To appear in *Optimization Methods & Software*
Vol. 00, No. 00, Month 20XX, 1–25

RESEARCH PAPER

Modeling Demand Response in Organized Wholesale Energy Markets

Michael C. Ferris[a] and Yanchao Liu[b]

[a]*Department of Computer Sciences, University of Wisconsin, Madison, WI, USA*; [b]*Department of Industrial and Systems Engineering, University of Wisconsin, Madison, WI, USA*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

We propose a bi-level optimization model for demand response in organized wholesale energy markets. In this model, the lower level performs the economic dispatch of energy and generates the price and the upper level minimizes the total amount of demand response subject to a net benefit requirement. In an economic sense, demand response is a trade of "consuming rights" instead of a sale of energy. Therefore it must be traded separately from the energy market. Although a bi-level optimization model is very hard to solve in general, we demonstrate that realistic power networks have characteristics that can be exploited to reduce the effective size of the problem instance. In particular, we transform the nonconvex net benefit test constraint to an equivalent linear form, and reformulate the nonconvex complementarity conditions of doubly bounded variables using SOS2 constraints. For realistic instances of the MPEC, we employ a three-phase approach that exploits the fast local solution from a nonlinear programming solver as well as LP-based bound strengthening within a mixed integer/SOS2 formulation. The model is tested against various data cases and settings, and generates useful insight for demand response dispatch operations in practice.

Keywords: Bi-level Program, Mixed Integer Program, Energy Market, Demand Response

1. Introduction

On March 15, 2011, the Federal Energy Regulatory Commission issued Order No. 745 [15], the "Final Rule" attempting to settle the yearlong rule-making debate on how to compensate demand response resources that participate in an organized wholesale energy market administered by a Regional Transmission Organization (RTO) or an Independent System Operator (ISO). This Order has proven very controversial. In May 2014, The U.S. Court of Appeals for the District of Columbia Circuit vacated the Order 745, agreeing with a group of electricity generators that FERC had overstepped its legal authority and was encroaching on the states' exclusive legal right to regulate retail electricity markets. However, on January 25, 2016, the Supreme Court majority disagreed with the Court of Appeals, ruling that demand response is primarily a wholesale market function and FERC Order 745 only addresses wholesale market transactions [4]. In this paper, we outline the Order, propose a model to implement its ruling and demonstrate computational approaches for numerical solutions.

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

This paper investigates a dynamic dispatch approach that incorporates the demand response dispatch and compensation rules as described in Order 745. The remainder

of this section introduces the background of the subject, in particular, the motivation for promoting demand response in the wholesale market and the key elements of the Order that pose challenges for implementation and thus motivate our work. Section 2 presents our main contribution: modeling the demand response problem as a bi-level optimization problem and proposing different reformulation and solution approaches. We transform the bi-level model into a linear program with complementarity constraints, which is further reformulated as a mixed integer program for global solutions. In Section 3, we develop a graphical approach to determine the DR cost-effectiveness condition in a congestion-free network, which is to be ued to validate the bi-level model in experiments. This simple method is also a straightforward way to estimate the monthly threshold to trigger demand response, as specified in the Order. Section 4 presents extensive numerical experiments covering model validation, general solvability, a bound strengthening method and useful observations on practical demand response operations. For realistic instances, we employ a collection of new and existing reformulations and tools that make an otherwise intractable model solvable in reasonable time. In particular, our three-phase approach can obtain global solutions for test cases with more than 2000 buses within 30 minutes. Section 5 concludes the paper and briefly discusses future work.

All occurrences of the term ISO in the rest of the paper should be taken as ISO/RTO. Electric power means the real (instead of the reactive or apparent) power. Node, bus and location mean the same thing. We use the units MW and $/MW to measure the power and the price. It is understood that the unit $/MW really means $/MW per hour, since we only consider a snapshot model representing an hour; and hence it is equivalent to $/MWh.

1.1 *Motivation of Demand Response*

The motive, if not the action, of consumers' response to electricity prices has existed since spot pricing was adopted in the electricity market. In the book [34], two essential types of consumers' response were identified: reduce usage if the price in a given hour is high, and reschedule usage if the price is high in some hours and low in other hours. Another early work [11] studied how a storage-type consumer could respond to the spot pricing of electricity by determining an optimal schedule of electricity usage given a predetermined electricity price schedule.

However, demand response is not an inherent element of competitive energy markets; rather, it is a recourse measure to account for the market imperfection caused by some unusual characteristics of the underlying commodity, electric energy. Specifically, electricity supply and demand over the grid must match closely at every instant in time, so the market must clear in real time. Such frequent market clearing cannot happen naturally (at the discretion of the "invisible hand"), but requires coordination of a central dispatcher, namely, an ISO. The ISO attempts to clear the market efficiently, i.e., maximize the social welfare, and therefore needs to know explicitly how much the suppliers and demanders value each increment of supply and demand (the supply and demand curves, as depicted in Figure 1). This information is conveyed to the ISO via supply offers and demand bids, see, for example [3] and [36].

While the supply curve is usually easy to estimate, it is difficult for the majority of the demand-side to identify the marginal value of electricity and hence bid a meaningful demand curve, see [25]. We observe that such difficulty is not unique to electricity, but is present in many other commodity markets, e.g., markets of farm produces and consumer products, etc. However, those markets do not require instantaneous clearing, therefore, consumers' response to price signals, an alternative expression of the demand curve, can have enough time to settle in and keep market equilibrium at the efficient point [28].

Figure 1.: Electricity supply and demand curves.

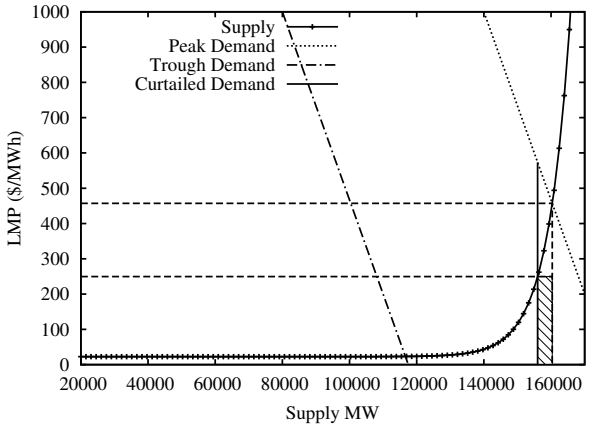
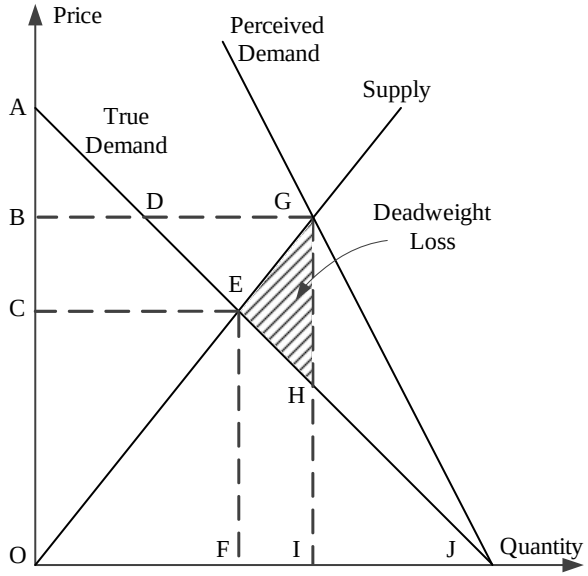


Figure 2.: Market inefficiency caused by imperfect demand information.



This is not the case for the ISO-run energy markets over the grid. In the absence of an accurate demand curve, social welfare cannot be accurately characterized, let alone being optimized.

It is commonly recognized that the demand elasticity, or the willingness and ability of the demand-side to reduce consumption in times of high prices, is actually higher than what is perceived by the ISO from the demand bids, see, for example, [24], [39] and [8]. The consequence of under-perceiving the demand elasticity is illustrated in Figure 2. With the true demand and supply curves, the market equilibrium is at point E with supplier's surplus being the area of COE and the demander's surplus being the area of ACE. However, if the elasticity of the demand was underestimated, as represented by the Perceived Demand curve in the figure, the market would operate at point G, resulting in a supplier's surplus of BOG and a demander's surplus of ABD minus DHG. The net effect is a surplus transfer from the demander to the supplier by the amount $BCEG$ and a social welfare loss of EHG.

A demand response mechanism, if appropriately designed and implemented, can serve to overcome the market inefficiency by migrating the market equilibrium point from G to E in Figure 2. This is illustrated in Figure 1, in which the supply curve is plotted

according to the Modeled Supply Function for June, 2012, published by PJM on its website. Note that the curve approximates the aggregated supply capability within the RTO and is not to be taken as the physical cost curve of any individual power plant. There is not a single demand curve because the demand for electricity is practically cyclical and the demand curve shifts from left to right depending on the time of day, day of the week or week of the year. For demonstration purposes, the two (dotted and dot-dashed) slanted lines are fictitious demand curves to represent the peak and trough demand scenarios, respectively. A similar depiction can be found in [25]. The curtailed demand line (the vertical line that stems downward from the middle of the peak demand curve) represents a fictitious example of demand response, corresponding to an amount measured by the distance from the vertical dashed line to this line. The demand response yields a reduction in locational marginal price (LMP) from around \$460/MWh to about \$250/MWh. The shaded area is the compensation received by the curtailed demand as per the FERC Order. It can be seen that the effect of demand response on the market equilibrium is aligned with what is needed to correct the inefficiency as depicted in Figure 2. As a side effect, demand response can also thwart the "all but irresistible temptation for generators to manipulate the market, sending prices soaring" as depicted by [35]. The following paragraph in FERC Order 745-A (on page 15) briefly summarizes the above point:

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

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

1.2 *Understanding the FERC Order*

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

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

Conventional wisdom would regard demand response as consumers' voluntary action to curtail consumption to cut down energy bills during periods of high prices. This is not the same notion of DR as discussed in the Order. The Order clearly refers to demand

[1]Interested readers are referred to [26] for a discussion of the DR policy from an alternative perspective other than a direct compensation viewpoint.

response as a service procured by, and a dispatchable resource of the ISO, which means that the dispatch decision, i.e., when, how much and which resources to dispatch, is under the jurisdiction of the ISO, not of the DR providers. A DR provider, on the other hand, needs to inform the ISO (via bids) of its capability and willingness to follow the ISO's dispatch.

For the second question, abundant evidence in the literature suggests that the answer is no. In short, a DR provider is not entitled to sell (in the energy market, day-ahead or real-time) the energy curtailed from its baseline consumption, without physically or contractually owning the baseline amount of energy, see [33], [8], [6] and [17]. Instead, DR can be treated as a sale of the "consuming rights" from certain consumers (DR provider) to other consumers (the remaining load). In particular, as implied by the Order, the remaining consumers pay the DR provider to reduce consumption. When the supply curve is steep, such trades among the demand-side can be beneficial to all consumers, including DR providers who get compensation from the remaining load, and the remaining load who enjoys lower LMP. This trading of consuming right is done outside the energy market so there is not an issue about energy entitlement. From a modeler's perspective, the simultaneous clearing of DR and energy requires either an iterative process, or a hierarchical model. This is the main subject of study in this paper, and will be elaborated in subsequent sections.

The answer to the third question has been indicated in the Order. Specifically, the Order recognizes and stresses the "billing unit effect", a phenomenon that, depending on the change in LMP relative to the size of the energy market, dispatching demand response resources may result in an increased cost per unit (dollars/MW) to the remaining wholesale load associated with the decreased amount of load paying the bill. See footnote 119 in the Order [15] for a numerical example. The Order states that billing unit effect should be avoided when an ISO dispatches the demand response.

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

$$\text{AvgPrice} = \frac{\sum_k (g_k + r_k)\lambda_k}{\sum_k (d_k - r_k)}, \tag{1}$$

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

The fourth question is critical for an economically sound DR program. As pointed out by [16], "if demand response is improperly compensated, hoped-for increases in efficiency may not materialize, as either too much or too little demand response may be developed." Better than nothing, the Order mentions a price level or threshold such that when the market price exceeds this level, the dispatch of demand response will be considered. Note that the "market price" here is meant to be a single price across all locations. We believe that this price is best defined as the demand weighted average LMP across all nodes

(short for AvgLMP), calculated by the following formula,

$$\text{AvgLMP} = \frac{\sum_k d_k \lambda_k}{\sum_k d_k}. \tag{2}$$

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

The ideal level of AvgLMP should be that of the market clearing point resulted from a perfect-information scenario, i.e., level C in Figure 2. The determination of such a point is of great importance to social welfare, and is not an easy task in practice. We refer the readers to [37] for a dedicated research and case study on this topic.

Curtailable load resources have existed for decades and there is an extensive literature on how system operators dispatch them, including, for example, [32], [7] and [1]. Moreover, many ISOs had been using demand response resources similar to that discussed in the Order before the Order was issued. Since February 2008, ERCOT has been operating an Emergency Interruptible Load Service (EILS) provided by loads (customers) willing to interrupt during an electric grid emergency in exchange for a payment [13]. The California ISO (CAISO) has been offering Demand Response products since 2010. CAISO's market participants can provide demand response in two ways, serving as a Proxy Demand Resource or serving as a Participating Load. A proxy demand resource can submit bids into the wholesale day-ahead and real-time markets and respond to dispatches at the direction of the CAISO, while a participating load is an entity providing Curtailable Demand, demand that can be curtailed at the direction of CAISO in the real-time dispatch [21, 22].

Since the issuance of the FERC Order, ISO/RTOs have made various localization efforts to preserve the economic efficiency of demand response. For example, ERCOT has implemented an "LMP minus proxy G" approach to avoid the double-payment problem, where G is a proxy for the purchase price or contract price that is generally representative of what retail customers would pay for their energy adjusted for risk [14]. PJM RTO treats the emergency demand response as a reliability resource where curtailment must strictly follow the RTO's dispatch orders. For voluntary load reduction, compensation is only assessed when the wholesale price is higher than a net benefit price published monthly by the RTO [18, 19]. The net benefit price represents the price at which the benefits incurred by a reduction in wholesale prices from the economic demand response will exceed the cost to pay for the economic demand response [18].

Some rules in the Order, such as the price at which DR resources are compensated, may be subject to further modification [20]. We would like to stress that this paper is not about the policy or arguing in favor of the Order. The main point is to contribute a solution methodology for dispatching DR resources within, but not dependent on, the existing operational context.

2. Modeling the Demand Response

The demand response problem arises from the ISO's practice of clearing the energy market, where economic dispatch is at the center of this practice. Research on this topic abounds in the power systems literature. In the development of this work, we find [30], [29], [2], [38] and [12] useful for understanding the subject matter. We briefly develop the economic dispatch model and then proceed to the demand response modeling.

Table 1.: Notations for the economic dispatch model.

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

2.1 *Preliminaries*

In the modeled power network, there is a set $\mathcal{B}$ of buses (or nodes), which are further distinguished by two subsets, i.e., $GEN \subset \mathcal{B}$ for generating buses and $LOAD \subset \mathcal{B}$ for load buses. A generating bus is one with an attached generating unit so that it may inject electricity into the network. A load bus is one that has no generating capability and can only withdraw electricity from the network. Buses are interconnected by transmission lines. In some cases there is more than one line connecting two buses, and each line is called a circuit[2]. Let CIR denote the set of circuit numbers, then every transmission line (or arc in graph theory terminology) in the network can be uniquely identified by the triple (k, l, c), where $k < l \in \mathcal{B}$, $c \in CIR$. Let $\mathcal{A}$ denote the set of all arcs in the network, and use the symbol a as a substitute for the arc triple (k, l, c) in subscripts when context allows. More notations are listed in Table 1, of which the upper half lists the parameters and the lower half lists the decision variables.

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

$$\min_{g,z,\delta} \quad \sum_{k\in\mathcal{B}} \alpha_k g_k^2 + \beta_k g_k \tag{3}$$

$$\text{s.t.} \quad z_{(k,l,c)} - b_{(k,l,c)}(\delta_l - \delta_k) = 0, \qquad \forall (k,l,c) \in \mathcal{A} \tag{4}$$

$$g_k - \sum_{\substack{(l,c):\\(k,l,c)\in\mathcal{A}}} z_{(k,l,c)} + \sum_{\substack{(l,c):\\(l,k,c)\in\mathcal{A}}} z_{(l,k,c)} = d_k, \qquad \forall k \in \mathcal{B} \tag{5}$$

$$\underline{g}_k \le g_k \le \bar{g}_k, \qquad \forall k \in \mathcal{B} \tag{6}$$

$$\underline{z}_a \le z_a \le \bar{z}_a, \qquad \forall a \in \mathcal{A} \tag{7}$$

In ED1, the objective function (3) is the total generation cost, with $\alpha_k \ge 0, \forall k$ ensuring the convexity of the function while β_k represents the intercept of the linear marginal cost curve, usually a non-negative quantity. Constraints (4) are the defining equations for the power flow z_a. These power flow quantities participate in the constraints (5), the nodal power balance equations. The equations in (5) say that at each bus k, the net generation $(g_k - d_k)$ must equal to the sum of the outbound power flow from bus k along all lines

[2]The unusual usage of the term "circuit" here is inherited from the IEEE Common Data Format, see `https://www.ee.washington.edu/research/pstca/formats/cdf.txt`. It should be noted that in a general power engineering context, a circuit refers to an entire path along which the electrons flow. The path may include several lines, along with other devices such as resisters and capacitors.

adjacent to k. Constraints (6) are the lower and upper bounds on the power generation, with $\bar{g}_k \geq \underline{g}_k \geq 0$. A load node $k \in LOAD$ that does not generate power is enforced by setting $\bar{g}_k = \underline{g}_k = 0$ in the data. Constraints (7) represent the thermal limits on the transmission lines, that is, the magnitude of the power flowing on an arc a should not exceed the arc's thermal limit $\bar{z}_a$ (whereas $\bar{z}_a = -\underline{z}_a$). Note that for a connected network which we assume here, the row rank of the linear system (4) to (5) was one less than full, which would leave an undesirable extra degree of freedom. For example, given g and z, we would be unable to determine δ. To overcome this issue, practitioners usually select a bus k at which the phase angle δ_k is artificially set to zero, and serve as the reference to the angles at all other buses. This bus is called the swing bus. In ED1 and all the subsequent models, the variable fixing is not expressed in the model but will be handled at the solution stage.

An important by-product of solving ED1 is the LMP. Take the bus k for example. The LMP at node k, denoted by λ_k, is by definition the sensitivity of the optimal value of the objective function to the demand d_k. Since ED1 is a convex quadratic programming model for which the KKT conditions are both necessary and sufficient for optimality, it is not difficult to verify that λ_k are the optimal multipliers on the nodal power balance constraints (5).

2.2 *Demand Response Model*

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

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

$$\frac{\sum_{k\in\mathcal{B}} d_k\lambda_k}{\sum_{k\in\mathcal{B}} d_k} \leq C_1 \tag{8}$$

Let C_2 be the AvgPrice before dispatching any DR resources. It is a parameter that can be calculated from the results of ED1, prior to computing the DR dispatch (i.e., applying (1) with $r_k = 0, \forall k \in \mathcal{B}$). Then the DR cost-effectiveness condition could be expressed as

$$\frac{\sum_{k\in\mathcal{B}}(g_k + r_k)\lambda_k}{\sum_{k\in\mathcal{B}}(d_k - r_k)} \leq C_2 \tag{9}$$

Within the boundaries of the net benefit test and LMP threshold constraints, there is leeway regarding the dispatch decisions of DR (i.e., which DR provider to dispatch and how much to dispatch), which can also have a substantial impact on economic efficiency. Such decisions will be guided by the objective function of the ISO's DR dispatch algorithm, for which the Order does not have a specification. Since the intended price-suppressing goal of DR is fully represented in the constraints, the objective, on the other hand, should aim to discourage over-suppressing of the price, or equivalently over-dispatching of the demand response, so as to prevent uneconomic consequences (as an example, in Figure 2 if the price is suppressed to a level below C, the deadweight loss

will emerge again). A myriad of functions can capture the "extent of DR dispatch", and the choice is up to the individual ISO. At present, we find no strong reason for the objective function to go beyond a linear form, so we will minimize a linear function $L(r)$ as the objective of the demand response dispatch. In subsequent analysis, we take $L(r) = \sum_{k\in\mathcal{B}} r_k$ to minimize the total amount of DR dispatch. Note that in cases where DR providers are allowed to bid a valuation, e.g., v_k, in addition to the upper bound $\bar{r}_k$, then $L(r) = \sum_{k\in\mathcal{B}} v_k r_k$ can be an appropriate objective function[3]. This and other variants of the model will be demonstrated in Section 4.6.

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

$$Q = \begin{bmatrix} 2\alpha_1 & & \\ & \ddots & \\ & & 2\alpha_{|\mathcal{B}|} \end{bmatrix} \quad c = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{|\mathcal{B}|} \end{bmatrix} \quad e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad B = \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_{|\mathcal{A}|} \end{bmatrix} \quad A = \begin{bmatrix} 1 & -1 & & \\ & & \ddots & \ddots \\ -1 & & & 1 \end{bmatrix}$$

The demand response model, DR1, is as follows.

$$\min_{g,z,\delta,r,\lambda} \quad L(r) \tag{10}$$

$$\text{s.t.} \quad d^T\lambda \le C_1 e^T d \tag{11}$$

$$(g + r)^T\lambda + C_2 e^T r \le C_2 e^T d \tag{12}$$

$$0 \le r \le \bar{r} \tag{13}$$

$$(g, z, \delta) \in \arg\min_{g',z',\delta'} \quad \frac{1}{2}(g')^T Q g' + c^T g' \tag{14}$$

$$\text{s.t.} \quad z' - BA\delta' = 0 \qquad (\perp \lambda^z) \tag{15}$$

$$g' - A^T z' = d - r \qquad (\perp \lambda) \tag{16}$$

$$g' \in [\underline{g}, \bar{g}] \qquad (\perp \eta^{\text{lo}}, \eta^{\text{up}} \ge 0) \tag{17}$$

$$z' \in [\underline{z}, \bar{z}] \qquad (\perp \mu^{\text{lo}}, \mu^{\text{up}} \ge 0) \tag{18}$$

where λ is the multiplier of (16).

DR1 is a bi-level model. Readers could consult [5] for a thorough treatment of bi-level optimization models, whereas [10] provides a useful survey on this subject. The lower level, i.e., (14) to (18), is an economic dispatch model (ED1) that takes the demand response variable r as given. Note that in an auction-based market context, dispatchable demand bids are abstractly represented as negative generation offers in the lower-level model. As discussed in Section 2.1, the LMP is the optimal multiplier λ on the nodal power balance constraint (16). The upper level minimizes the total MW amount

[3]Price bids from DR resources serve to better estimate the value of lost load (VOLL), which has been arbitrarily set to \$1000/MWh in most ISOs.

of demand response subject to the LMP threshold constraint (11), the cost-effectiveness constraint (12), DR bound constraints (13) and that (g, z, δ) solves the lower level problem so that λ represents the true LMP. Note that even if the objective function (10) is instantiated with DR valuations v_k, DR providers are not equivalent to price-sensitive demand bidders. In this case, DR providers present themselves as inelastic loads in the energy market and claim their "rights to be served" on equal footing as other inelastic loads, while in the meantime offering to trade such rights at a price should load shedding become necessary as determined by the central dispatcher. In contrast, a price-sensitive demand bidder directly bids in the energy market with price-quantity bids, in the same way as generators do. This includes any virtual bidders not backed by physical resources. In an ideal world, all loads should express price sensitivity. In the current market reality, however, few loads actively participate in the energy market. In this circumstance, an extra layer of coordination via the upper-level DR market is used to mitigate inefficiency.

An alternative model would be a single-level model formed by preserving all the constraints in DR1 and combining the two objectives in DR1 into one by summing them up. Let us name such a model DR1a:

$$
\begin{aligned}
\min_{g,z,\delta,r,\lambda} \quad & 1/2 g^T Q g + c^T g + L(r) \\
\text{s.t.} \quad & (11) - (13) \\
& z - BA\delta = 0 \\
& g - A^T z = d - r \quad (\perp \lambda) \\
& g \in [\underline{g}, \bar{g}] \\
& z \in [\underline{z}, \bar{z}]
\end{aligned}
$$

By comparing the two models, we argue that the bi-level model is more appropriate for the problem at hand. First, in DR1a, it is not justifiable to simply take the multiplier λ of the constraint (16) as the LMP. By definition, LMP is the cost of serving the next increment of demand, i.e., the derivative of the Lagrangian function with respect to the demand evaluated at a KKT point. The extra constraints (11) and (12) would complicate the expression of the derivative and disrupt this definition. In contrast, DR1 encapsulates the original ED model in its lower level and therefore the multiplier λ remains to represent the true LMP. Secondly, although the generation cost and the DR objective function $L(r)$ both need to be minimized, they are not simply additive in a single objective function. In fact, minimization of the two objective functions is intrinsically hierarchical in that the core business remains to be the economic dispatch given the demand data as well as a particular DR decision, and on top of that, we seek a "minimal" dispatch of DR to satisfy the LMP threshold constraint and the net benefit test. The bi-level DR1 exactly serves this purpose.

The economic dispatch model assumes that the nodal forecast loads, i.e., d_k, are fixed within the applicable time frame of the decision, be it an hour as in the day-ahead dispatch problem or five minutes as in the real-time dispatch problem. In practice, especially when time-coupling unit commitment decisions are involved, a planning horizon may consist of multiple ED time frames, e.g., the day-ahead plan consists of 24 hourly dispatch decisions. The inherently iterative process of optimizing DR resides in each ED time frame. The formulation (10) - (18) is independent of the planning horizon and therefore applicable for both the day-ahead and real-time markets.

2.3 *Model Reformulation*

The parameterized economic dispatch model in the lower level is a convex quadratic program, hence can be replaced by its KKT conditions, and therefore DR1 becomes an MPEC (mathematical program with equilibrium constraints) model. Specifically, the KKT conditions of the lower level problem include (15) to (18), as well as the following equalities and inequalities,

$$A\lambda - \lambda^{\mathrm{z}} - \mu^{\mathrm{lo}} + \mu^{\mathrm{up}} = 0 \tag{19}$$

$$Qg + c - \lambda - \eta^{\mathrm{lo}} + \eta^{\mathrm{up}} = 0 \tag{20}$$

$$(BA)^T \lambda^{\mathrm{z}} = 0 \tag{21}$$

$$\eta_k^{\mathrm{lo}}(g_k - \underline{g}_k) = 0,\ \eta_k^{\mathrm{lo}} \geq 0,\ \forall k \in \mathcal{B} \tag{22}$$

$$\eta_k^{\mathrm{up}}(\bar{g}_k - g_k) = 0,\ \eta_k^{\mathrm{up}} \geq 0,\ \forall k \in \mathcal{B} \tag{23}$$

$$\mu_a^{\mathrm{lo}}(z_a - \underline{z}_a) = 0,\ \mu_a^{\mathrm{lo}} \geq 0,\ \forall a \in \mathcal{A} \tag{24}$$

$$\mu_a^{\mathrm{up}}(\bar{z}_a - z_a) = 0,\ \mu_a^{\mathrm{up}} \geq 0,\ \forall a \in \mathcal{A} \tag{25}$$

where λ's and η's are dual variables, and their correspondence to the primal constraints (15) to (18) is marked in the parentheses following the constraints in DR1.

Two difficulties remain for the global solution of DR1: the nonconvexity of the net benefit test constraint (12), and the nonconvexity of the complementarity conditions in (22) to (25). We will address them below.

2.3.1 Transforming constraint (12)

The bilinear term $(g+r)^T\lambda$ in the net benefit test constraint (12) can be converted into a linear expression of the dual variables, as follows.

$$\begin{aligned}
(g+r)^T\lambda &= (A^Tz + d)^T\lambda && \text{by (16)}\\
&= z^TA\lambda + d^T\lambda\\
&= z^T(\lambda^{\mathrm{z}} + \mu^{\mathrm{lo}} - \mu^{\mathrm{up}}) + d^T\lambda && \text{by (19)}\\
&= \delta^T(BA)^T\lambda^{\mathrm{z}} + z^T\mu^{\mathrm{lo}} - z^T\mu^{\mathrm{up}} + d^T\lambda && \text{by (15)}\\
&= 0 + \underline{z}^T\mu^{\mathrm{lo}} - \bar{z}^T\mu^{\mathrm{up}} + d^T\lambda && \text{by (21),(24)-(25)}
\end{aligned}$$

Therefore, constraint (12) is reduced to a linear inequality:

$$\underline{z}^T\mu^{\mathrm{lo}} - \bar{z}^T\mu^{\mathrm{up}} + d^T\lambda + C_2e^Tr \leq C_2e^Td \tag{26}$$

2.3.2 Implementing constraints (22)-(25)

We investigate three approaches to implement the bilinear equations in (22)-(25). The first approach is taking the bilinear equations "as-is" to form a nonlinear program (NLP), then using an NLP solver to obtain a local solution. The second approach involves linearizing them using binary variables. For instance, the relation

$$\eta_k^{\mathrm{lo}}(g_k - \underline{g}_k) = 0 \tag{27}$$

Table 2.: Modeling complementarity using an SOS2 set.

	η^{lo}	s^{up}	s^{lo}	η^{up}
Case 1	+	+	0	0
Case 2	0	+	+	0
Case 3	0	0	+	+
Case 4	+	0	0	0
Case 5	0	+	0	0
Case 6	0	0	+	0
Case 7	0	0	0	+
Case 8	0	0	0	0

is equivalent to

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

where $\bar{\eta}_k^{\text{lo}}$ is the upper bound on η_k^{lo} and v_k^{lo} is a binary variable. The third approach takes advantage of the special ordered sets (SOS) capability of MIP solvers such as CPLEX and Gurobi. Special ordered sets are a device used in branch and bound methods for more intelligent branching on variables. Even when all members of an SOS are continuous variables, the model containing one or more such sets is a discrete optimization problem that requires a mixed integer solver for its solution. We do not address the branching technicalities in this paper; instead, SOS is primarily used as a means of specifying the discrete nature of complementary slackness without providing an explicit big-M bound. In particular, for each generator k let us define two positive variables $s_k^{\text{lo}} := g_k - \underline{g}_k$ and $s_k^{\text{up}} := \bar{g}_k - g_k$ and put the ordered quadruple $\{\eta_k^{\text{lo}}, s_k^{\text{up}}, s_k^{\text{lo}}, \eta_k^{\text{up}}\}$ in an SOS2 set, indicating that at most two members of the set can be positive and the positive members must be adjacent. The SOS2 set specified above has eight realizations regarding its members' positivity, as enumerated in Table 2. It is straightforward to see that all of these combinations are possible under complementary slackness and any other combination is impossible. Specifically, η^{lo} and s^{lo} cannot be both positive and likewise, η^{up} and s^{up} cannot be both positive. Note that cases 4 to 8 are when strict complementarity does not hold at both bounds of g_k.

To obtain global solutions, the first two approaches require explicit upper bounds for the multipliers η^{lo}, η^{up}, μ^{lo} and μ^{up}. The boundedness of these multipliers is not guaranteed since the Mangasarian - Fromovitz constraint qualification (MFCQ) [27, 31] does not necessarily hold at every optimal solution of the lower-level QP. Take a fictitious one-bus system for example: assume that a generator with \$1/MWh marginal cost and 20 MW capacity serves a load of D, so the dispatch solution becomes $\arg\min_g\{1/2g^2|g = D, 0 \leq g \leq 20\}$. When $D = 20$, the optimal solution g^* is 20, at which point the MFCQ does not hold. Such nonsmooth situations are conceivably rare in practice. However, even if the multipliers are assumed bounded, the bounds are unknown and have to be set heuristically in computation. The following bounds have been tested and shown to be effective in experiments. First, it is observed (and makes practical sense) that the dispatch of DR at any node would not result in an increase in the highest nodal LMP, and that the LMP usually does not drop below the marginal cost of the cheapest generator. Therefore, we set the upper bound for λ in the DR model as $\bar{\lambda} := ||\lambda^*||_\infty e$, the highest nodal LMP resulted from the ED1 model, and set the lower bound by $\underline{\lambda} := ||Q\underline{g} + c||_\infty e$, where e is a vector of 1's. From the bounds on λ, and by (20) and the fact that η^{lo} and η^{up} cannot

be positive simultaneously, we can set

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

For the bound on μ^{lo} and μ^{up}, we use $\bar{\mu}^{\text{lo}} := \bar{\mu}^{\text{up}} := 2||\mu^*||_\infty e$, where μ^* is the optimal multiplier (a vector of size $2|\mathcal{A}|$) on constraint (7) in the ED1 solution, and e is a $|\mathcal{A}|$-sized vector of 1's.

3. A Graphical Approach for Congestion-Free Networks

In this section, we propose a graphical approach for implementing the DR rules in a simplified setting which is often the case in practice. The approach is also used for validating the bi-level model in Section 4.1. The material is not essential to understanding the bi-level method hence its details can be reviewed at a later time.

It is well-known that when there is no congestion and no losses in the network, the LMPs at all buses will be identical and equal to the marginal supply offer at the market clearing point. In particular, absent constraints (7), ED1 could be simplified to ED2, as follows.

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

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

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

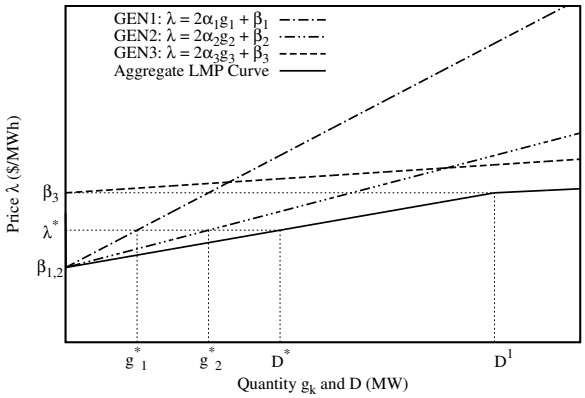
In Figure 3a, we draw the marginal cost lines of three generators, and the aggregate supply curve, which we call the LMP curve. In this example, β_1 and β_2 are equal, so we mark them by $\beta_{1,2}$ in the figure. This figure reveals the relationship among the demand, generator dispatch and price. For example, given a total demand D^*, we can read off from the figure the corresponding LMP, which is λ^*, and the generator dispatch solution, which is $(g_1^*, g_2^*, 0)$. It can be seen that the third generator will kick in when the demand is beyond the level of D^1, which marks a kink point on the piecewise linear LMP curve.

For the subsequent analysis of demand response on an LMP curve, we use superscript "old" and "new" on a symbol to indicate that it is a quantity before and after the demand response, respectively. Figure 3b shows two points on an LMP curve following this superscripting convention.

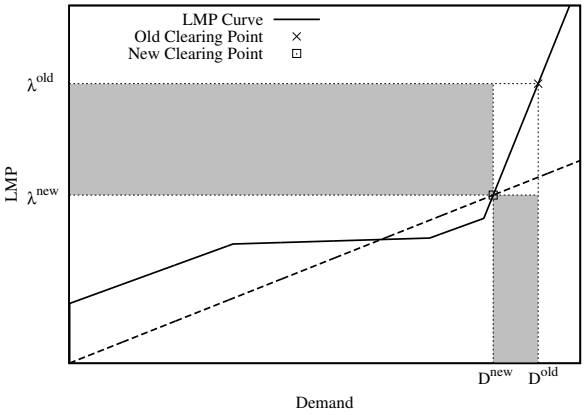
Let p denote the average price, which is defined by Equation (1). The demand reduction of $\Delta D = D^{\text{old}} - D^{\text{new}}$ results in a reduction in LMP by $\Delta\lambda = \lambda^{\text{old}} - \lambda^{\text{new}}$, and the average price after the demand response is $p^{\text{new}} = \lambda^{\text{new}} D^{\text{old}}/D^{\text{new}}$, while the average price before the demand response is $p^{\text{old}} = \lambda^{\text{old}}$. The cost-effectiveness condition requires $p^{\text{new}} \le p^{\text{old}}$,

Figure 3.: A graphical approach to identify cost-effective demand response solutions.

(a) LMP curve in a 3-generator example.



(b) Cost-effectiveness test on an LMP curve.



which gives $\lambda^{\text{new}} D^{\text{old}} \leq \lambda^{\text{old}} D^{\text{new}}$. Seen from Figure 3b, this inequality is equivalent to

$$\lambda^{\text{new}} \Delta D \leq D^{\text{new}} \Delta\lambda, \tag{28}$$

the left and right hand sides of which are the lower-right and upper-left shaded areas in Figure 3b. Since all the quantities involved in (28) are positive, we can equivalently write the inequality in the form

$$\frac{\lambda^{\text{new}}}{D^{\text{new}}} \leq \frac{\Delta\lambda}{\Delta D}. \tag{29}$$

In (29), the left hand side is the slope of the line passing through the origin and the point $(D^{\text{new}}, \lambda^{\text{new}})$, i.e., the dashed line in Figure 3b, and the right hand side is the slope of the line segment connecting $(D^{\text{old}}, \lambda^{\text{old}})$ and $(D^{\text{new}}, \lambda^{\text{new}})$. We summarize this observation in the following rule.

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

Applying Rule 1, we can see that Figure 3b shows a case where it is cost effective to dispatch the ΔD amount of demand response from the current demand level of D^{old}.

It is also easy to derive the "local" cost-effectiveness condition for the demand response,

in other words, whether the demand response is immediately cost effective as the DR amount ΔD increases from zero. As D^{new} approaches D^{old} from below, the left hand side of (29) becomes

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

which is the slope of the line passing through the origin and $(D^{\text{old}}, \lambda^{\text{old}})$, while the right hand side of (29) becomes

$$\lim_{D^{\text{new}}\uparrow D^{\text{old}}} \frac{\Delta\lambda}{\Delta D} = \lim_{D^{\text{new}}\uparrow D^{\text{old}}} \frac{\lambda^{\text{old}} - \lambda^{\text{new}}}{D^{\text{old}} - D^{\text{new}}} = \lim_{D^{\text{new}}\uparrow D^{\text{old}}} \frac{\lambda(D^{\text{old}}) - \lambda(D^{\text{new}})}{D^{\text{old}} - D^{\text{new}}} = \partial_{-}\lambda(D^{\text{old}}), \tag{30}$$

which is the left derivative of $\lambda(D)$ at D^{old}. If we make an convention that, at the demand level D^* which corresponds to a range of indefinite λ, let $\lambda(D^*)$ be the minimum value in the range, then $\lambda(D)$ becomes a function (1-to-1 mapping) of D, and (30) is just the slope of the LMP curve to the immediate left of the point $(D^{\text{old}}, \lambda^{\text{old}})$. Therefore, determining whether the demand response is locally cost effective at demand level D amounts to comparing the slopes of two lines that cut through the (D, λ) point in the LMP curve. This is noted in Rule 2.

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

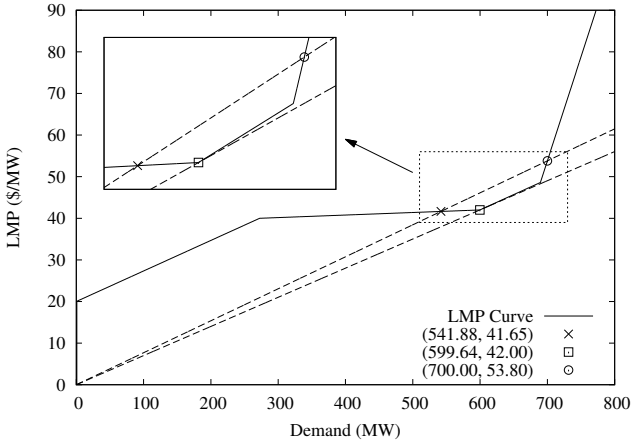
As outlined in the Order, there are two options or steps for an ISO to implement the DR policy: (1) develop an approximation to determine, and publish on a monthly basis, a price level at which the dispatch of demand response resources will be cost-effective; (2) develop a dynamic approach to include demand response dispatch in both the day-ahead and real-time energy markets. The graphical approach is useful for implementing Option 1 since network and other engineering constraints are commonly ignored for this level of economic analysis, as corroborated by the PJM example in Figure 1 as well as the CAISO's approach in [23, 40]. However, the graphical approach is not suitable for Option 2 where network constraints bring the "locational" dimension into play. In the general case, ED1 cannot be reduced to ED2 and the graphical approach would require a complete survey of all faces of the polytope prescribed by (4) - (7), the complexity of which would outweigh the purpose it intends to serve. The bi-level model has no such limitations and is therefore more suitable in these contexts.

4. Numerical Experiments

4.1 *Validation*

We validate the model DR1 by comparing its solutions to those obtained by the graphical method developed in Section 3. We will use the well-known 14-bus case [9] with the generator cost parameters coming from the Matpower [41] data sets.

Figure 4.: LMP curve for the 14-bus case without line limits.



The congestion-free LMP curve for the 14-bus case is given by

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

At $D = 599.642$, $\lambda = 0.073421 \times 599.642 - 2.02661 = 42.00$. Since the slope $42.00 \div 599.642 = 0.070042$ falls in between 0.006112 and 0.073421, we can identify the demand level 599.642 as a threshold for the cost-effectiveness of DR. In other words, DR is locally cost-effective only when the demand level is higher than 599.642 MW, see Figure 4. With this knowledge, we can fabricate some scenarios on which the results are predictable and test if the solutions found by DR1 on these scenarios match our predictions.

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

4.1.1 Scenario 1:

Set the demand to 599.642 MW, the economic dispatch model ED1 gives the current AvgLMP λ of 42.00 and the current AvgPrice C_2 of 42.00. We know that at this demand level, any positive DR level would violate the cost-effectiveness condition. So, if the LMP threshold C_1 is set to 42.00, we would expect DR1 to be just feasible thus optimal at $R_k = 0, \forall k$; and for a slightly lower LMP threshold, i.e., $C_1 = 41.99$, DR1 would become infeasible. Furthermore, similar results are expected to occur for any demand level that is lower than 599.624, which is also confirmed by experiments on demand levels sampled within the range [0,599.642], as demonstrated in Table 3.

Table 3.: DR1 results for cost-ineffective demand levels.

D	LMP (λ)	Price (C_2)	LMP Cap (C_1)	R_k
599.642	42.00	42.00	42.00	0
			41.99	infes
500	41.392	41.392	41.392	0
			41.391	infes
400	40.780	40.780	40.480	0
			40.779	infes
300	40.169	40.169	40.169	0
			40.168	infes
200	34.685	34.685	34.685	0
			34.684	infes

Table 4.: DR1 results for cost-effective demand levels with different C_1 values.

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

4.1.2 Scenario 2:

Set the demand to a level above 599.642 MW, for example, 600 MW, then ED1 gives $\lambda = C_2 = 42.03$. If we set $C_1 = 42.00$, an AvgLMP level corresponding to 599.642 MW demand, we would expect $600 - 599.642 = 0.358$ MW of demand response to be dispatched. The DR1 result confirmed this expectation, dispatching exactly this amount of DR at bus 2. We carry out a series of experiments on demand levels above 599.642 coupled with various C_1 levels. The results are summarized in Table 4. All results produced by DR1 match those generated by the graphical approach.

The $D = 700$ case is also illustrated in Figure 4. The line connecting the origin and the point (700, 53.8) intersects the LMP curve at (541.88, 41.65), so the maximum amount of cost-effective demand reduction from 700 MW is 700 - 541.88 = 158.12 MW, and the corresponding LMP is 41.65. These assertions are verified by DR1.

Another important point could be observed. As shown above, the demand can be cost-effectively reduced from 700 to 541.88. However, we have noted in Scenario 1 (and also by examining Figure 4) that any demand reduction from a demand level below 599.64 would have been cost-ineffective. The rationale for these clashing observations lies in the fact that the cost-effectiveness judgement depends on the current (starting) demand level. For example, at $D = 700$ the AvgPrice is 53.80, so a demand reduction that could yield

Table 5.: Setting and solution of IEEE test cases.

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

an average price no higher than 53.80 would be deemed cost-effective; but at $D = 599.64$, a demand reduction would have to yield an average price less than or equal to 42.00 in order to be cost-effective.

4.2 *General solvability*

We experiment different formulations and solvers on five IEEE test cases [9] to demonstrate the general solvability of the model. In particular, we run the NLP formulation using CONOPT, BARON and GLOMIQO, and run the binary and SOS formulation using CPLEX and GUROBI. Two congestion conditions are examined for each test case: free and congested. In order to make feasible yet simple DR cases, we need to scale up the demand to certain levels, and set appropriate line limits for the congested scenarios. The setting and solutions are presented in Table 5, in which λ^{ED} and λ^{DR} stand for the AvgLMP computed from the ED and DR solutions, and all C_1 levels were set to $0.9\lambda^{\text{ED}}$ for simplicity.

Table 6 lists the computation time (in seconds) for each solver to find the solution, where "-" indicates not finishing within an hour. The computer is a Dell R710 server with two 3.46G X5690 Xeon Chips, 12 Cores and 288GB Memory. For BARON, GLOMIQO and the binary formulation, we apply the big-M bounds discussed in Section 2.3.2 to pursue the global solution within the bounds. Note that the SOS formulation does not require artificial variable bounds, thus can be trusted to provide the true global solution. The fact that all solvers obtained the same solution is evidence for the validity of our choice of variable bounds. In all cases, CONOPT consistently provides a good local solution very quickly, which can serve as a starting point for other global solvers. We use this as part of a three-phase solution strategy to be discussed below.

4.3 *Solving Realistic Instances: A Three-phase Approach*

We proceed to test DR1 on larger cases based on the Polish network. While the nodal demands are scaled up (by a factor between 1.05 to 1.2) to make feasible DR cases, we adopted the realistic line ratings given in the network data. It is observed that in realistic cases most lines will never reach their thermal limits. We exploit this observation in a three-phase solution approach as outlined below.

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

Table 6.: Solution time (in seconds) of different formulations and solvers.

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

2. Bound and fix: For each line $a \in \mathcal{A}$, we find the lowest/highest level that the flow z_a can possibly reach (let us call such a level an effective bound), by minimizing/maximizing z_a subject to (15), (16), (17), (18) and the inequality $\sum_{k\in\mathcal{B}} r_k \le R^*$. If the effective lower bound of z_a is greater than $\underline{z}_a$, then μ_a^{lo} (which belongs to a SOS2 set) can be fixed to zero in the DR1 model; likewise, if the effective upper bound of z_a is less than $\bar{z}_a$, μ_a^{up} can be fixed to zero in the DR1 model. Such a "bound and fix" step could significantly reduce the effective number of discrete variables in the MIP (binary or SOS) formulation of DR1, making it easier to solve. Note that exploring the effective bounds requires solving $2|\mathcal{A}|$ linear programs, which is computationally inexpensive and is efficiently parallelizable (we used 40 parallel processes in the experiments).

3. Solving the MIP: After the variable fixing, we now solve the MIP (using either the binary or SOS2 formulation of Section 2.3.2) with CPLEX, taking the local solution from phase 1 as an initial integer feasible solution (i.e., enabling the *mipstart* option in CPLEX).

The performance of this approach is demonstrated in Table 7. The solution times of each step are listed in the last three columns of the table. We can see that for each of the five cases, CONOPT obtains the local solution within about 10 seconds and the bound strengthening time is well within 2.5 minutes. Both of the binary and SOS2 formulations obtained the same solution but the binary formulation solves much faster. Here are the settings used in the experiments: In order to realistically control the size of the instances, the qualified DR buses are set to be those having an original demand level within a certain interval, e.g., $[30, \infty]$ MW, and the DR upper bound r_k is set to 10% of the original demand D_k. Since the Polish data do not contain generators' quadratic cost coefficients, we artificially set them to 0.1 for all generators in all cases. Table 8 summarizes the case-specific setting, i.e., total number of lines, total demand, number of DR qualified buses, total available MW of DR. The last two columns are the number of lines which can possibly reach their upper bound and lower bound, respectively, determined by the bounding procedure. It can be seen that most lines will never reach their bounds, hence the corresponding multipliers are fixed to zero in the subsequent MIP solve. We have also tested various cases in which CONOPT reported infeasible. For such cases, CPLEX was able to terminate with infeasibility quickly (in less than 1 minute), which globally verifies that the cases are indeed infeasible.

We acknowledge that the DR model contains unavoidable nonconvex constraints and there can be no general guarantee that a global solution is always obtainable within a reasonable amount of time – it is an extremely hard problem. However, realistic instances/data usually have exploitable characteristics such as the ones exploited above,

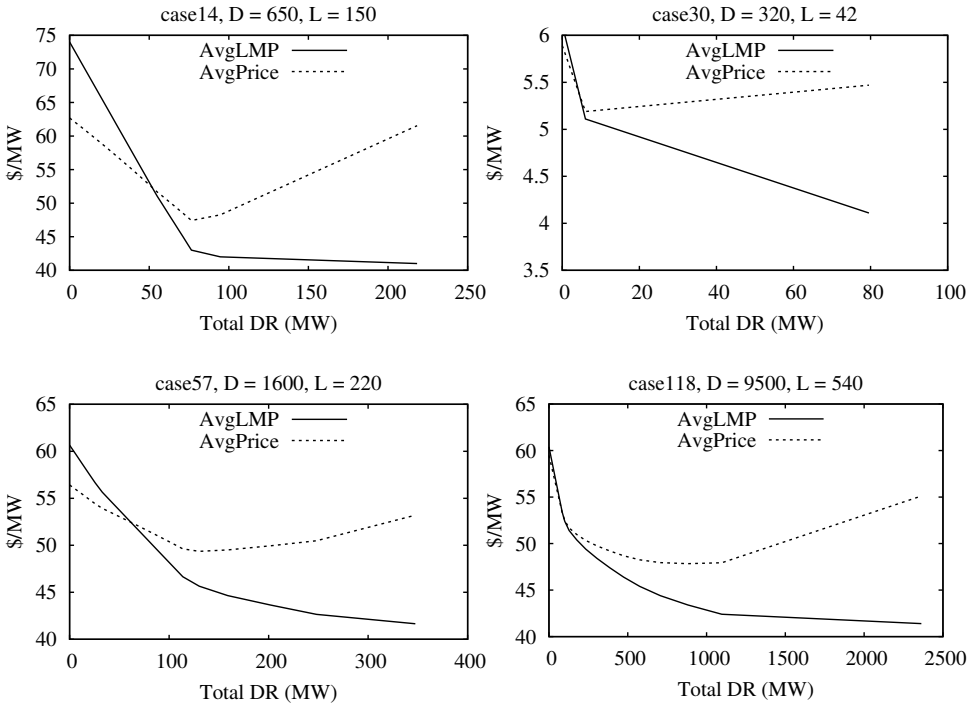
Table 7.: DR test results on Polish networks.

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

Table 8.: Settings and bounding results on Polish networks.

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

Figure 5.: Optimal solutions for decrementing C_1 levels on different data cases.

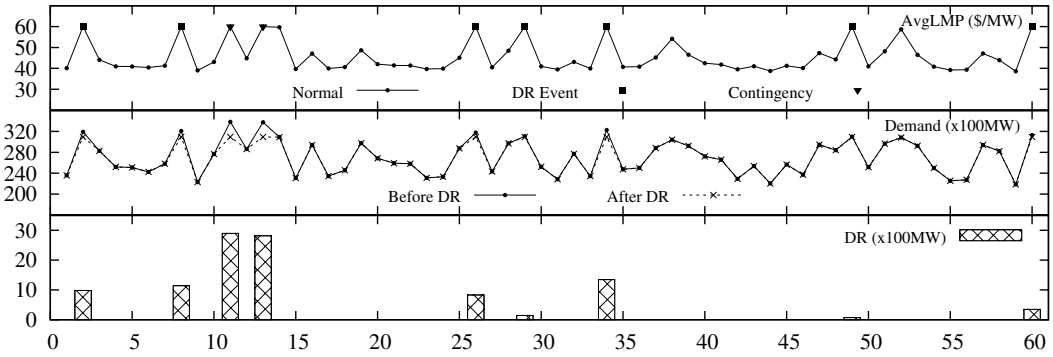


and use of SOS2 formulation (without need of artificial bounds) or big-M formulations (for faster solution) are acceptable for practice. Case-specific data analysis and simplification are necessary complements to the practical deployment of the model.

4.4 *On the cost-effectiveness condition*

We design experiments on four data cases to shed some light on the DR cost-effectiveness condition. The results are presented in Figure 5. Each subplot in the figure represents a series of experiments on the same data scenario (as annotated by the subplot title)

Figure 6.: Simulation results for the 300-bus case.



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

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

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

(2) The AvgPrice curves are convex shaped. As the Total DR level increases, the AvgPrice first decreases and then increases, and as long as it has not surpassed the original AvgPrice level, a feasible solution exists. The rationale is explained at the end of Section 4.1. However, for "good practice", we recommend the ISO consider dispatching demand response (i.e., running the DR1 model) when the AvgLMP is not too much above the LMP threshold C_1, or equivalently, set the threshold so that the resulting Total DR is substantially away from (i.e., lower than) the value that makes the AvgPrice start increasing.

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

4.5 *Simulation*

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

We do a series of 60 dispatch experiments each with a random demand profile. We

randomize the demand by multiplying the base case demand by a random scale factor uniformly distributed between 0.90 and 1.42. The LMP threshold is set to \$60/MW as we regard this as a reasonable level for demonstration. Given a random demand profile, in each run we first execute ED1, and take one of the following three actions depending on the outcome of ED1. Specifically, if ED1 gives an optimal solution and the corresponding AvgLMP is below the threshold, then no demand response is needed thus DR1 is not executed. If ED1 gives an optimal solution and the AvgLMP is above the threshold, then DR1 is executed to find an optimal DR (and implicitly ED) solution with a satisfactory AvgLMP. Finally, if ED1 fails to give an optimal solution, which indicates that the demand has exceeded the generation capacity and we cannot compute a value for the AvgPrice, then DR1 is executed with $C_2 = \infty$. In the last case, by executing DR1 we hope to not only control the AvgLMP below the threshold but also restore a feasible ED solution, and in exchange for this ambitious goal, we compromise the cost-effectiveness requirement by setting C_2 to infinity.

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

The simulation results, i.e., AvgLMP, demand and DR levels, are plotted in Figure 6. We mark three different system events: "Normal" if no DR is needed, "DR Event" if DR is dispatched to bring down the AvgLMP, and "Contingency" if DR is dispatched to restore the system feasibility. As shown in the experiments, DR1 is always successful to maintain the desired level of AvgLMP when a demand surge occurs (seven occurrences in the 300-bus case), and never fails to restore the system feasibility when needed (two occurrences). Furthermore, there is an apparent positive correlation between the demand level and the AvgLMP level, which indicates that demand response is indeed an effective way to control the market prices.

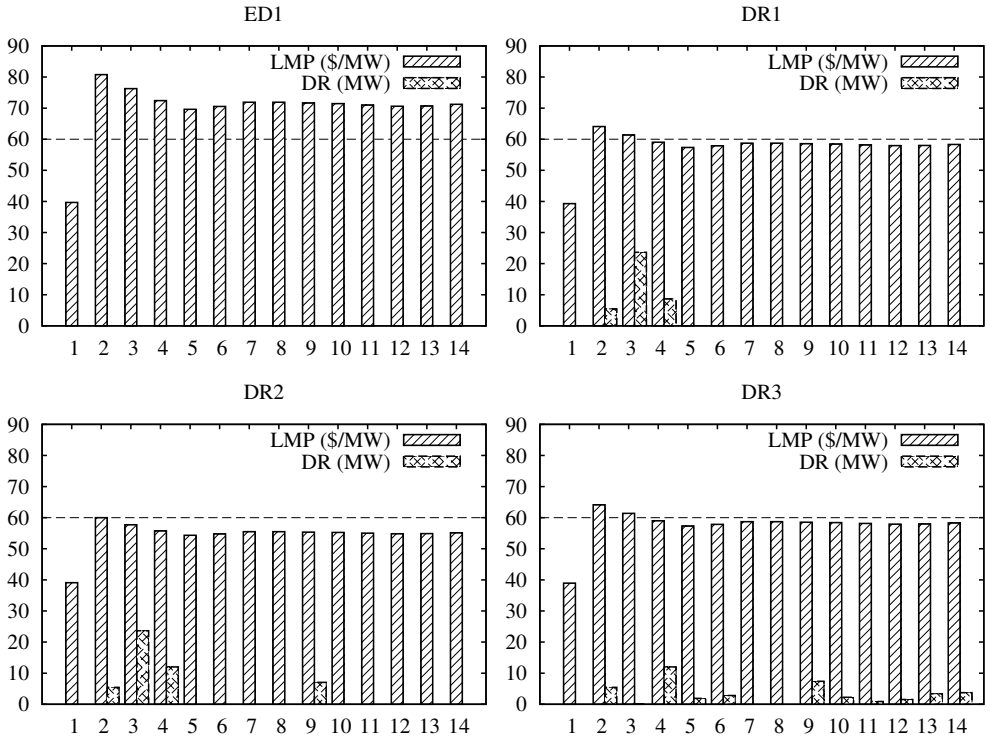
4.6 *Variants*

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

- DR2: Replace the constraint (11) by $\lambda_k \leq C_1, \forall k \in \mathcal{B}$ to impose the LMP threshold on every nodal LMP instead of on the AvgLMP.
- DR3: Set the objective function (10) by $L(\lambda) = \sum_{k\in\mathcal{B}} v_k r_k$ to account for the valuation v_k that the DR provider k places on a MW of demand reduction.

An illustrative experiment is performed on the 14-bus case with results presented in Figure 7. The total demand level is set to 650 MW and the line limit is 150 MW on every line. While ED1 gives an AvgLMP of \$74.01/MW, we set C_1 to \$60/MW, as depicted by the horizontal dotted lines in the subplots. For DR3, we set $v_3 = 200$ and $v_k = 100, \forall k \in \mathcal{B}/\{2\}$ to express a higher reluctance to dispatch demand response at node 3 compared to other nodes. For each node indicated on the horizonal axis, the bar on the left represents the LMP level and the bar on the right (if exists) represents the dispatched DR level at this node. Note that the LMP and DR levels share the same scale along the vertical axis but have different units, i.e., LMP is measured in \$/MW and DR in MW.

Figure 7.: Comparison of DR model variants on the 14-bus case.



As seen in the figure, DR1 was able to reduce the AvgLMP by dispatching a total of 37.7 MW of DR at nodes 2, 3 and 4. DR2 dispatched more (totaling about 48.1 MW) demand response at nodes 2, 3, 4 and 9, thus was successful to keep the maximum nodal LMP under \$60/MW as intended. DR3 apparently took into account the higher valuation v_3, and as a result dispatched much less DR at node 3 (about 0.02 MW) but dispatched more at various other nodes. These variants show that the bi-level DR model behaves sensibly and is flexible for further customizations.

5. Conclusions

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

A bi-level structure is necessary to model the interdependency between the energy price and the dispatch of demand response. To obtain a global solution, we have transformed the nonconvex Net Benefit Test constraint in the upper level into a linear form and investigated different methods to reformulate the complementarity relations arising from the lower level. We have developed a three-phase solution approach for large scale instances. Corroborated by extensive computational results, we conclude that:

(1) Local NLP solutions are always quick to compute and are useful to generate starting points.
(2) Realistic problems, despite their large scale, are not necessarily prohibitive to solve if data characteristics are sufficiently understood and exploited.
(3) The bi-level model is able to produce practical DR solutions and is readily extensible for other DR compensation rules.

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

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

Acknowledgements

This work is supported in part by Air Force Grant FA9550-10-1-0101, DOE grant DE-SC0002319, and National Science Foundation Grant CMMI-0928023.

References

[1] I. Aho, H. Klapuri, J. Saarinen, and E. M?kinen, *Optimal load clipping with time of use rates*, International Journal of Electrical Power & Energy Systems 20 (1998), pp. 269 – 280, Available at `http://www.sciencedirect.com/science/article/pii/S0142061597000665`.

[2] G. Andersson, *Modelling and Analysis of Electric Power Systems*, Eidgenössische Technische Hochschule Zürich (2008), lecture 227-0526-00, ITET ETH Zürich.

[3] J.M. Arroyo and A.J. Conejo, *Multiperiod auction for a pool-based electricity market*, IEEE Transactions on Power Systems 17 (2002), pp. 1225–1231.

[4] G. Bade, *Updated: Supreme Court uploads FERC Order 745, affirming federal role in demand response*, Utility Dive (2016), Available at `http://www.utilitydive.com/news/updated-supreme-court-upholds-ferc-order-745-affirming-federal-role-in-de/412668/`.

[5] J.F. Bard, *Practical Bilevel Optimization: Algorithms and Applications*, Nonconvex Optimization and its Applications, Vol. 30, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[6] R. Borlick, *Paying for Demand-Side Response at the Wholesale Level: The Small Consumers' Perspective*, The Electricity Journal 24 (2011), pp. 8–19.

[7] D.W. Caves and J.A. Herriges, *Optimal dispatch of interruptible and curtailable service options*, Oper. Res. 40 (1992), pp. 104–112, Available at `http://dx.doi.org/10.1287/opre.40.1.104`.

[8] H. Chao, *Demand Management in Restructured Wholesale Electricity Markets*, ISO New England (2010), Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.8726&rep=rep1&type=pdf`.

[9] R. Christie, *Power systems test case archive* (1993), http://www.ee.washington.edu/research/pstca.

[10] B. Colson, P. Marcotte, and G. Savard, *An overview of bilevel optimization*, Annals of Operations Research 153 (2007), pp. 235–256.

[11] B. Daryanian, R. Bohn, and R. Tabors, *Optimal demand-side response to electricity spot prices for storage-type customers*, IEEE Transactions on Power Systems 4 (1989).

[12] H.W. Dommel and W.F. Tinney, *Optimal power flow solutions*, IEEE Transactions on Power Apparatus and Systems PAS-87 (1968).

[13] ERCOT, *Interruptible Loads in ERCOT: A Brief History* (2009).

[14] ERCOT, *LOADS IN SCED Version 2, Preserving LMP Minus G*, ERCOT Demand Side Working Group (2015), Available at `http://www.ercot.com/content/wcm/key_documents_lists/51725/LMP_G_Concept_Paper_Outline_040615.docx`.

[15] FERC, *Demand Response Compensation in Organized Wholesale Energy Markets* (2011), http://www.ferc.gov/EventCalendar/Files/20110315105757-RM10-17-000.pdf.

[16] W.W. Hogan, *Providing Incentives for Efficient Demand Response*, PJM Demand Response (2009), FERC Docket EL09-68-000.

[17] W.W. Hogan, *Demand Response Pricing in Organized Wholesale Markets*, ISO/RTO Council (2010a), FERC Docket RM10-17-000.

[18] P. Interconnection, *Retail Electricity Consumer Opportunities for Demand Response in*

PJMs Wholesale Markets, Available at https://www.pjm.com/~/media/markets-ops/dsr/end-use-customer-fact-sheet.ashx.

[19] P. Interconnection, *PJM Manual 11: Energy & Ancillary Services Market Operations, Revision: 57* (2012).

[20] P. Interconnection, *The Evolution of Demand Response in the PJM Wholesale Market* (2014), Available at https://www.pjm.com/~/media/documents/reports/20141007-pjm-whitepaper-on-the-evolution-of-demand-response-in-the-pjm-wholesale-market.ashx.

[21] C. ISO, *MARKET PARTICIPATING LOAD TECHNICAL STANDARD FOR SUMMER 2000* (2000), Available at http://www.caiso.com/Documents/AttachmentE1-MarketParticipatingLoadTechnicalStandard_Summer2000.pdf.

[22] C. ISO, *Demand Response & Proxy Demand Resource C Frequently Asked Questions* (2011), Available at http://www.caiso.com/documents/demandresponseandproxydemandresourcesfrequentlyaskedquestions.pdf.

[23] C. ISO, *Monthly Demand Response Net Benefits Test Results November 2014* (2014), Available at http://www.caiso.com/Documents/DemandResponseNetBenefitTestResultsNovember2014.pdf.

[24] P.L. Joskow, *California's electricity crisis*, Oxford Review of Economic Policy (2001).

[25] D.S. Kirschen, *Demand-side view of electricity markets*, IEEE Transactions on Power Systems 18 (2003), pp. 520–527.

[26] Y. Liu, J.T. Holzer, and M.C. Ferris, *Extending the bidding format to promote demand response*, Energy Policy 86 (2015), pp. 82 – 92, Available at http://www.sciencedirect.com/science/article/pii/S0301421515002487.

[27] O. Mangasarian and S. Fromovitz, *The fritz john necessary optimality conditions in the presence of equality and inequality constraints*, Journal of Mathematical Analysis and Applications 17 (1967), pp. 37 – 47, Available at http://www.sciencedirect.com/science/article/pii/0022247X67901631.

[28] N.G. Mankiw, *Principles of Economics*, sixth ed., South-Western College Pub, 2011.

[29] J.A. Momoh, M.E. El-Hawary, and R. Adapa, *A review of selected optimal power ftow literature to 1993*, IEEE Transactions on Power Systems 14 (1999).

[30] A. Monticelli, M. Pereira, and S. Granville, *Security-constrained optimal power flow with post-contingency corrective rescheduling*, IEEE Transactions on Power Systems PWRS-2 (1987).

[31] J. Nocedal and S.J. Wright, *Numerical Optimization*, 2nd ed., Springer, 2006.

[32] S.S. Oren and S.A. Smith, *Design and management of curtailable electricity service to reduce annual peaks*, Oper. Res. 40 (1992), pp. 213–228, Available at http://dx.doi.org/10.1287/opre.40.2.213.

[33] L.E. Ruff, *Economic Principles of Demand Response in Electricity*, Edison Electric Institute,Washington D. C. (2002).

[34] F.C. Schweppe, M.C. Caramanis, R.D. Tabors, and R.E. Bohn, *Spot Pricing of Electricity*, Kluwer international series in engineering and computer science: Power electronics & power systems, Kluwer Academic, 1988, Available at http://books.google.com/books?id=Sg5zRPWrZ_gC.

[35] K. Spees and L.B. Lave, *Demand response and electricity market efficiency*, The Electricity Journal 20 (2007), pp. 69–85.

[36] C.L. Su and D. Kirschen, *Quantifying the effect of demand response on electricity markets*, IEEE Transactions on Power Systems 24 (2009), pp. 1199 – 1207.

[37] R. Walawalkar, J. Apt, and R. Mancini, *Economics of electric energy storage for energy arbitrage and regulation in New York*, Energy Policy 35 (2007), pp. 2558 – 2568.

[38] H. Wang, C.E. Murillo-Sánchez, R.D. Zimmerman, and R.J. Thomas, *On computational issues of market-based optimal power flow*, IEEE Transactions on Power Systems 22 (2007), pp. 1185–1193.

[39] H.J. Wellinghoff and D.L. Morenoff, *Recognizing the importance of demand response: The second half of the wholesale electric market equation*, Energy Law Journal 28 (2007).

[40] L. Xu, *Demand Response Net Benefits Test*, California ISO (2011), Available at http://www.caiso.com/documents/finalproposal-demandresponsenetbenefitstest.pdf.

[41] 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 26 (2011).