Pricing Data Center Demand Response

Zhenhua Liu, Iris Liu, Steven Low, Adam Wierman
California Institute of Technology
Pasadena, CA, USA
{zliu2,iliu,slow,adamw}@caltech.edu

ABSTRACT

Demand response is crucial for the incorporation of renewable energy into the grid. In this paper, we focus on a particularly promising industry for demand response: data centers. We use simulations to show that, not only are data centers large loads, but they can provide as much (or possibly more) flexibility as large-scale storage if given the proper incentives. However, due to the market power most data centers maintain, it is difficult to design programs that are efficient for data center demand response. To that end, we propose that prediction-based pricing is an appealing market design, and show that it outperforms more traditional supply function bidding mechanisms in situations where market power is an issue. However, prediction-based pricing may be inefficient when predictions are inaccurate, and so we provide analytic, worst-case bounds on the impact of prediction error on the efficiency of prediction-based pricing. These bounds hold even when network constraints are considered, and highlight that prediction-based pricing is surprisingly robust to prediction error.

Categories and Subject Descriptors

J.2 [**Computer Applications**]: Physical sciences and engineering

Keywords

data center, demand response, prediction based pricing, power network

1. INTRODUCTION

Demand response is widely recognized as a crucial tool for incorporating renewables into the grid, e.g., see recent reports from the National Institute of Standards and Technology (NIST) and the Department of Energy (DoE) [13,42]. Demand response programs provide incentives for customers to adapt their electricity demand to supply availability, for example, reducing their consumption in response to a peak load warning signal or request from the utility. Thus, demand response programs can help the grid transition from the paradigm of "generation follows demand" to one where, at least partially, "demand follows generation." Such a transition is fundamental to the integration of renewable energy because generation is becoming more intermittent and less controllable as renewable penetration increases.

SIGMETRICS'14, June 16–20, 2014, Austin, Texas, USA.
ACM 978-1-4503-2789-3/14/06 ...$15.00..

In this paper, we consider a promising demand response resource: data centers. Data centers are particularly well-suited for demand response. First, data centers represent *large loads* for the grid. In 2011, they consumed approximately 1.5% of all electricity worldwide and individual data centers can be 50 MW, or more [1, 20, 43]. Further, the energy consumption of data centers is *growing quickly*, by approximately 10-12% per year [1, 20, 30]. This growth is crucial for keeping pace with the growth of renewable adoption predicted for the coming years. Third, and most importantly, data centers are extremely *flexible loads*. Data centers are highly automated and monitored, e.g., the power load and state of IT equipment and cooling facilities can be continuously monitored and panoramically adjusted. For example, a recent empirical study by LBNL has quantified the flexibility in power usage of four data centers under different management approaches [20]. They find that 5% of the load can typically be shed in 5 minutes and 10% of the load can be shed in 15 minutes; and that these can be achieved *without* changes to how the IT workload is handled, i.e., via temperature adjustment and other building management approaches. Further, if workload management approaches are considered, the degree of flexibility can be larger, without additional time needed to shed the load. Significant research has recently gone into the design of such workload management, e.g., [10, 18, 21, 34, 39, 49, 52, 53].

Data center demand response today. Despite wide recognition of the demand response potential of data centers, the current reality is that data centers perform little, if any, demand response [20, 43].

In particular, the most common demand response program available for data centers is Coincident Peak Pricing (CPP), which is required for medium and large industrial consumers in many regions. These programs work by charging a very high price for usage during the coincident peak hour, often over 200 times higher than the base rate.[1] It is common for the coincident peak charges to account for 23% or more of a customer's electric bill according to Fort Collins Utilities [48]. Hence, a customer has a strong incentive to reduce usage during the peak hour. Although it is impossible to accurately predict exactly when the peak hour will occur, many utilities identify potential peak hours and send warning signals to customers (5-10 per month), which helps customers manage their loads and make decisions about their energy usage. For more details about CPP see [48].

Unfortunately, CPP programs are poorly designed from the perspective of data center demand response. Not providing response may incur a very large charge and providing a response may not actually result in any savings if the coincident peak does not occur during the warning period. As a

[1]The coincident peak hour is defined as the hour when the most electricity is demanded from the load serving entity (LSE).

result, even when they are forced to participate in such programs, data centers tend not to actively respond to signals. Further, even if they do respond, such programs extract very little flexibility from data centers. At best they obtain curtailment of usage a few times per month. This wastes the potential responsiveness of data centers.

Demand response market design. Although researchers have begun to focus on new market designs for data center demand response, e.g., [20, 28, 45, 46, 48], a clear vision remains elusive.

This is also true outside of the domain of data centers. Recently, the design of demand response programs has received considerable attention in a variety of settings, e.g., electric vehicles, pool pumps, and air conditioner cycling. Broadly speaking, the demand response programs that have emerged can be classified into two categories based on the interaction with users: either (i) users bid some degree of flexibility (supply) into the market, usually via a parameterized supply function, or (ii) users respond to a posted price, which was chosen using predictions about the available flexibility (e.g., supply functions). We term these approaches "supply function bidding" and "prediction-based pricing", respectively. Examples of proposed designs that use supply function bidding include [29, 50], and examples of prediction-based pricing designs include [12, 32, 40].

While each of these design approaches has pluses and minuses (as we discuss in Section 3), our focus on data centers motivates us to focus on prediction-based pricing programs.

In particular, a key assumption in the design and analysis of supply function bidding demand response programs is that users are *price takers*, i.e., they do not anticipate their impact on the price. Under this assumption, such designs can minimize the aggregate user cost while achieving the desired curtailment of demand. However, if this assumption is violated, and users act strategically, then inefficiency emerges in the market. Data centers are a canonical example of a user with market power – data centers can make up 50% of the load of the distribution circuits they are on, e.g., Facebook's data center in Crook County, Oregon. Thus, it is dangerous to treat them as price takers.

In contrast, prediction-based pricing is not nearly as impacted by market power issues. It is, however, highly dependent on the accuracy of the predictions of the user response to prices. Thus, there are still significant challenges in the design of such programs, and these issues are the focus of this paper.

Contributions of this paper. This paper makes two main contributions: (i) it quantifies the potential of data center demand response through a comparison with large-scale storage, and (ii) it presents and analyzes a novel design for prediction-based pricing of data center demand response. We discuss each of these in more detail in the following.

The potential of data center demand response: To quantify the potential of data center demand response we perform numerical case studies that compare the value of the flexibility provided by data centers with that provided by large-scale storage. In particular, in Section 2, we ask: *How much (optimally placed) storage can a data center replace?*

Interestingly, our results highlight that the flexibility provided by data centers is as valuable as, and often more valuable than, the flexibility provided by large-scale storage when it comes to ensuring that a distribution network meets its voltage constraints in the presence of a large-scale solar (PV) installation (see Figures 6). For example, the voltage violation frequency that comes from using a 30MW data center, which can provide 20% flexibility, is roughly equivalent to that of 1MWh of optimally-placed storage in the 46 bus distribution network from Southern California Edison that we consider. This is a quite conservative comparison because we assume storage with infinite charging speed (see Figure 5 for the impact of the charging rate). Further, the benefit of data center flexibility is robust to the placement within the distribution network – there are very few locations where the effectiveness of the data center drops considerably (see Figure 7).

Additionally, we look at the impact of a growing dichotomy in how IT companies address the sustainability of their data centers. Some companies, e.g., Apple [24], have invested heavily in on-site renewable generation; while others, e.g., Google [25], have tended to invest in renewable generation that is not co-located with their data centers. Both approaches have merits. Providing renewable generation on-site ensures that it is available where a very large and flexible load is located, but if renewable generation is not placed on site it can be placed in locations with better generation quality and/or cheaper installation costs.

Interestingly, our case studies highlight that co-location of data centers and large-scale PV installations is very efficient. In particular, the voltage violation frequency when the data center is placed at the same bus as the PV in a distribution network is within 4% of optimal. However, it is worth noting that a data center with local PV is not nearly as efficient at helping manage a large-scale PV installation as a data center without local PV. In particular, a 20MW data center with 20% flexibility and a co-located 5MW solar installation provides the same voltage violation frequency as 0.3MWh of optimally-placed storage, i.e., 25% less than a 20MW data center with no local PV. Thus, having PV at the location of the data center is better than having it elsewhere, due to the complementary diurnal patterns of each, but a data center without local renewables is a more valuable resource for grid management than a data center with local renewables.

Prediction-based pricing: Given the potential of data center demand response identified in the first half of the paper, the second half of the paper focuses on designing a demand response program that can extract this flexibility. As we have already discussed, prediction-based pricing is an appealing candidate given the market power data centers maintain. Thus, in Sections 4 and 5 we present and analyze a design for prediction-based pricing. Section 4 introduces the design in a context without the constraints imposed by the distribution network, and then Section 5 incorporates the network constraints into the design and analysis.

The analysis in these sections is focused on three issues. First, we focus on the impact of the accuracy of predictions on the efficiency of the market design. This is, perhaps, the most crucial issue for prediction-based pricing programs. Our results provide an analytic characterization of worst-case efficiency bounds under the assumption of quadratic objective functions (Theorem 2), which is a common assumption in the power system literature. In particular, we derive tight bounds on the competitive ratio of prediction-based pricing that highlight the impact of the variability of the prediction error.

The second issue is the contrast between prediction-based pricing and supply function bidding. As we have mentioned, prediction error hurts the former while market power hurts the latter. Thus, the natural question becomes: Under which settings is prediction-based pricing appropriate? By contrasting our results with those of [50] on the efficiency of supply function bidding, we give an explicit characterization in terms of market power and prediction error of when prediction-based pricing outperforms supply function bidding (Figure 10). Broadly speaking, the comparison highlights that prediction-based pricing is appropriate for data center demand response when prediction errors are moderate and the data center has significant, local market power.

Finally, the third issue our analysis focuses on is the impact of network constraints on the design and efficiency of prediction-based pricing. In our analysis, the network constraints manifest themselves as a chance constraint on the price that ensures that voltage violations in the network are rare. But, despite constraints on the prices, we prove that

the efficiency of prediction-based pricing is not impacted by the network constraints, i.e., the competitive ratio remains unchanged (Theorem 5). This represents the first analytic bound on the efficiency of prediction-based pricing in the presence of network constraints.

2. QUANTIFYING THE POTENTIAL OF DATA CENTER DEMAND RESPONSE

Before looking at the design of market programs to extract flexibility from data centers, it is crucial to quantify the potential of such programs. In this section, we accomplish this by contrasting the flexibility provided by data centers with that provided by large-scale storage.

Often, when people think of the challenges for grid management that result from renewable energy, the thought is: "if only we had large-scale storage..." The problem is that large-scale storage is expensive, which leads to the consideration of demand response. But, besides cost, demand response also has other benefits over storage. In particular, storage needs to be pre-charged to be ready for use, while demand response has no such requirement. However, storage has benefits as well. First, the placement of storage is more flexible than that of data centers. Second, apart from pre-charging, storage does not bring with it any electricity demand, whereas data center demand response inherently requires the presence of a large load in the distribution network.

In the experiments that follow, we study the impact of these competing factors in order to understand how the potential of data center demand response compares to large-scale storage. In particular, we ask: *How much (optimally placed) storage can a data center replace?* Since we focus on bounding the potential of data center demand response in this section, we do not model market factors. Rather, we assume that the load serving entity (LSE) can call on the data center and storage as needed. Market design is considered in the second half of the paper.

2.1 Setup

To quantify the potential of data center demand response, we study a situation where a distribution network has a large-scale solar installation and either large-scale storage or a data center to help manage the intermittency of the solar installation.

The performance objective we consider is that of minimizing the frequency of violations of voltage constraints in the distribution network. To measure this frequency we sum the number of buses with voltage violations at each time slot and over time, i.e., the number of buses that result in voltages outside the tolerance bounds given by the network. For instance, a violation frequency 0.1 means on average, each bus experiences voltage violation in 10% of the time. We contrast the frequency of voltage violations when a data center is present and when large-scale storage is present.

Distribution network. We consider two distribution networks in our experiments. Both are distribution networks from the Southern California Edison (SCE) utility company. The first is a 47 bus network (Figure 1) and the second is a 56 bus network (Figure 2). Both are described in detail in [15].

There is no conventional generation on these distribution networks. All power comes from the substation bus, a.k.a., the zero bus, and the solar installation (which we describe later). The demands are taken from SCE load profiles [23], except for the data center, for which the demand is described later.

Given these settings, a significant amount of the solar generation can be transmitted out of the distribution network through the substation bus. However, because we consider a large-scale solar installation, when the installation has near

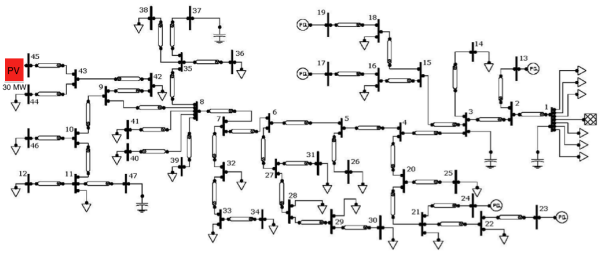


Figure 1: SCE 47 bus network.

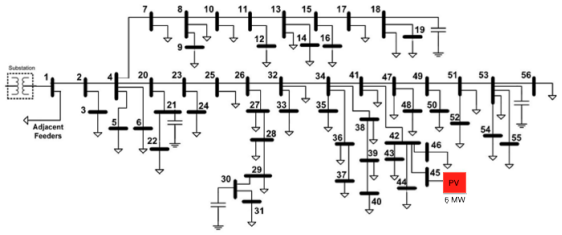


Figure 2: SCE 56 bus network.

peak generation, the network constraints become binding and voltage violations are common. Note that the voltage constraint we consider is taken directly from the network tolerance specifications, and is 3%. The number of violations in our simulations are consistent with previous work on these networks, e.g., [14, 15]. The presence of storage or the data center is used to help avoid such violations.

For our simulations, given the network, the power flow is computed for a sequence of discrete time steps $t = 1, \ldots, T$ using MatPower [54]. Then, we analyze the voltages for each time step and determine the number of buses that have voltage violations. Finally, we sum the voltage violation events from all buses over all time steps, and use it to calculate the violation frequency. The length of the time steps that we consider is one minute.

Renewable energy. To model a solar installation placed within a distribution network, we use solar irradiance data from Los Angeles, CA in February 2012 [26] to alter the power load at the bus where the solar (PV) generation is located. Thus, irradiance data acts like an installed solar capacity. The trace is illustrated in Figure 3(a).

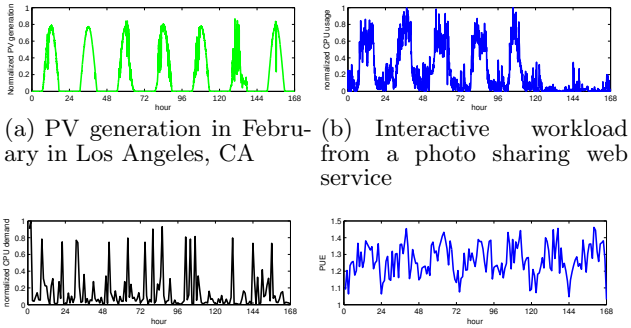
For the experiments reported, the PV is placed at bus 45 and sized at 30MW for the 47 bus network, and also placed at bus 45 but sized at 6MW for the 56 bus network[2]. The results do not qualitatively change when other locations and sizes are considered.

Data center model. To incorporate a data center into the experiments, we need to model two aspects: the power usage of the data center over time and the flexibility in the power usage of the data center.

To model the power usage of a data center, we adopt the model used in [3, 33, 34, 38], which provides a simple but representative characterization. In particular, we model the power demand of the data center as a function of the workload, including interactive (inflexible) and delay-tolerant (flexible) workloads, and the cooling efficiency, as measured by the Power Usage Effectiveness (PUE).

To model the workload we use two traces. The interactive workload trace is from a popular web service application with more than 85 million registered users in 22 countries

[2]We use different size of PV because the capacities of these two networks are different.



(a) PV generation in February in Los Angeles, CA

(b) Interactive workload from a photo sharing web service

(c) Facebook Hadoop workload

(d) Google data centers PUE

Figure 3: One week traces for (a) PV generation, (b) inflexible workload, (c) flexible workload, and (d) cooling efficiency.

(see Figure 3(b)). The trace contains average CPU utilization and memory usage as recorded every 5 minutes. The peak-to-mean ratio of the interactive workload is about 4. The delay-tolerant workload information comes from a Facebook Hadoop trace (see Figure 3(c)). The total demand ratio between the interactive workload and batch jobs is 1:1. This ratio can vary widely across data centers, but we choose this ratio as representative based on discussions in [35].

To model the data center power efficiency including cooling efficiency, we use a trace of the PUE from Google data centers. As shown in the figure, the PUE varies between 1.05 to 1.45, and has strong diurnal pattern, i.e., higher around noon because outside air temperature is higher.

To combine the workload traces and the PUE to obtain a model of the total power demand of the data center, we use the following relationship.

$$v(t) = PUE(t)(a(t) + b(t)),$$

where $a(t)$ is the power demand from the inflexible workload and $b(t)$ is power demand from the flexible workload demand. Note that the data center power demand has the same average value as the PV generation with the same capacity in the distribution network.

The second aspect of the data center model that we must include is the flexibility of the power demand. For this, our model is informed by the recent empirical study [20], which we have discussed in the introduction.

To model the range of flexibility in our experiments, we denote the demand flexibility of the data center by e and allow the data center to have demand within

$$[(1-e)v(t), \min\{(1+e)v(t), C_d\}],$$

where C_d is the capacity of the data center and $v(t)$ is the data center power demand at time t if no demand response is called upon. Thus, $e = 0.10$ could be achieved without workload management, and $e = 0.20$ can be achieved with some workload management, e.g., quality degradation or load deferral. When demand response is required from the data center, the load that minimizes the voltage violation rate is provided by the data center.

Since a downside of data center demand response is that the LSE cannot control the placement of the data center, the placement of the data center is varied during our experiments in order to understand robustness to "bad" data center locations. Note that we assume there is no cost associated with the demand shaping of data center; however the cost of this could be incorporated easily if desired.

Storage model. To incorporate large-scale storage into our model, we adopt a standard model, e.g., from [19,27,31,47]. In order to provide a conservative estimate of the potential of data center demand response we assume perfect storage, i.e., no loss or leakage. This means that, at all times t, the storage level for the next time step is $L(t+1) = L(t)+u(t)$, where $u(t)$ is the energy change in the level at time t. Note that $u(t)$ is positive if we are charging the storage and negative if we are discharging. Of course, $L(t) \in [0, C_s]$ for all t, where C_s is the storage capacity. So, $u(t) \in [-L(t), C_s - L(t)]$, where C_s is the storage capacity. This range quantifies the amount of flexibility that can be called upon by the LSE. As in the case of the data center, the LSE will call upon a feasible $u(t)$ that minimizes the voltage violations. Although more advanced energy storage management policy could be used to further improve the benefit, here we use this simple greedy strategy for both data center and energy storage for comparisons.

For most of the experiments we assume that the storage can completely charge and discharge in one time step. This is, of course, unrealistic, but it allows us to give a conservative estimate of the benefits of data center demand response. We do evaluate the impact of limitations on the charging rate in Figure 5 in order to highlight the degree to which this assumption leads to an underestimate of the value of data center demand response.

As we have already mentioned, a benefit of storage is that it can be placed optimally within a network. The optimal placement of the storage is at bus 44 for the 47 bus network and bus 53 for the 56 bus network. Note that the optimal placement is robust as we adjust the capacity of the storage in our experiments.

2.2 Case studies

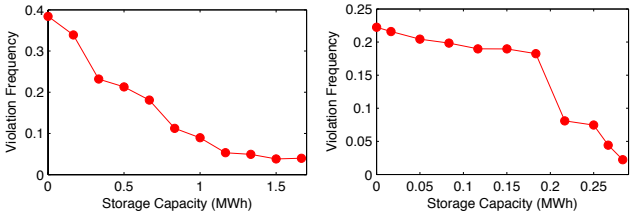
Using the setting described above, our focus is on two comparisons that each sheds light on the potential of data center demand response: (i) a comparison between data center demand response and large-scale storage, and (ii) a study of the impact of on-site renewable generation on data center demand response.

Data center demand response versus large-scale storage. To contrast large-scale storage with data center demand response, we first need to quantify the benefits from large-scale storage. This is done in Figures 4 and 5, which show the impacts of the storage capacity and the storage charging rate on the voltage violation rate in the two distribution networks. Figure 4 highlights that, as expected, the voltage violation rate decreases as storage capacity grows. However, it also shows that this relationship is nonlinear and depends strongly on the network structure. Similarly, Figure 5 highlights that, as expected, a smaller charging rate increases the frequency of voltage violations. However, the impact of a smaller charging rate is, perhaps, more significant than expected. Note that for our experiments we conservatively estimate the value of data center demand response by comparing it with storage having a charging rate of 1, i.e., we assume that the storage can completely charge and discharge in one minute. This is unrealistic, but provides a lower bound on the value of data center demand response.

Given the characterization of storage, we can now highlight the value of data center demand response in terms of the "equivalent" storage capacity, i.e., in terms of the capacity of optimally-placed large-scale storage necessary to provide the same voltage violation frequency. The results of this comparison are shown in Figure 6.

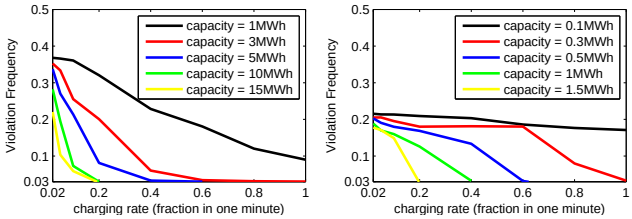
Naturally, the amount of storage equivalent to data center demand response grows with the size of the data center. However, the capacity plateaus after the data center size grows beyond 35MW for the SCE 47 bus network and beyond 6MW for the SCE 56 bus network. Note that this is a consequence of two differences between the networks – the structure and the size of the PV installation (30MW vs. 6MW).

But, in both networks, Figure 6 highlights that data center demand response has a significant potential. In particular,



(a) SCE 47 bus network (b) SCE 56 bus network

Figure 4: Impact of energy storage capacity, C_s, on the voltage violation rates.



(a) SCE 47 bus network (b) SCE 56 bus network

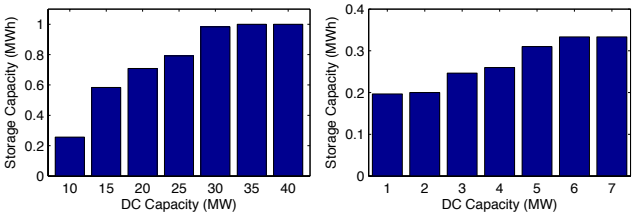
Figure 5: Impact of energy storage charging rate on the voltage violation rates.

recall that the comparison in this plot assumes storage with infinite charging speed, i.e., a charging rate of 1, and is thus quite conservative (as illustrated in Figure 5). Additionally, the cost of storage is upwards of $500/kWh for lithium-ion batteries (which have small charging rates) and upwards of $5000/kWh for technologies with fast charging rates, such as flywheels. Thus, the flexibility provided by one 30MW data center is worth upwards of $500,000 - $5,000,000. These numbers are conservative estimates, and grow considerably if a slower charging rate is used in the simulations or if the flexibility of the data center, e, is increased.

Figures 7 and 8 delve into the comparison of data center demand response and large-scale storage in more detail for each of the networks. In Figure 7, we fix the capacity of the data center to 20MW, which is a representative size for today's IT companies, and then investigate the impact of the degree of data center flexibility, e, and the placement of the data center. For example, Figures 7(a)-7(c) highlight that the voltage violation rates decrease as data center power demand becomes more flexible. In particular, a 20MW data center with 20% power demand flexibility placed at the PV location is equivalent to 0.67MWh of optimally-placed storage in the 47 bus distribution network. Further, Figure 7(d) shows that the benefit of data center flexibility is robust to the placement of the network in the distribution network, i.e., there are very few locations where the effectiveness of the data center drops considerably and many locations that are near-optimal, e.g., placing the data center at the location of the PV (Figure 7(b)). Figure 7(d) also illustrates that a 20MW data center is better than 0.33MWh of storage pretty much uniformly. The results in a SCE 56 bus network are similar, as shown in Figure 8.

Should data centers invest in co-located renewables?. There is a dichotomy right now in how IT companies address the sustainability of their data centers. Some companies, e.g., Apple [24], have invested heavily in on-site renewable generation; while others, e.g., Google [25], have tended to invest in renewable generation that is not co-located with their data centers.

Both approaches have merits, as we have discussed in the introduction. For the purpose of this paper, the key distinction is how on-site renewable generation impacts data center demand response. This context highlights another benefit of on site renewable generation – it ensures that the data center is placed close to the renewables, which is very often a near-optimal placement for demand response purposes.



(a) SCE 47 bus network (b) SCE 56 bus network

Figure 6: Diagram of the capacity of storage necessary to achieve the same voltage violation frequency as data centers of varying sizes. The data center has flexibility $e = 0.2$.

First, Figure 7(d) highlights that co-location of data centers and large-scale PV installations is very efficient. In particular, the voltage violation frequency when the data center is placed as the same bus as the PV in a distribution network is within 4% of optimal.

However, it is worth noting that a data center with local PV is not nearly as efficient at helping to manage a large-scale PV installation as a data center without local PV, by comparing Figure 7(c) with 9(a). In particular, a 20MW data center with 20% flexibility and a 5MW solar installation provides the same voltage violation frequency as 0.3MWh of optimally-placed storage when helping to manage 30MW of PV elsewhere on the distribution network, i.e., 25% less than a data center with the same flexibility but no local PV.

Thus, having PV at the location of the data center is better than having it elsewhere, due to the complementary diurnal patterns of each, but a data center without local renewables is a more valuable resource for grid management than a data center with local renewables.

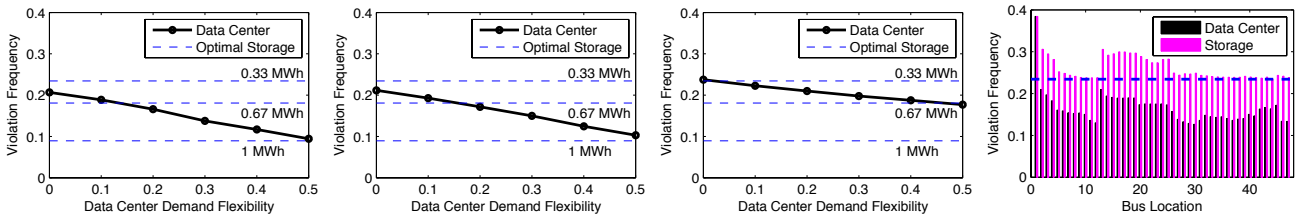
3. MARKET CHALLENGES FOR DATA CENTER DEMAND RESPONSE

The previous section highlights that data centers have the potential to be as useful as, if not more useful than, storage for demand response. However, realizing this potential is challenging. Data centers today tend not to participate in demand response programs and, if they do, they tend to participate passively.

For example, the most common program for data center demand response today is coincident peak pricing and, though many data centers are forced to participate, they typically do not actively respond to the warnings issued by the utility. Further, even if they did, this would mean that the data center provided flexibility only 5-10 times a month, which is far from the amount of available flexibility. Such limited signaling from the LSE to the data center cannot possibly extract the potential flexibility illustrated in Section 2. On the other hand, if the utility company sends too many warning signals, data centers simply will not respond to them.

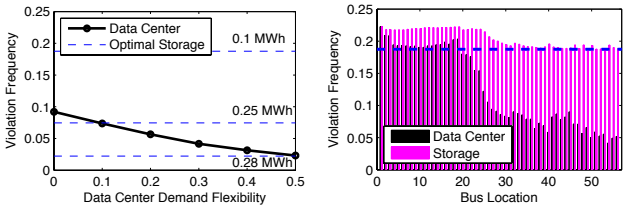
Thus, realizing the potential of data center demand response requires new market programs. While the design of market programs for data centers is only beginning to receive attention, there has been considerable work on the design of demand response programs in other contexts in recent years, e.g., [2, 9, 16, 17, 22, 29, 37, 50]. Much of this work focuses on the design of residential programs for, e.g., electric vehicles, pool pumps, and air conditioner cycling.

Broadly speaking, the demand response programs that have emerged can be classified into two categories based on the interaction with users: either (i) users bid some degree of flexibility (supply) into the market, usually via a parameterized supply function, or (ii) users respond to a posted price, which was chosen using predictions about the avail-



(a) Data center placed at the optimal storage location (b) Data center placed at the PV location (c) Data center placed at bus 2 (d) Data center vs. storage

Figure 7: Comparison of a 20MW data center to large-scale storage in a 47 bus SCE distribution network. (a)-(c) show the violation frequency as a function of the amount of data center flexibility, e, and compare to optimally placed storage, for different locations of the data center. (d) shows the violation frequency resulting from a data center with $e = 0.2$ versus 0.33MWh of storage, for each location.



(a) Data center placed at bus 53 (b) Data center vs. storage

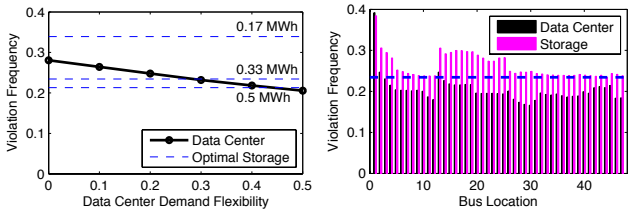
Figure 8: Comparison of a 4MW data center to large-scale storage in a 56 bus SCE distribution network. (a) shows the violation frequency as a function of the amount of data center flexibility, e, and compare to optimally placed storage. (b) shows the violation frequency resulting from a data center with e = 0.2 compared to 0.07MWh of storage at each location.

able flexibility (e.g., supply functions). We discuss each of these approaches below and highlight the challenges of each when it comes to data center demand response.

Supply function bidding. In this approach to market design each user announces a bid to the load serving entity (LSE) that specifies the amount load will be curtailed as a function of the price, a.k.a., a supply function. The form of the supply function is typically fixed to have some parametric form and the bid specifies the parameter. The LSE then chooses a market clearing price that achieves the demand response target. Examples of market designs of this form include [29, 50] and the references therein.

Typically, a key assumption in the design and analysis of such markets is that users are *price takers*, i.e., they do not anticipate their impact on the price. Under this assumption, such designs can minimize the aggregate user cost while achieving the desired curtailment of demand. However, if this assumption is violated, and users act strategically, then inefficiency emerges. Recent work has begun to characterize this inefficiency, and the basic conclusion is that it can be extreme [50].

While the assumption that users are price takers is natural in many demand response settings, e.g., residential pool pump and air conditioner programs; it is quite problematic in the case of data centers. A residential user does not have the power to manipulate prices, i.e., does not have *market power*, but a large data center can make up 50% of the load of the distribution circuits they are on, e.g., Facebook's data center in Crook County, Oregon. Thus, data centers are a canonical example of an agent with market power. This observation motivates the consideration of prediction-based pricing in the current paper.



(a) Data center placed at bus 2. (b) Data center vs. storage

Figure 9: Comparison of a 20MW data center with a co-located 5MW PV installation to large-scale storage in a 47 bus SCE distribution network. (a) depicts the data center located at bus 2. (b) shows the violation frequency resulting from a data center with $e = 0.2$ compared to 0.33MWh of storage, for each location.

Prediction-based pricing. In this approach to market design, the LSE presents the user a price that they will pay the user for curtailment, and then the user responds. Examples of designs of this type can be found in [12, 32, 40] and the references therein. The challenge in such a program is how the LSE should determine the price.

If the LSE knew the supply function of the users, then it could easily set a price to extract the desired curtailment. However, the LSE does not have this information, and since it is not provided by the user (as in the supply function bidding approach), the LSE must *predict* the user supply functions. Then, using the predicted supply functions, the LSE can determine an appropriate price to induce the desired curtailment.

Clearly, one should expect prediction-based pricing to only be appropriate if supply functions can be predicted accurately. This is a challenge in the data center environment since the supply functions of the data center may depend on the workloads and weather (among other things), each of which is highly non-stationary.

The key task in the remainder of the paper is to characterize how accurate predictions must be for the prediction-based pricing approaches to be useful. Interestingly, the contrast between the performance of prediction-based pricing and supply function bidding depends on the balance between the market power of data centers and the accuracy of supply function prediction. We discuss this in Section 4.3 by contrasting our results with those in [50].

4. PREDICTION-BASED PRICING FOR DATA CENTER DEMAND RESPONSE

In this section, we develop a market program for extracting flexibility from data centers. Given the discussion in

Section 3, our focus is on prediction-based pricing. In particular, the goal of this section is (i) to optimally design prediction-based pricing programs for data center demand response, (ii) to quantify the efficiency loss created by prediction error in such programs, and (iii) to contrast prediction-based pricing with supply function bidding. We do this in the context of a classic supply function model in this section, and then show how to incorporate distribution network constraints in Section 5.

4.1 Model formulation

The setting we consider here is where an LSE wishes to procure a total amount D of load reduction from a set of users indexed by $1, 2, \ldots, n$. We focus on one time step and ignore the network constraints in this section.

To procure this load reduction, the LSE announces a price p and pays user i the amount ps_i when user i reduces consumption by $s_i \geq 0$. The market design task is to design p so that the LSE achieves the desired amount of curtailment.

To model the user reaction to the price, we assume that each user i incurs a cost $C_i(d_i)$ when she reduces her consumption by an amount $d_i \geq 0$. We assume some parameter(s) of the cost function $C_i(\cdot)$ are random so that for each $d_i \geq 0$, $C_i(d_i)$ is a random variable. This randomness captures the fact that, in practice, the LSE does not know the parameter(s) of $C_i(\cdot)$ exactly. However, the LSE may be able to estimate the parameters from historical consumption data and the effect of estimation error can be modeled through the distribution of the random parameter(s) in $C_i(\cdot)$.

We assume that user i strategically reduces her consumption when faced with a price p in a profit maximizing manner. Let $s_i(p)$ denote the unique cost minimizing curtailment. Specifically, for each realization of $C_i(\cdot)$, denoted by $c_i(\cdot)$, user i solves

$$\min_{d_i \geq 0} \quad c_i(d_i) - pd_i, \tag{1}$$

which gives

$$s_i(p) \;=\; c_i'^{-1}(p). \tag{2}$$

To ensure that a unique solution $s_i(p) \geq 0$ always exists, we impose that each realization $c_i(\cdot)$ of the random cost function $C_i(\cdot)$ is non-negative, increasing, strictly convex, twice continuously differentiable, and has $c(0) = 0$. Additionally, note that we have implicitly assumed that the randomness in $C_i(d_i)$ is independent of the price p. These are standard assumptions in the electricity market literature, e.g., [4, 8, 41, 51].

Given the model above, the total demand response the LSE achieves with price p is the random quantity $\sum_i s_i(p)$. Given the uncertainty about the user costs, this curtailment likely does not exactly match the demand response target D. We assume that the penalty for deviation from the target is captured through a penalty function $h(\cdot)$. In particular, the penalty is $h\left(D - \sum_i s_i(p)\right)$. We assume this penalty function $h(\cdot)$ is convex, non-negative, has a global minimum $h(0) = 0$, and is continuously differentiable with $h'(0) = 0$. These assumptions ensure that the optimal price is well-defined, see Theorem 1.

4.2 The efficiency of prediction-based pricing

Given the setting described above, our task is to first understand how to price, and then to understand the efficiency loss due to prediction error. We start with the case where the LSE has perfect predictions of the data center supply functions, i.e., with perfect foresight. Then, we move to the case where the LSE has only predictions of the data center supply functions. Finally, we quantify the efficiency loss that results from this uncertainty.

Throughout, to evaluate the efficiency of the LSE's choice of p we use a notion of social cost defined as the sum of the penalty of deviation from the demand response target D and the total user costs, i.e.,

$$G(p) \;:=\; h\left(D - \textstyle\sum_i s_i(p)\right) + \sum_i C_i(s_i(p)). \tag{3}$$

Note that the social cost $G(p)$ is random from the LSE's perspective for two reasons: both $C_i(d_i)$ and the user responses $s_i(p)$ are random. But, the randomness in both of these originates from the randomness of the user cost functions $C_i(\cdot)$.

Pricing with perfect foresight. Before looking at the design of prediction-based pricing, it is informative to consider how an LSE with perfect foresight would price. In particular, consider an LSE that is clairvoyant, i.e., has perfect knowledge about the cost function, and can choose $p(\omega)$ to minimize $G(p)$ for the realization on instance ω. We use ω here to highlight this price is for each realization ω. In this situation, the price chosen by the LSE is summarized in the following theorem.

Theorem 1. *For each realization ω, there exists a unique minimizer p^* such that*

$$p^*(\omega) \;=\; h'\left(D - \sum_i s_i(p^*(\omega))\right), \tag{4}$$

and $0 \leq p^* < \overline{p}$, where $\overline{p}$ satisfies $\sum_i s_i(\overline{p}) = D$.

An interesting aspect of this theorem is that the optimal price is strictly lower than any price $\overline{p}$ that would exactly satisfy the demand response target.

Of course, using p^* in practice is infeasible. However, it provides an important benchmark for the performance of prediction-based pricing without perfect foresight. Note p^* is random from LSE's perspective, since the cost function realizations are random. Thus, the strategy yields an expected cost which we denote as follows

$$\mathbb{E}\left[G(p^*)\right] \;=\; \mathbb{E}\left[\min_{p \geq 0} G(p)\right]. \tag{5}$$

Prediction-based pricing. In practice, the LSE does not know the exact realization of the user cost function, thus it can only use predictions of the cost functions in order to choose a price $\hat{p}$. Here, we focus on the case where the LSE chooses $\hat{p}$ in order to minimize the expected cost that results, i.e.,

$$\hat{p} \in \operatorname*{argmin}_{p \geq 0} \mathbb{E}\left[G(p)\right]. \tag{6}$$

This yields the following

$$\mathbb{E}\left[G(\hat{p})\right] \;=\; \min_{p \geq 0} \mathbb{E}\left[G(p)\right]. \tag{7}$$

Of course, other objectives that include some form of risk management may also be interesting to consider in future work. Note that we assume that users know their own cost function, and can therefore choose their curtailment amount $s_i(p)$ based on the true cost function $c_i(\cdot)$ (cf. (2)). This means the random events that determine the $C_i(\cdot)$ are revealed only to individual users, but not to the LSE (or other users).

The efficiency of prediction-based pricing. Clearly the cost when pricing with perfect foresight is no larger than the cost when using prediction-based pricing. Here, our goal is to understand how much is lost because of uncertainty about the cost function.

To quantify this efficiency loss, we study the worst-case ratio between the cost of prediction-based pricing and the cost of pricing with perfect foresight. This is a *competitive*

ratio. In particular, let F be the joint distribution of all random variables in the model, and $\mathcal{F}$ be a set of permissible distributions. Then the competitive ratio we consider is formally defined as $CR = \max_{F\in\mathcal{F}} \frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*)]}$.

To evaluate the competitive ratio, we need to restrict ourselves to the quadratic penalty function and cost functions, i.e.,

$$h\left(D - \sum_i s_i(p)\right) := \frac{q}{2}\left(D - \sum_i s_i(p)\right)^2 \text{ and } (8)$$

$$C_i(d_i) := \frac{1}{2X_i}d_i^2, \quad (9)$$

where $q > 0$ is known, but $X_i > 0$ are random variables to the LSE. Note that this may seem restrictive, but this form is standard within the electricity markets literature, e.g., [4, 8, 41, 51].

Then, for each realization, we can explicitly compute the curtailments of the users. Specifically, from (2):

$$s_i(p) = X_i p \quad \text{and} \quad C_i(s_i(p)) = \frac{1}{2}X_i p^2 \quad (10)$$

Now, we can state the main theorems of this section, which bound the competitive ratio of prediction-based pricing in terms of the variability of prediction errors (Theorem 2) and show that the bound is tight (Theorem 3). Let $X := \sum_i X_i$, denote the variance of X by $\mathbb{V}[X]$, and denote the squared coefficient of variation of X by $\mathbb{C}^2[X] = \mathbb{V}[X]/(\mathbb{E}[X])^2$.

Theorem 2. *Suppose the penalty function and cost functions are given by* (8) *and* (9)*, respectively. Then the competitive ratio is upper bounded by*

$$\frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*(\omega))]} \leq 1 + \frac{(q\mathbb{E}[X])^2\mathbb{C}^2[X]}{1 + (q\mathbb{E}[X])(\mathbb{C}^2[X]+1)}. \quad (11)$$

Moreover $\hat{p} \leq \mathbb{E}[p^*]$*, with equality if and only if* $\mathbb{V}[X] = 0$.

Theorem 3. *Under the conditions of Theorem 2 the bound in* (11) *is asymptotically tight, i.e., for all* $\epsilon > 0$*, there exists a probability density function* $f(X)$ *such that*

$$\frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*(\omega))]} \geq 1 + \frac{(q\mathbb{E}[X])^2\mathbb{C}^2[X]}{1 + (q\mathbb{E}[X])(\mathbb{C}^2[X]+1)} - \epsilon. \quad (12)$$

Before moving on, it is worth making a few remarks about these theorems.

First, the results apply both when the prediction errors from users are independent and when they are correlated.

Second, the competitive ratio decreases as the variability of X decreases. This means that a better prediction can provide better performance. In the extreme case, when there is no randomness in X, i.e., perfect foresight, then Theorem 2 guarantees that the competitive ratio is 1. Moreover $\hat{p} = p^*(\omega)$ and $G(\hat{p}) = G(p^*)$. In contrast, when there is prediction error, the LSE tends to have lower price to prevent over provisioning. This is because the attained curtailment $\sum_i s_i(p)$ is an increasing function of the price p. Specifically, we have

$$\hat{p} = \frac{q\mathbb{E}[X]\,D}{q\mathbb{E}[X^2] + \mathbb{E}[X]} \text{ and } p^* = \frac{qD}{qX+1}, \quad (13)$$

which both increase with q.

Third, it is interesting to note that the competitive ratio does not depend on the particular distributional form beyond the first and second moments of an aggregated value. This is due to the quadratic nature of both the user cost functions $C_i(\cdot)$ and the penalty function $h(\cdot)$. One should expect that if these functions were polynomials with higher order then higher order moments would show up in the competitive ratio.

Finally, it is important to consider the impact of the number of users, n, on the competitive ratio, i.e., on the efficiency of prediction-based pricing. This does not show up explicitly in Theorem 2, but it is possible to extract the information via a slightly more detailed analysis.

Consider a simple case where all X_i are i.i.d. with mean $\mathbb{E}[X_i] = \alpha$ and variance $\mathbb{V}[X_i] = \sigma^2$. Then, the mean and variance of the random variable $X(n) := \sum_{i=1}^n X_i$ are given by:

$$\mathbb{E}[X(n)] = n\alpha \quad \text{and} \quad \mathbb{V}[X(n)] = n\sigma^2. \quad (14)$$

As n increases, the central limit theorem guarantees that $\frac{X(n)-n\alpha}{\sqrt{n}\sigma}$ tends to a Gaussian random variable with zero mean and unit variance. Hence, informally, $X(n)$ tends to a Gaussian random variable with its mean and variance growing linearly in n as in (14).

Note, however, that (14) only imposes conditions on the first two moments of $X(n)$ and does not require $X(n)$ to be Gaussian nor their distributions to depend on just the first two moments. To highlight the dependence on n, let $G_n, g_n, p^*(n), \hat{p}(n), X(n), \hat{X}(n)$ etc. denote the corresponding quantities when there are n users. Then, we can prove the following corollary of Theorem 2, which shows that the competitive ratio exceeds 1 by an amount upper bounded by the normalized variance $q\sigma^2/\alpha$.

Corollary 4. *Suppose the first two moments of* $X(n)$ *are given by* (14)*. Under the conditions of Theorem 2, the bound on the competitive ratio is increasing in* n*. Moreover*

$$\frac{\mathbb{E}[G_n(\hat{p}(n))]}{\mathbb{E}[G_n(p^*(n))]} \leq 1 + \frac{q^2\alpha^2}{\frac{q\alpha^3}{\sigma^2} + \left(\frac{\alpha^2}{\sigma^2} + q\alpha\right)/n}$$

$$\to 1 + \frac{q\sigma^2}{\alpha} \text{ as } n \to \infty.$$

Note that the competitive ratio increases as the number of users increases. That is because the cost $h(\cdot)$ is based on the sum, not mean, of the users' elasticities. A system with a small number of users is identical to a system with a larger number of users in which some are entirely inelastic, which has lower uncertainty than the large system in which all users have random elasticity.

However, the analysis above should be taken with a grain of salt because, in practice, users are correlated. For example, on a hot day, many users will be more reluctant to turn their cooling systems off. We can illustrate the impact of such correlations with the following simple model.

$$X_i = \epsilon X_0 + X_i',$$

where X_i' are i.i.d. and independent of the common random variable X_0. In this case, given $\epsilon > 0$, $\mathbb{E}[X] = \Theta(n)$, $\mathbb{V}[X] = \Theta(n^2)$, so $\mathbb{C}^2[X] = \Theta(1)$, and

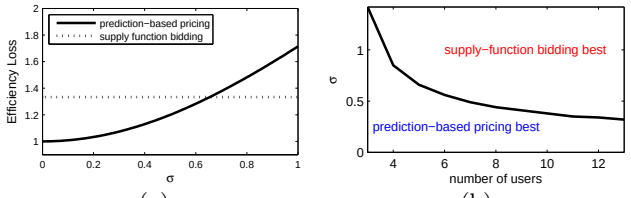
$$\frac{\mathbb{E}[G_n(\hat{p}(n))]}{\mathbb{E}[G_n(p^*(n))]} = \Theta(n).$$

This highlights that correlation among users can magnify the impact of prediction errors compared to the uncorrelated case, which has a negative impact on the performance of prediction-based pricing.

Such effects are not too worrying in the case of data center demand response, since it is unlikely for there to be a large number of data centers on any given distribution network. However, we have included the discussion in order to highlight a danger of using prediction-based pricing in other demand response contexts.

4.3 Prediction-based pricing versus supply function bidding

The previous results highlight that if predictions are accurate, then prediction-based pricing can be an effective



(a) (b)

Figure 10: Comparison of prediction-based pricing and supply function bidding demand response programs. (a) shows the efficiency loss as a function of the prediction error with $n = 5$. (b) shows the prediction error at which prediction-based pricing begins to have worse efficiency than supply function bidding for each n.

market design; however, if predictions are poor the market is highly inefficient. We now contrast the efficiency of prediction-based pricing with the supply function bidding approach discussed in Section 3.

Recall that previous work has concluded that supply function bidding is an efficient market design when agents have limited market power [29, 50]. Thus, which design is appropriate depends on the degree to which participants have market power and the accuracy of the predictions of supply functions made by the LSE.

To concretely illustrate the comparison between these two approaches, we contrast the competitive ratio derived above with the parallel results in [50]. Formally, Theorem 5.1 in [50] bounds the efficiency loss from strategic behavior of customers, i.e., price of anarchy (PoA), by $1 + \frac{\mathbf{min}\{D_m, D\}}{-D + \sum_{i \neq m} D_i}$, where D_i is the exogenous limit on consumer iÕs load reduction and D_m is the largest one, i.e., $m \in \mathbf{argmax}\{D_i\}$. This result is tight when the number of customers is no smaller than 2. Therefore, if there is only two large customers such like data centers or one large customer and some small customers considered together, then the efficiency loss can be very high. Generally, the loss decreases when more customers enter the market.[3]

The results of the comparison are shown in Figure 10. Specifically, Figure 10(a) shows the efficiency loss of both prediction-based pricing and supply function bidding. The impact of prediction error (in terms of the standard deviation σ of X_i when fixing $\mathbb{E}[X_i] = 1$) can be seen in the figure, where we assume the prediction errors of customers are independent. In particular, the figure highlights that the efficiency loss increases as the prediction error increase. When the number of users is small (5 in the figure), and thus market power is an issue, even with large prediction error (up to 60%), the prediction-based approach can still provide better performance than supply function bidding.

Figure 10(b) shows how this changes as the number of users grows, and thus market power becomes less of an issue. In particular, the figure shows the standard deviation threshold where prediction-based pricing becomes worse than supply function biding. Naturally, this threshold decreases as the number of users increases. However, even with 10 users, prediction-based pricing tolerates more than 30% prediction error before providing worse efficiency than supply function bidding. This emphasizes that prediction-based pricing is an appealing approach for demand response since it is unlikely to have more than a few data centers on a given distribution circuit.

[3]When this is only one customer, the approach in [50] does apply. Roughly speaking, in this case, the customer is a monopoly, so it can force the utility company to pay as much as possible if meeting the total demand reduction is enforced.

5. INCORPORATING NETWORK CONSTRAINTS

The previous section introduces prediction-based pricing in a context without a power network. In that context, the results highlight that prediction-based pricing is an appealing approach for data center demand response, since the efficiency of the mechanism is robust to errors in prediction as long as there are not a large number of correlated agents. In this section, our goal is to add an additional degree of realism to the model, power network constraints, and to investigate how these constraints impact the performance of prediction-based pricing.

5.1 Modeling the network

The setting we consider in this section is the same as in Section 4, except for the addition of network constraints. Typically, when electricity market issues like demand response are considered, the network constraints are either ignored entirely or a linear approximation, termed the "DC model," is used. See [44] for an introduction. However, due to our focus on reducing voltage violations with data center demand response, the DC model is not appropriate; it assumes the voltages at all buses are fixed at the reference value, which is seldom true in distribution networks.

As a result, we adopt a different model, called the "branch flow" model, which is commonly used for modeling distribution systems, e.g., [5, 11]. This model still uses a linear approximation of the power constraints, but now voltage variations are allowed at all buses except the root bus.

The branch flow model is defined as follows. The power network is represented by a directed, connected tree $\mathcal{G} = (N, E)$, where each node in $N := \{0, 1, ..., n\}$ represent a bus with the root at bus 0, each edge in E represents a line. Denote an edge by (i, j) or $i \to j$ if it points from bus i to bus j. The orientation of edges is fixed to be from the root to the leaves for $\mathcal{G}$.

For each edge $(i, j) \in E$, let $z_{ij} := r_{ij} + \mathbf{i}x_{ij}$ be the complex impedance on the line, and let $S_{ij} := P_{ij} + \mathbf{i}Q_{ij}$ be the sending-end complex power from bus i to bus j. This is the same as the receiving end power since lines are assumed to be lossless.

Let $s_j = P_j + \mathbf{i}Q_j$ be the complex net load (load minus generation) on bus j. Here P_j is the real power consumption, which can be further written as $P_j^0 - s_j(p)$, where P_j^0 is the real power consumption without demand response and $s_j(p)$ is the demand reduction given price p. Under our model, $s_i(p) = X_i p, \forall i$. Q_j is the reactive power consumption on bus j and we assume $Q_j = \beta_j P_j, \forall j$. The branch flow model is defined by the following set of power flow equations.

$$S_{ij} - s_j = \sum_{k:j \to k} S_{jk}, \forall j, \tag{15}$$

$$v_i - v_j = 2\mathbf{Re}(z_{ij}^* S_{ij}), \forall i, j, \tag{16}$$

where $\mathbf{Re}(\cdot)$ is the real part of a given complex number. Here (15) balances the power on each bus, and (16) characterizes the voltages across line (i, j) according to Ohm's law.

The constraint for the voltage on each bus is

$$\underline{v}_i \leq v_i \leq \overline{v}_i, \forall i. \tag{17}$$

5.2 Prediction-based pricing in networks

The incorporation of the network has a significant consequence for the design of prediction-based pricing. Due to the randomness of the cost functions, it is impossible for the voltage constraints to be always satisfied. This motivates a chance constraint where the goal of the LSE when setting

price $\hat{p}$ is now

$$\begin{aligned}\mathbb{E}\left[G(\hat{p})\right] = \min_{p} \quad & \mathbb{E}\left[G(p)\right] \\ \text{s.t.} \quad & p \geq 0 \\ & \mathbf{P}\{\text{voltage violation}|p\} \leq \epsilon.\end{aligned} \tag{18}$$

To determine more concretely what the set of feasible prices is for the chance constraint above, we first need to transform the power network constraints into constraints on feasible prices. To accomplish this, note that (15) gives that $S_{ij} = \sum_{k\in T_j} s_k$, where T_j is the tree rooted at bus j (including bus j). Then, we can rewrite (16) as

$$\begin{aligned}v_i - v_j &= 2\mathbf{Re}(z_{ij}^* S_{ij}) \\ &= 2\mathbf{Re}\left((r_{ij} - \mathbf{i}x_{ij})\sum_{k\in T_j} s_k\right) \\ &= 2\left(r_{ij}\sum_{k\in T_j} P_k + x_{ij}\sum_{k\in T_j} Q_k\right) \\ &= 2\left(r_{ij}\sum_{k\in T_j}\left(P_k^0 - X_k p\right) + x_{ij}\sum_{k\in T_j}\beta_k\left(P_k^0 - X_k p\right)\right) \\ &= 2\sum_{k\in T_j}\left(r_{ij} + x_{ij}\beta_k\right)P_k^0 - 2\sum_{k\in T_j}\left(r_{ij} + x_{ij}\beta_k\right)X_k p \\ &:= M_{ij} - N_{ij}p.\end{aligned}$$

Note that M_{ij} is a constant here, while N_{ij} is a random variable due to the uncertainties in X_k.

Next, assuming that E_k is the set of edges from root to bus k, we have (using $v_0 = 1$)

$$\begin{aligned}v_k &= 1 - \sum_{(i,j)\in E_k}\left(M_{ij} - N_{ij}p\right) \\ &= 1 - \sum_{(i,j)\in E_k} M_{ij} + \sum_{(i,j)\in E_k} N_{ij}p.\end{aligned}$$

Therefore $\underline{v}_k \leq v_k \leq \overline{v}_k$ becomes

$$\underline{v}_k \leq 1 - \sum_{(i,j)\in E_k} M_{ij} + \sum_{(i,j)\in E_k} N_{ij}p \leq \overline{v}_k,$$

which further implies

$$\frac{\underline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij}}{\sum_{(i,j)\in E_k} N_{ij}} \leq p \leq \frac{\overline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij}}{\sum_{(i,j)\in E_k} N_{ij}}.$$

This condition should hold for all buses, and therefore the feasible set is

$$\max_k \frac{\underline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij}}{\sum_{(i,j)\in E_k} N_{ij}} \leq p \leq \min_k \frac{\overline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij}}{\sum_{(i,j)\in E_k} N_{ij}}. \tag{19}$$

We can simplify the feasible set further by assuming that the voltage constraints (17) are satisfied when there is no demand response, i.e.,

$$\underline{v}_k \leq 1 - \sum_{(i,j)\in E_k} M_{ij} \leq \overline{v}_k, \forall k. \tag{20}$$

This implies that the feasible range in (19) is nonempty.

Additionally, since we only consider demand reduction with $p \geq 0$,[4] and $\underline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij} \leq 0, \forall k$, and we assume $X_k \geq 0$, we can further simplify the feasible set to

$$p \leq \min_k \frac{\overline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij}}{\sum_{(i,j)\in E_k} N_{ij}}. \tag{21}$$

Again, recall that N_{ij} is random. Therefore, the constraint above is on realizations. Importantly, for each realization, the constraints are linear, and therefore we can translate the constraints into a bound on the fraction of violation for each bus as follows.

$$\mathbf{P}\left\{\sum_{(i,j)\in E_k} N_{ij}p \geq \overline{v}_k - 1 + \sum_{(i,j)\in E_k} M_{ij}\right\} \leq \epsilon, \forall k. \tag{22}$$

The above equation can be viewed as a concrete specialization of the voltage violation constraint in (18). Note that it has a number of interesting properties. In particular, the violation probability is a strictly increasing function of p that equals 0 when $p = 0$ and approaches $\mathbf{P}\left\{\sum_{(i,j)\in E_k} N_{ij} > 0\right\}$ as $p \to \infty$. Therefore, if $\mathbf{P}\left\{\sum_{(i,j)\in E_k} N_{ij} > 0\right\}$ is smaller than ϵ, the chance constraint is satisfied for all $p \geq 0$[5]. Otherwise there is a threshold p_ϵ at which point the violation probability exceeds ϵ. In this case, the feasible pricing space is $[0, p_\epsilon]$, and the optimizing price becomes the projection of the unconstrained price derived in Section 4 onto this interval.

5.3 The efficiency of prediction-based pricing in networks

The previous analysis highlights that the necessary adjustment in the price used by the LSE due to network constraints can be achieved via a projection onto a feasible space of prices, which we have characterized in (22). The goal of this section is to understand the impact of network constraints, i.e., the projection into the feasible space of prices, have on the efficiency of the resulting price.

The main message of what follows is that network effects do not reduce the efficiency of prediction-based pricing, when efficiency is measured by the competitive ratio.

In particular, let us compare our algorithm with the clairvoyant algorithm that uses the same feasible set $[0, p_\epsilon]$ for each realization. This makes the offline algorithm weaker than the one considered in Section 4, i.e., the performance is strictly worse.

Recall that we denote by $G(\hat{p})$ and $G(p^*(\omega))$ the cost of our algorithm and the clairvoyant algorithm in Section 4 where network constraints are not considered. Let us now denote by $G(\hat{p}_\epsilon)$ and $G(p_\epsilon^*(\omega))$ the cost of our algorithm and the clairvoyant algorithm with the same feasible set $[0, p_\epsilon]$, defined as a function of the network constraints.

Our goal is to compare the competitive ratio without network constraints, i.e., $\frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*(\omega))]}$, to the competitive ratio under network constraints, i.e., $\frac{\mathbb{E}[G(\hat{p_\epsilon})]}{\mathbb{E}\left[G(p_\epsilon^*(\omega))\right]}$.

The following theorem highlights that constraints on the pricing space actually reduce the efficiency loss from uncertainty, and so the competitive ratio of prediction-based pricing remains unchanged when network constraints are considered. In the statement, we consider the feasible price set $R := [\underline{p}, \overline{p}]$ and denote by $g(\hat{p}_R)$ and $g(p_R^*)$ the cost of our algorithm and the clairvoyant algorithm with the same feasible set for a convex function $g(\cdot)$, e.g., a realization of the random function $G(\cdot)$. Proof is given in the appendix.

Theorem 5. *Consider any positive, convex function $g(\cdot)$ that is a realization of the random function $G(\cdot)$ and any*

[4]Note that all the results here can be easily extended to the case where we allow p to be negative.

[5]This does not happen in our case because we assume X_i's are positive, therefore $\mathbf{P}\left\{\sum_{(i,j)\in E_k} N_{ij} > 0\right\} = 1$

non-empty feasible set $R := [\underline{p}, \overline{p}]$. *Then,*

$$\frac{g(\hat{p})}{g(p^*)} \geq \frac{g(\hat{p}_R)}{g(p^*_R)}, \tag{23}$$

and thus

$$\frac{\mathbb{E}\left[G(\hat{p})\right]}{\mathbb{E}\left[G(p^*(\omega))\right]} \geq \frac{\mathbb{E}\left[G(\hat{p}_\epsilon)\right]}{\mathbb{E}\left[G(p^*_\epsilon(\omega))\right]}. \tag{24}$$

A key distinction between this theorem and Theorem 2 is that the feasible price set of both the optimal and the algorithm are fixed to $R := [\underline{p}, \overline{p}]$. This implies that we are not comparing with the "true" offline optimal, which may have different feasible sets for the price for different realizations. Instead, we are comparing with the weaker offline optimal that, because of uncertainty, optimizes over the same feasible price set as our online algorithm, but then has the foresight necessary to choose optimally given these price constraints. This is a common choice for comparison when studying the competitive ratio of online algorithms in situations where clairvoyance yields different feasible action spaces.

6. CONCLUDING REMARKS

In this paper we have highlighted two main points. First, that data center demand response has significant potential and, second, that prediction-based pricing is an appealing mechanism with which to extract this potential.

More concretely, we have illustrated that, not only are data centers large loads to target with demand response programs, they can provide nearly the same degree of flexibility for LSEs as large-scale storage if properly incentivized. However, this last caveat is crucial – it is much harder to extract flexibility from data centers than from storage.

To that end, this paper has argued that prediction-based pricing is a promising market design for this context. While, in general, prediction-based pricing may be less efficient than supply function bidding (due to prediction errors), because data centers typically have significant market power on their distribution networks, supply function bidding can be very inefficient whereas prediction-based pricing is less influenced.

In particular, the analytic results in Sections 4 and 5 highlight that the efficiency of prediction-based pricing is favorable to that of supply function bidding when market power is an issue – even when predictions are error prone. These analytic results are the first, to our knowledge, that provide bounds on the competitive ratio of prediction-based pricing programs, and also the first to provide an analysis of prediction-based pricing programs in a context where network constraints are considered.

However, much work still remains before prediction-based pricing can be used in practice. In particular, in this paper we have adopted quadratic objectives, and it is important to understand the impact of this. For example, in the context of internet congestion management, [36] has studied the impact of convexity of costs on the contrast between time-of-use pricing and fixed-budget rebates. A similar study in the context of predictive pricing and supply function bidding is crucial.

Further, it is important to do an empirical study to understand how predictable the response of data centers will be in demand response programs. Initial pilot studies along these lines are proceeding in some demand response markets, but these have yet to focus on data centers specifically. Depending on the result of such studies, it may be natural to consider hybrid mechanisms that combine predictions and bidding in order to extract supply function information from data centers.

Additionally, many practical aspects of prediction-based pricing programs still require careful thought. For example, what is the appropriate time-scale at which prices should be adjusted? The time-scale chosen allows for a balance between the responsiveness desired by the LSE and the risk-aversion of the data center. Further, in this paper we have assumed a scalar price. One could also investigate location dependent prices in distribution networks, similar to locational marginal prices (LMPs) for transmission networks. While these are not currently used, the extra geographical flexibility they provide could be valuable. Finally, there are interesting exploration-exploitation tradeoffs that come up when setting prices in prediction-based pricing programs. We have not addressed this issue in this paper due to the complexities of the power network, but work in the operations research community has begun to study this in other contexts [6, 7], using "regret" as the performance measure. It would be interesting for future work to incorporate this issue into the demand response context.

Acknowledgment

This work was supported by NSF grants CCF 0830511, CNS 0911041, and CNS 0846025, DoE grant DE-EE0002890, ARO MURI grant W911NF-08-1-0233, Microsoft Research, Bell Labs, the Lee Center for Advanced Networking, and ARC grant FT0991594. We are grateful to Qiuyu Peng and Yunjian Xu for the helpful discussion. We also thank the anonymous reviewers and our shepherd, Patrick Loiseau, for their valuable comments and help.

7. REFERENCES

[1] Report to congress on server and data center energy efficiency. 2007.

[2] D. J. Aigner and J. G. Hirschberg. Commercial/industrial customer response to time-of-use electricity prices: Some experimental results. *The RAND Journal of Economics*, 16(3):341–355, 1985.

[3] B. Aksanli, J. Venkatesh, L. Zhang, and T. Rosing. Utilizing green energy prediction to schedule mixed batch and service jobs in data centers. *ACM SIGOPS Operating Systems Review*, 45(3):53–57, 2012.

[4] B. Allaz and J.-L. Vila. Cournot competition, forward markets and efficiency. *Journal of Economic theory*, 59(1):1–16, 1993.

[5] M. Baran and F. F. Wu. Optimal sizing of capacitors placed on a radial distribution system. *IEEE Transactions on Power Delivery*, 4(1):735–743, 1989.

[6] O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

[7] O. Besbes and A. Zeevi. On the minimax complexity of pricing in a changing environment. *Operations research*, 59(1):66–79, 2011.

[8] D. W. Cai and A. Wierman. Inefficiency in forward markets with supply friction. In *Proc. of CDC*, 2013.

[9] L. Chen, N. Li, S. H. Low, and J. C. Doyle. Two market models for demand response in power networks. In *IEEE SmartGridComm*, pages 397–402, 2010.

[10] Y. Chen, D. Gmach, C. Hyser, Z. Wang, C. Bash, C. Hoover, and S. Singhal. Integrated management of application performance, power and cooling in data centers. In *Proc. of NOMS*, 2010.

[11] H.-D. Chiang and M. E. Baran. On the existence and uniqueness of load flow solution for radial distribution power networks. *IEEE Transactions on Circuits and Systems*, 37(3):410–416, 1990.

[12] A. J. Conejo, J. M. Morales, and L. Baringo. Real-time demand response model. *IEEE Transactions on Smart Grid*, 1(3):236–242, 2010.

[13] Department of Energy. The smart grid: An introduction.

[14] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy. Inverter var control for distribution systems with renewables. In *IEEE SmartGridComm*, pages 457–462, 2011.

[15] M. Farivar, R. Neal, C. Clarke, and S. Low. Optimal inverter var control in distribution systems with high pv penetration. In *IEEE Power and Energy Society General Meeting*, pages 1–7, 2012.

[16] L. Gan, U. Topcu, and S. H. Low. Optimal decentralized protocol for electric vehicle charging. In *IEEE CDC*, pages 5798–5804, 2011.

[17] L. Gan, A. Wierman, U. Topcu, N. Chen, and S. H. Low. Real-time deferrable load control: handling the uncertainties of renewable generation. In *Proc. of ACM eEnergy*, pages 113–124, 2013.

[18] A. Gandhi, Y. Chen, D. Gmach, M. Arlitt, and M. Marwah. Minimizing data center SLA violations and power consumption via hybrid resource provisioning. In *Proc. of IGCC*, 2011.

[19] N. Gast, J.-Y. Le Boudec, A. Proutière, and D.-C. Tomozei. Impact of storage on the efficiency and prices in real-time electricity markets. In *Proceedings of the fourth international conference on Future energy systems*, pages 15–26. ACM, 2013.
[20] G. Ghatikar, V. Ganti, N. Matson, and M. Piette. Demand response opportunities and enabling technologies for data centers: Findings from field studies. 2012.
[21] J. Heo, P. Jayachandran, I. Shin, D. Wang, T. Abdelzaher, and X. Liu. Optituner: On performance composition and server farm energy minimization application. *Parallel and Distributed Systems, IEEE Transactions on*, 22(11):1871–1878, 2011.
[22] Y.-Y. Hsu and C.-C. Su. Dispatch of direct load control using dynamic programming. *IEEE Transactions on Power Systems*, 6(3):1056–1061, 1991.
[23] https://www.sce.com/wps/portal/home/regulatory/load-profiles.
[24] http://www.apple.com/environment/renewable-energy.
[25] http://www.google.com/green/energy/.
[26] http://www.nrel.gov/midc/lmu/.
[27] L. Huang, J. Walrand, and K. Ramchandran. Optimal demand response with energy storage management. In *IEEE SmartGridComm*, pages 61–66, 2012.
[28] D. Irwin, N. Sharma, and P. Shenoy. Towards continuous policy-driven demand response in data centers. *Computer Communication Review*, 41(4), 2011.
[29] R. Johari and J. N. Tsitsiklis. Parameterized supply function bidding: Equilibrium and efficiency. *Operations research*, 59(5):1079–1089, 2011.
[30] J. Koomey. Growth in data center electricity use 2005 to 2010. *Oakland, CA: Analytics Press. August*, 1:2010, 2011.
[31] J.-Y. Le Boudec and D.-C. Tomozei. A demand-response calculus with perfect batteries. In *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*, pages 273–287. Springer Berlin Heidelberg, 2012.
[32] N. Li, L. Chen, and S. H. Low. Optimal demand response based on utility maximization in power networks. In *IEEE Power and Energy Society General Meeting*, pages 1–8, 2011.
[33] M. Lin, Z. Liu, A. Wierman, and L. Andrew. Online algorithms for geographical load balancing. In *Proc. of IGCC*, 2012.
[34] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In *Proc. of INFOCOM*, 2011.
[35] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser. Renewable and cooling aware workload management for sustainable data centers. In *Proc. of ACM Sigmetrics*, 2012.
[36] P. Loiseau, G. Schwartz, J. Musacchio, S. Amin, and S. S. Sastry. Incentive mechanisms for internet congestion management: Fixed-budget rebate versus time-of-day pricing. 2013.
[37] Z. Ma, D. Callaway, and I. Hiskens. Decentralized charging control for large populations of plug-in electric vehicles. In *IEEE CDC*, pages 206–212, 2010.
[38] A. H. Mahmud and S. Ren. Online capacity provisioning for carbon-neutral data center with demand-responsive electricity prices. *ACM SIGMETRICS Performance Evaluation Review*, 41(2):26–37, 2013.
[39] D. Meisner, C. Sadler, L. Barroso, W. Weber, and T. Wenisch. Power management of online data-intensive services. In *Proc. of ISCA*, 2011.
[40] A.-H. Mohsenian-Rad and A. Leon-Garcia. Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Transactions on Smart Grid*, 1(2):120–133, 2010.
[41] F. Murphy and Y. Smeers. On the impact of forward markets on investments in oligopolistic markets with reference to electricity. *Operations research*, 58(3):515–528, 2010.
[42] National Institute of Standards and Technology. NIST framework and roadmap for smart grid interoperability standards. NIST Special Publication 1108, 2010.
[43] NY Times. Power, Pollution and the Internet.
[44] N. S. Rau. *Optimization principles: practical applications to the operation and markets of the electric power industry*. John Wiley & Sons, Inc., 2003.
[45] B. Urgaonkar, G. Kesidis, U. V. Shanbhag, and C. Wang. Pricing of service in clouds: Optimal response and strategic interactions. 2013.
[46] R. Urgaonkar, B. Urgaonkar, M. Neely, and A. Sivasubramaniam. Optimal power cost management using stored energy in data centers. In *Proc. of the ACM Sigmetrics*, 2011.
[47] P. Van de Ven, N. Hegde, L. Massoulie, and T. Salonidis. Optimal control of residential energy storage under price fluctuations. In *Proc. of ENERGY*, pages 159–162, 2011.
[48] www.fcgov.com/utilities/business/rates/electric/coincident-peak.
[49] H. Xu and B. Li. Cost efficient datacenter selection for cloud services. In *Proc. of ICCC*, 2012.
[50] Y. Xu, N. Li, and S. Low. Demand response with parameterized supply function bidding.
[51] J. Yao, S. S. Oren, and I. Adler. Two-settlement electricity markets with price caps and cournot generation firms. *European Journal of Operational Research*, 181(3):1279–1296, 2007.
[52] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely. Data centers power reduction: A two time scale approach for delay tolerant workloads. In *Proc. of INFOCOM*, pages 1431–1439, 2012.
[53] Q. Zhang, M. Zhani, Q. Zhu, S. Zhang, R. Boutaba, and J. Hellerstein. Dynamic energy-aware capacity provisioning for cloud computing environments. In *ICAC*, 2012.
[54] R. D. Zimmerman, C. E. Murillo-Sánchez, and D. Gan. A matlab power system simulation package, 2005.

APPENDIX

A. APPENDIX: PROOFS

Proof of Theorem 1

To begin, we compute as follows:

$$\begin{aligned} g'(p) &= -h'\left(D-\sum_i s_i(p)\right)\sum_i s_i'(p)+\sum_i c_i'(s_i(p))s_i'(p) \\ &= \sum_i s_i'(p)\left(c_i'(s_i(p))-h'\left(D-\sum_i s_i(p)\right)\right) \\ &= \left(p-h'\left(D-\sum_i s_i(p)\right)\right)\sum_i s_i'(p), \end{aligned}$$

where the last equality follows from $c_i'(s_i(p))=p$ for all i. Our assumptions imply that $s_i'(p)=\left(c_i''(s_i(p))\right)^{-1}>0$, and hence $g'(p)=0$ if and only if

$$v(p) := p-h'\left(D-\sum_i s_i(p)\right)=0.$$

Now, $v(0)=-h'(D)\le 0$, $v(\overline{p})=\overline{p}>0$, and $v(p)$ is strictly increasing. Hence a unique $0\le p^*<\overline{p}$ satisfies $v(p^*)=0$. Moreover $g'(p)<0$ for $p<p^*$ and $g'(p)>0$ for $p>p^*$ implying that p^* is the unique minimizer of $g(p)$. □

Proof of Theorem 2

We first evaluate $\mathbb{E}[G(p^*)]$. From Theorem 1

$$\begin{aligned} p^* &= h'\left(D-\sum_i s_i(p^*)\right) \\ &= q\left(D-p^*\sum_i X_i\right) \\ &= qD-qXp^*. \end{aligned}$$

Hence

$$p^* = \frac{q}{1+qX}D, \tag{25}$$

which is a random (optimal) price.

Next, from (10) and (25) we have

$$\mathbb{E}[G(p^*)] = \frac{qD^2}{2}\mathbb{E}\left[\frac{1}{1+qX}\right], \tag{26}$$

where the expectation is taken over X.

To evaluate (7) we have, using (10),

$$\begin{aligned} \mathbb{E}[G(\hat{p})] &= \min_{p\ge 0}\mathbb{E}\left[\frac{q}{2}(D-Xp)^2+\frac{1}{2}Xp^2\right] \\ &= \min_{p\ge 0}\frac{1}{2}\left(\left(q\mathbb{E}\left[X^2\right]+\mathbb{E}[X]\right)p^2-2qD\mathbb{E}[X]\,p+qD^2\right). \end{aligned}$$

Consequently, the unique minimizer $\hat{p}$ and the optimal value of (7) are

$$\hat{p} = \frac{q\mathbb{E}[X]}{q\mathbb{E}[X^2]+\mathbb{E}[X]}D, \quad (27)$$

$$\mathbb{E}[G(\hat{p})] = \frac{qD^2}{2}\frac{\mathbb{E}[X]+q\mathbb{V}[X]}{\mathbb{E}[X]+q\mathbb{E}[X^2]}. \quad (28)$$

We can now quantify the competitive ratio using (26) and (18). Jensen's inequality implies $\mathbb{E}[G(p^*)] \geq \frac{qD^2}{2}\frac{1}{1+q\mathbb{E}[X]}$. Thus,

$$\begin{aligned}\frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*)]} &\leq \frac{\mathbb{E}[X]+q\mathbb{V}[X]}{\mathbb{E}[X]+q\mathbb{E}[X^2]}(1+q\mathbb{E}[X]) \\ &= 1+\frac{q^2\mathbb{E}[X]}{\mathbb{E}[X]+q\mathbb{E}[X^2]}\mathbb{V}[X].\end{aligned}$$

Rewriting the above in terms of the square coefficient of variation $\mathbb{C}^2[X]$ gives:

$$\frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*)]} \leq 1+\frac{(q\mathbb{E}[X])^2\mathbb{C}^2[X]}{1+(q\mathbb{E}[X])(\mathbb{C}^2[X]+1)}.$$

Finally, to compare $\hat{p}$ in (27) with p^* in (25) we can rewrite $\hat{p}$ as

$$\hat{p} = \frac{qD}{1+q\mathbb{E}[X](\mathbb{C}^2[X]+1)}.$$

Hence

$$\begin{aligned}\mathbb{E}[p^*] &= \mathbb{E}\left[\frac{qD}{1+qX}\right] \\ &\geq \frac{qD}{1+q\mathbb{E}[X]} \geq \frac{qD}{1+q\mathbb{E}[X](\mathbb{C}^2[X]+1)} = \hat{p},\end{aligned}$$

where the first inequality follows from the Jensen's inequality and the second inequality follows from $\mathbb{C}^2[X] \geq 0$. Both of these are equalities if and only if X has zero variance. □

Proof of Theorem 3

To show tightness we focus on the only inequality used in the proof of Theorem 2, which is

$$\mathbb{E}[G(p^*)] \geq \frac{qD^2}{2(1+\mathbb{E}[X])}.$$

We need to show that, for any $\epsilon > 0$, there exists a probability distribution $f(X)$ with mean $\mathbb{E}[X]$ and variance $\mathbb{V}[X]$ such that

$$\mathbb{E}[G(p^*)] \leq \frac{qD^2}{2(1+\mathbb{E}[X])} + \epsilon.$$

We define such a probability distribution as follows. For any $0 < x < 1$, let $d_1 := \mathbb{E}[X] - \sqrt{\mathbb{V}[X](1-x)/x}$ and $d_2 := \mathbb{E}[X] + \sqrt{\mathbb{V}[X]x/(1-x)}$. Then define the following probability density function:

$$f_x(X) = x\delta(\mathbb{E}[X]-d_1) + (1-x)\delta(\mathbb{E}[X]-d_2), \quad (29)$$

where

$$\delta(a) := \begin{cases}\infty & \text{if } a = 0 \\ 0 & \text{otherwise}\end{cases}$$

and $\int \delta(a)da = 1$.

Note that for any $0 < x < 1$, the probability distribution defined in (29) has mean $\mathbb{E}[X]$ and variance $\mathbb{V}[X]$ and

$$\lim_{x\to 1}\mathbb{E}[G(p^*)] = \frac{qD^2}{2(1+\mathbb{E}[X])}.$$

Thus, the bound is tight. □

Proof of Corollary 4

Given $\mathbb{E}[X(n)] = n\alpha$ and $\mathbb{V}[X(n)] = n\sigma^2$, we have $\mathbb{C}^2[X] = \frac{\sigma^2}{n\alpha^2}$. Thus, we can compute as follows.

$$\begin{aligned}\frac{\mathbb{E}[G(\hat{p})]}{\mathbb{E}[G(p^*)]} &\leq 1+\frac{(q\mathbb{E}[X])^2\mathbb{C}^2[X]}{1+(q\mathbb{E}[X])(\mathbb{C}^2[X]+1)} \\ &= 1+\frac{q^2n^2\alpha^2\frac{\sigma^2}{n\alpha^2}}{1+qn\alpha\left(\frac{\sigma^2}{n\alpha^2}+1\right)} \\ &= 1+\frac{q^2\alpha^2}{\frac{q\alpha^3}{\sigma^2}+\left(\frac{\alpha^2}{\sigma^2}+q\alpha\right)/n} \\ &\to 1+\frac{q\sigma^2}{\alpha} \ \ as \ \ n\to\infty.\end{aligned}$$

□

Proof of Theorem 5

To prove that the competitive ratio of prediction-based pricing does not become larger when there are constraints on the space of prices, i.e., $p \in [\underline{p}, \overline{p}]$, we consider two cases. The cases are diagramed in Figure 11.

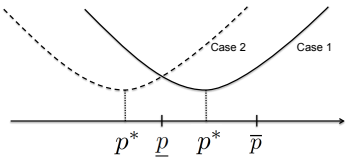


Figure 11: Diagram of cases for proof of Theorem 5.

Case 1: The price p^* picked by the clairvoyant algorithm is within the feasible set $[\underline{p}, \overline{p}]$, i.e., $p^* \in [\underline{p}, \overline{p}]$. We have $p^*_R = p^*$ and therefore $g(p^*_R) = g(p^*)$. If the price picked by our algorithm $\hat{p} \in [\underline{p}, \overline{p}]$, then we have $\hat{p}_R = \hat{p}$ and therefore $g(\hat{p}_R) = g(\hat{p})$. Hence $\frac{g(\hat{p})}{g(p^*)} = \frac{g(\hat{p}_R)}{g(p^*_R)}$.

Otherwise $\hat{p} \notin [\underline{p}, \overline{p}]$. We have $\hat{p}_R = \underline{p}$ if $\hat{p} < \underline{p}$ and $\hat{p}_R = \overline{p}$ if $\hat{p} > \overline{p}$. In either case $g(\hat{p}_R) \leq g(\hat{p})$, and therefore $\frac{g(\hat{p})}{g(p^*)} \geq \frac{g(\hat{p}_R)}{g(p^*_R)}$.

Case 2: The price p^* picked by th clairvoyant algorithm is outside the feasible set $[\underline{p}, \overline{p}]$. Without loss of generality, we assume $p^* < \underline{p}$, as shown in the figure. We have $p^*_R = \underline{p}$ and $g(p^*_R) \geq g(p^*)$. If the price picked by our algorithm $\hat{p} \in [\underline{p}, \overline{p}]$, then we have $\hat{p}_R = \hat{p}$ and therefore $g(\hat{p}_R) = g(\hat{p})$. Hence $\frac{g(\hat{p})}{g(p^*)} \geq \frac{g(\hat{p}_R)}{g(p^*_R)}$.

Otherwise $\hat{p} \notin [\underline{p}, \overline{p}]$. We have $\hat{p}_R = \underline{p}$ if $\hat{p} < \underline{p}$ and $\hat{p}_R = \overline{p}$ if $\hat{p} > \overline{p}$. In the first case we have $\hat{p}_R = p^*_R = \underline{p}$ and therefore $g(\hat{p}_R) = g(p^*_R)$, hence $\frac{g(\hat{p})}{g(p^*)} \geq \frac{g(\hat{p}_R)}{g(p^*_R)} = 1$. In the second case we have $g(\hat{p}_R) \leq g(\hat{p})$, and therefore $\frac{g(\hat{p})}{g(p^*)} \geq \frac{g(\hat{p}_R)}{g(p^*_R)}$. □