate</td>
</tr>
<tr>
<td>$k_2$</td>
<td>30,000 MW</td>
<td>mid-merit capacity</td>
</tr>
<tr>
<td>$s_2$</td>
<td>$100$/MW</td>
<td>mid-merit start-up cost</td>
</tr>
</tbody>
</table>

The process for demand shocks is an important part of the model. Again, in order to
reduce computation time and complexity, simplifying assumptions are made. Hourly ERCOT load data for 3 summer months of 2008 are used to construct demand states and transition probabilities across states. The load data is put into bins of 1,000 MWh, ranging from 28,000 to 62,000 MWh. A demand state is an (hour-of-day, load) pair; there are 325 demand states. A transition probability matrix for demand states is estimated using the frequency distribution from the data. There is a linear inverse demand function for each state given by, \( P(Q, \theta) = \theta - bQ \), where \( \theta \) is the demand shock for that state and \( b \) is a common coefficient. An average price elasticity of 0.1 for wholesale demand is assumed. Using the average load and average wholesale price, this induces a value for \( b = 0.01 \). A value for \( \theta \) is determined for each demand state by assuming that load equals the quantity demanded at the average wholesale price.

Based on Propositions 3.1 and 3.2, we compute a solution for the planner’s problem, and the allocation and resulting price process is a competitive equilibrium. Given our assumptions, the amount of ‘on’ baseload capacity is fixed at total baseload capacity. That is, \( x_1 = k_1 \). Given that capacities \( k = (k_1, k_2) \) are fixed, the value function for the planner, \( W(x, k, \theta) \) is then a function of \( x_2 \in [0, k_2] \) for each \( \theta \in \Theta \). We use Chebyshev collocation to approximate the value function \( W \) as a function of the continuous variable \( x_2 \), for each possible \( \theta \) [see Judd [1998] and Miranda and Fackler [2002] for expositions of collocation methods and computation strategies]. A fairly coarse collocation grid of 5 nodes is used for these computations. This means that the planner’s value function is approximated by a fifth degree polynomial in \( x_2 \), for each possible \( \theta \) in the simulations.

We assume a 10% annual interest rate. This translates to an hourly discount factor of \( \delta = 0.999989 \). This is of course a very high discount factor. In these preliminary computations we use successive approximations to find the planner’s value function. Convergence is slow because the discount factor is high. In future work we will look for a faster computation method.

We simulate the model for a sequence of 500 time periods (hours), beginning from initial demand state #1 and mid-merit ‘on’ capacity equal to 1/2 total mid-merit capacity. Figure 1 displays a frequency distribution of equilibrium prices for the simulation; price is on the
horizontal axis and frequency is on the vertical axis. Roughly 1/2 of prices are either at baseload marginal cost ($30/MWh) or mid-merit marginal cost ($60/MWh). About 1/3 of equilibrium prices are between these two marginal costs. The remaining 17% are above mid-merit marginal cost, and about 4% are above $95/MWh. The emergence of equilibrium prices above, in some cases well above, maximum marginal generation cost is an important empirical implication of this type of model. This is in sharp contrast to the predictions of the (static) competitive benchmark model. For the parameters used here, competitive benchmark predicted prices are all in the interval $30 - $60/MWh. These results suggest two things. First, the dynamic model with start-up costs may be able to explain/rationalize at least some part of the price volatility observed in wholesale electricity markets. Second, it suggests that using price markups over marginal generation cost may be a misleading measure of market power, likely to overstate the extent of actual market power. While the simulations thus far a limited, future additions to the paper will build off these results to examine how the short-run dynamic constraints affect investment decisions under environmental policies such as the subsidization of renewables or pricing carbon.
Figure 1: Histogram of Equilibrium Prices
Figure 2: 10 Day Simulation of Equilibrium and Competitive Benchmark Prices (in $/MWh)

Figure 3: 10 Day Simulation of Equilibrium Startups (in MWs)
6 Conclusion

In this paper we have developed a computationally tractable model of dynamic competition in electricity markets which allows us to characterize equilibrium outcomes with fixed generating capacities. We show a correspondence between the dynamic competitive equilibrium and a social planners problem. This allows us to establish equilibrium existence and also provides an attractive computational platform for numerical analysis. We then extended the model to allow for endogenous generating capacities. This allows us to characterize long-run steady-state capacity choices in equilibrium.

Using the framework for numerical analysis, we find that incorporating dynamic constraints, such as startup costs and minimum operating levels, yields a significantly more volatile equilibrium price distribution. Further computational work will expand the set of technologies in the consideration set as well as characterize both short-run and long-run equilibria under potential environmental regulation.

Appendix

Proof of Proposition 3.1

(⇒) Let feasible allocation \( \{a_t\} = \{q_t, y_t, z_t, x_t\} \) be a social optimum from initial state \((x_0, \theta_0)\). This allocation is a stochastic process that is measurable on the set possible histories of demand shocks. This allocation induces a price process, \( \{p_t\} \), where \( p_t = P(\sum_j q_{jt, \theta_t}) \).

Let \( \{a'_t\} \equiv \{q'_t, y'_t, z'_t, x'_t\} \) be some alternative feasible allocation from \((x_0, \theta_0)\). Define a family of allocations by, \( \{a^\lambda_t\} = \{\lambda a_t + (1 - \lambda)a'_t\} \), for \( \lambda \in [0, 1] \). \( \{a^\lambda_t\} \) is a feasible allocation since the set of feasible allocations is convex.

Since \( \{a_t\} \) is a social optimum, we have,

\[
E_0[\sum_t \delta^t H(q^\lambda_t, z^\lambda_t, \theta_t)] - E_0[\sum_t \delta^t H(q_t, z_t, \theta_t)] \leq 0
\]
for $\lambda \in [0, 1]$. This implies that,

$$E_0[\sum_t \delta^t [H(q^\lambda, z^\lambda, \theta_t) - H(q_t, z_t, \theta_t)] \leq 0$$

for $\lambda \in (0, 1)$, and

$$\lim_{\lambda \downarrow 0} E_0[\sum_t \delta^t [H(q^\lambda, z^\lambda, \theta_t) - H(q_t, z_t, \theta_t)] \leq 0.$$

Gross benefit $B$ is bounded above, since $P(0, \theta)$ has a finite upper bound for all $\theta$ and quantities are limited by capacities. Therefore, $H$ is bounded above. Since $H$ is bounded above by a $\delta$-integrable function, we can pass the limit operator inside the expectation and $t$-summation operator. So,

$$E_0[\sum_t \delta^t \lim_{\lambda \downarrow 0} [H(q^\lambda, z^\lambda, \theta_t) - H(q_t, z_t, \theta_t)] \leq 0.$$

Taking the limit as $\lambda$ approaches zero yields,

$$E_0[\sum_t \delta^t \sum_j \left( \frac{\partial B}{\partial Q} \sum_j q_{jt} \theta_t \right) - c_j (q_{jt} - q_{jt}) - s_j (z_{jt} - z_{jt})] \leq 0,$$

and since the derivative of gross benefit is equal to price,

$$E_0[\sum_t \delta^t \sum_j [(p_t - c_j) q_{jt} - s_j z_{jt}]] \leq E_0[\sum_t \delta^t \sum_j [(p_t - c_j) q_{jt} - s_j z_{jt}]]. \quad (8)$$

The RHS of (8) is aggregate market profit at the allocation $\{a_t\}$ and price process $\{p_t\}$ induced by the solution to the planner’s problem. Since $\{a'_t\}$ is an arbitrary alternative feasible allocation, inequality (8) implies that, given $\{p_t\}$ and rational expectations on this process, there is no other feasible allocation that yields greater aggregate market profit than the allocation induced by the planner’s solution.

The last part of the proof is to connect aggregate market profit maximization to maximization of individual firms’ profits. Given the allocation $\{a_t\}$ we induce operating policies
for firms as follows. For period $t$ and history $\theta^t$ there is a vector $(q_t, y_t, z_t, x_t)$ associated with $\{a_t\}$. For type-$j$ firms, mass $x_{jt}$ have $\omega_t = 1$ and mass $k_j - x_{jt}$ have $\omega_t = 0$. If $x_{jt} > 0$ then we assign the fraction $y_{jt}/x_{jt}$ of ‘on’ firms $\alpha_t = 1$ and the remaining fraction of ‘on’ firms $\alpha_t = 0$. For type-$j$ firms with $\alpha_t = 1$, assign production $\gamma_t = q_{jt}/y_{jt}$. If $x_{jt} < k_j$ then we assign the fraction $z_{jt}/(k_j - x_{jt})$ of ‘off’ firms $\beta_t = 1$ and the remaining fraction of ‘off’ firms $\beta_t = 0$. Any such assignment of operating policies yields a stochastic process for $(\gamma_t, \alpha_t, \beta_t)$ that satisfies (2) for each possible demand shock history, for each firm. The assigned operating policy for a firm yields a (expected, discounted) profit for the firm, and aggregated profits of all firms from period zero are equal to aggregate market profit on the RHS of (8). This equivalence holds since the operating policy assignment used implies that the aggregated production quantities and start-up capacities of firms add up to the $q_{jt}$ quantities and $z_{jt}$ start-up capacities on the RHS of (8).

Suppose that there is an alternative policy for a firm that yields greater profit than the policy assigned based on allocation $\{a_t\}$. Furthermore, suppose that the measure of firms for which this is true is positive. Then an alternative market allocation can be constructed such that the firms that achieve greater profit with an alternative policy are assigned the alternative policy, and other firms retain the policy based on allocation $\{a_t\}$. This alternative market allocation is feasible, since there are no externalities in production across firms. The aggregate profit for this alternative market allocation exceeds aggregate profit associated with $\{a_t\}$ (the profit on the RHS of (8)). But this contradicts the result that aggregate profits are maximized using allocation $\{a_t\}$; that is, there cannot be a positive measure of firms that achieve greater profit than the profit associated with the assigned policy from allocation $\{a_t\}$. If there is an alternative policy for a firm that yields greater profit than the policy assigned based on allocation $\{a_t\}$, and the measure of firms for which this is true is zero, then the policy assignment for these firms can be changed to the alternative policy without altering the market allocation or the price process.

We have shown that social optimal allocation $\{a_t\}$, along with associated process $\{p_t\}$, satisfies the 3 conditions for a market equilibrium:

(i) The allocation is feasible from $x_0$, since a socially optimal allocation must be feasible,
(ii) There is an assignment of policies to firms that maximize profit for each individual firm,
(iii) The price process satisfies the market clearing condition, since prices are set to clear the
market for each period $t$ and history $\theta^t$.

$\Leftarrow$ Let $\{a_t\} = \{q_t, y_t, z_t, x_t\}$ and $\{p_t\}$ be the feasible allocation and price process, re-
spectively, for a market equilibrium. Suppose there is an alternative feasible allocation
$\{a'_t\} = \{q'_t, y'_t, z'_t, x'_t\}$ that yields greater total surplus than $\{a_t\}$. Define $\{a^\lambda_t\}$ as in the if
part of the proof for $\lambda \in [0, 1]$; we know that $\{a^\lambda_t\}$ is a feasible allocation. By the concav-
ity and differentiability properties of $H$ in $(q, z)$, using Lemmas 1 and 2 from Hopenhayn
[1990], we have the function $(H(q^\lambda_t, z^\lambda_t, \theta) - H(q_t, z_t, \theta))/\lambda$ is decreasing in $\lambda$, and increases
to $\sum_j [(p_t - c_j)(q'_jt - qjt) - s_j(z'_jt - zjt)]$ as $\lambda \downarrow 0$. In addition, concavity of $H$ implies that
$(H(q^\lambda_t, z^\lambda_t, \theta) - H(q_t, z_t, \theta))/\lambda$ is bounded below by $(H(q'_t, z'_t, \theta) - H(q_t, z_t, \theta))/\lambda$ for $\lambda \in (0, 1]$.
These results are used in the string of inequalies below.

$$E_0[\sum_t \delta^t H(q'_t, z'_t, \theta_t)] - E_0[\sum_t \delta^t H(q_t, z_t, \theta_t)] \geq 0$$

for $\lambda \in [0, 1]$ since $\{a'_t\}$ yields greater total surplus than $\{a_t\}$. This implies that,

$$E_0[\sum_t \delta^t \frac{H(q^\lambda_t, z^\lambda_t, \theta_t) - H(q_t, z_t, \theta_t)}{\lambda}] \geq 0$$

for $\lambda \in (0, 1]$, and

$$\lim_{\lambda \to 0} E_0[\sum_t \delta^t \frac{H(q^\lambda_t, z^\lambda_t, \theta_t) - H(q_t, z_t, \theta_t)}{\lambda}] \geq 0.$$ 

This implies that,

$$E_0[\sum_t \delta^t \sum_j [(p_t - c_j)q'_jt - s_jz'_jt]] \geq E_0[\sum_t \delta^t \sum_j [(p_t - c_j)qjt - s_jzjt]]. \quad (9)$$

In other words, we have shown that the supposition that there is some alternative feasible
allocation $\{a'_t\}$ that yields greater total surplus than a market equilibrium allocation $\{a_t\}$
implies that total market profits at \{a'_t\} exceed those at \{a_t\}. This is inconsistent with the property that individual firm profits are maximized at \{a_t\}, since (9) implies that at least some firms could earn greater profits by choosing a different policy.

**Proof of Proposition 3.2**

Recall that function \(H\) is bounded and is concave, and therefore continuous, in \((q, z)\) for each \(\theta\). In addition, the constraint set \(T(x)\) is compact-valued for each \(x\) and continuous in \(x\). Define an operator \(T\) as follows:

\[
(Tf)(x, \theta) = \max_{(q, y, z) \in T(x)} \{H(q, z, \theta) + \delta \sum_{\theta' = 1}^{\tilde{\theta}} \rho(\theta, \theta') f(y + z, \theta')\} \tag{10}
\]

Let \(C\) be the space of bounded functions that are continuous in \(x\) that map \(X(k) \times \Theta\) into \(\mathbb{R}\). Then by an argument similar to that in the proof of Theorem 4.6 in Stokey and Lucas [1989], the operator \(T\) maps elements in \(C\) into \(C\) and \(T\) has a unique fixed point, which is the value function \(W(x, k, \theta)\) that satisfies the Bellman equation (6) for the planner. An optimal policy associated with (6) yields a feasible allocation and a price process, which by Proposition 3.1 constitute a market equilibrium.
References


