Electricity Demand Response under Real-Time Pricing: A Multi-armed Bandit Game

Zibo Zhao* and Andrew L. Liu† and Yihsu Chen‡
* School of Industrial Engineering, Purdue University, West Lafayette, IN 47907. USA.
E-mail: zhao438@purdue.edu
† School of Industrial Engineering, Purdue University, West Lafayette, IN 47907. USA.
E-mail: andrewliu@purdue.edu. Corresponding author.
‡ Department of Technology & Information Management, University of California, Santa Cruz, CA 95064. USA.
E-mail: yihsuchen@ucsc.edu

***Abstract*—Real-time electricity pricing (RTP) for consumers has long been argued to be key to realize the many envisioned benefits of a smart energy grid. However, there has not been a consensus on how to best implement RTP in an organized, competitive wholesale market with active demand participation. Since most of such markets implement a two-settlement system, with day-ahead electricity price forecasts guiding physical transactions in the next day and real-time ex post prices settling any real-time imbalances, it is a natural idea to let consumers respond to the day-ahead prices. We show in this paper through simulation that naive responsive behaviors to the day-ahead price signals can lead to high price volatility, which will increase not only the risk of system instability, but also the financial risks faced by the consumers. To overcome this issue, we propose a game-theoretic framework in which each consumer solves a multi-armed bandit problem; that is, each consumer learns from the history of the game and attempts to minimize the cumulative regrets. We show through simulation that such a framework leads to drastically reduced volatility on real-time prices and much flatter load curves for the entire grid.**

I. Introduction

In traditional power systems, electricity demand is usually considered to be inflexible, as consumers have been used to the idea of consuming electricity whenever they need to. Since the reliable operation of a power system requires the supply and demand to be balanced at all time, demand inflexibility adds great pressures on the system to maintain enough redundancies in both generation and transmission capacities. Such redundancies are very costly, as they are capital intensive to build. In addition, the lack of flexibility on the demand side makes the power system less reliable and vulnerable to attacks, as the outage of a few large power plants and/or transmission lines may bring down a large part of the highly interconnected power grid (such as the U.S. Northeast blackout of 2003 [6]).

The visionary late MIT professor Fred Schweppe envisioned an energy future with real-time electricity pricing and actively engaged demand response back in 1978 [11], as he recognized the many benefits associated with flexible demand. With the advent of various smart grid technologies, such as smart meters and an array of information and communications technologies (ICTs), flexible demand is more than ever to be closer to reality. The direct benefits of flexible demand are huge, including saving a tremendous amount of money for consumers and making power systems more robust. There are also more subtle benefits, such as the potential environmental benefits of using flexible demand to better match outputs from renewable resources, such as wind and solar, and hence reducing air pollutant emissions from fossil fuel-fired power plants.

It is our belief that technologies alone, however, are not enough to seamlessly integrate flexible demand into a wholesale power market. There have to be changes to the current market operations, which can be on the side of system operators, utilities or individual consumers. Generally speaking, there are two fundamentally different approaches to bring demand flexibility: one is the centralized approach; the other is the decentralized approach. The former approach, as the name suggests, is to have the system operators or utilities directly manage their load. Various forms of such an approach already exist in the current system operations. For example, load shedding contracts have been around for many years. Such contracts provide the system operators the flexibility to cut off certain load during emergency situations. In return, the other side of the contracts, usually large industrial customers, will receive lower electricity rates in return. More recent example of centralized load control is to use smart household thermostats to reduce peak load [13].

While central load control may be effective, the amount of resources that system operators and utilities can control is limited, partially due to software and computing power limitation. In addition, many demand-side resources, such as microgrids and distributed generation resources (such as rooftop solar panels), are not under the control of system operators (unless certain contracts are in place). The centralized approach may also raise issues on privacy, as some consumers do not feel comfortable of having someone else manage their household's electricity usage.

The decentralized approach, on the other hand, depends on end-consumers to make their own electricity consumption decisions. Through certain incentives and information sharing, it is hoped that the collective consumers' actions may bring the desired demand flexibility from the system's perspective. This is commonly referred to as demand response (DR). Within the general term of DR, there are many different forms. Using the classification in [12], there are incentive-based DR and price-

based DR. For the former, it can be similar to the load shedding contract; namely, a consumer receives some incentives (such as lower rates) to promise to respond to utilities' call of reducing electricity consumption as needed. Other forms of incentive-based DR often involve a baseline; that is, consumers will receive some incentives if they can bring their energy consumption below a pre-defined baseline consumption level. Price-based DR, on the other hand, is usually completely voluntary (i.e., no contracts), and consumers alone make the decision on when to use electricity based on some electricity price information they receive from the system operators or utilities. The price information has to reflect to some degree of power system's conditions (such as high demand and low supply at a certain period). What exactly shall be contained in such price information vary greatly, ranging from the time-of-use pricing mechanism to real-time pricing (RTP). In the later, real-time wholesale electricity prices (such as hourly or half-hourly prices) are shared with the end consumers for them to decide their energy consumption. In this work, we focus on priced-based DR coupled with RTP; namely, consumers respond to real-time electricity prices. Such an approach may have the least resistance of being implemented as it is a completely decentralized approach that does not require any operation or rule changes on the system operator side.

There have been a large amount of works on how an individual consumer should make decisions under RTP. However, much fewer works exist to study the system-level impacts when a large amount of consumers respond to RTP. Roozbehai et al. [10] raise the concern that if not done properly, price-based DR plus RTP may increase price volatility and reduce system reliability. This is so since real-time electricity prices are only available after the actual supply and demand are realized. It is meaningless for consumers to respond to these ex-post prices. Hence, consumers can only respond to some price forecasts, which creates a closed-loop system with feedback, as the price forecasts will influence consumers' decisions, which in turn will impact the real-time electricity prices and likely will cause the real-time prices to diverge from the price forecasts. Any price-based DR implementation without considering such a closed-loop system is doomed to fail. While such a closed-loop system can be managed in a centralized approach (see [14], for example), there was a void in literature on how to realize a fully decentralized price-based DR without causing extreme price volatility or jeopardizing system reliability. To address such a void, the first two authors proposed a multi-armed bandit (MAB) game framework [15], in which each consumer solves a multi-armed bandit problem. The essence of this approach is that each consumer can learn from their past decisions based on the past electricity prices and their electric bills, and gradually arrive at a strategy that can minimize the consumers' regrets (of not making a better decision that would have lowered their electric bills). While preliminary simulation results have been shown in [15] (using test cases without transmission networks), in this work we focus on presenting new results based on a power system representing the New England region in the US, with capacity-constrained transmission lines. In addition to the similar effects of the MAB game approach as shown in [15], including volatility reducing and load-curve flattening, we show that transmission congestion costs can also be reduced in this completely decentralized approach (i.e., without the system operator to dispatch demand resources), even with exogenous uncertainty on wind plants' outputs. Consequently, social welfare is increased under the MAB game approach than in a naive response approach. In addition, we implemented another decentralized approach to implement DR based on the approach in [9], and show that the MAB game approach still compares favorably in all the considered measures, including price volatility, congestion costs, and social welfare.

To make this paper stand-alone, we will present the basic setup for the MAB game in Section II. We will also briefly describe the two other decentralized approaches to implement DR in the same section; namely, the naive-response approach and the approach introduced in [9]. Section III describes in detail the New England power system model and input data, and then presents the simulation results. Section IV discusses the limitations of the MAB game approach, and identifies several future research directions.

II. Models

A. General Market Setting

In a wholesale power market, electric power generators submit their bids to supply certain quantities of electricity at certain prices to an Independent System Operator (ISO), whose task is to dispatch electricity to match the demand[1] with the supply bids, while ensuring all physical constraints are met. The (wholesale) electricity price is the result of this supply and demand balancing, as illustrated in Fig. 1. The dispatch is done in a two-settlement fashion and is usually on an hourly basis. More specifically, at each day, the ISO solicits supply bids from power generators to meet the demand forecasts of each hour in the next day. This is the day-ahead (DA) market with the market clearing electricity prices referred to as the day-ahead price. In real time, the ISO matches any supply and demand deviations with additional generation resources. Such additional balancing produces the so-called real-time (RT) electricity prices. An ISO finds the optimal (i.e., cost minimizing) dispatch schedules through solving large-scale linear programming (or convex quadratic) problems, commonly referred to as the economic dispatch problem.

Here we assume that electricity consumers are charged the real-time electricity prices.[2] But since the real-time prices are determined after the fact, i.e, after the supply and demand have been realized (hence, the so-called ex-post prices), consumers cannot respond to the ex-post prices to determine how they

[1] As we focus on consumers' responses to real-time electricity prices, we do not consider active demand bidding into wholesale markets.

[2] On top of the wholesale rates, electric utility companies also impose additional charges to end users to cover the utilities' transmission and distribution (T&D) costs. But such charges are fixed; i.e., they do not vary over time. Hence, we do not consider any of the fixed charges to consumers in our models as such charges do not affect any of our research findings.

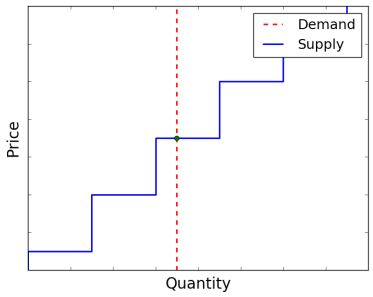

Figure 1: Wholesale electricity price as the market clearing price from supply bids and demand balancing.

would consume electricity in some future time. Hence, in the naive-response model (to be introduced in Section II-C), we assume that consumers can receive the day-ahead prices and then decide what to do the next day. In the MAB-game model, on the other hand, the consumers do not need to respond to any price forecasts, as their decisions are made through learning their past decisions and payoffs. Hence, the day-ahead prices (or any price forecasts) are irrelevant in the MAB-game model.

Some more details regarding the market setting are as follows.

Temporal resolution. Our simulations in both models contain repetitive daily market operations, with each day denoted by $d \in \{0, 1, ...\}$. Within each day, there are time periods denoted by t, and $t = \{1, 2, \ldots, T\}$. With each t there is a corresponding DA and RT price. Usually the t's are measured in hours or in even finer resolution in real markets.

Electricity demand. We consider a total of n end consumers, where the consumers are indexed by $k_c \in \{1, ..., n\}$. We assume that each of the consumers has both inflexible and flexible load. The consumption level and timing of the inflexible load is fixed. For the flexible load, its consumption level is also assumed to be fixed (across different days), which is denoted by θ^{k_c}. Hence the only decision each consumer needs to make is when to consume the flexible load to minimize their electric bills. Such a decision is denoted by $x_d^{k_c} \in \{1, ..., T\}$. An underlying assumption here is that a consumer only makes one decision in one day (hence, there is only a d index to the variable x, not a t index).

For the inflexible demand, there is no need to consider on the individual consumer level, and we let BD_t denote the aggregate inflexible load at time $t = \{1, \ldots, T\}$ in a day, and it does not change over days.

With the above notations, the system-wide real-time demand, denoted by $L_d^{RT}(t)$, is as follows for $t = 1, \ldots, T$ and $d = 1, 2, \ldots$:

$$L_d^{RT}(t) = BD_t + \sum_{k_c=1}^{n} \theta^{k_c} \mathbb{1}_{\{x_d^{k_c}=t\}}, \tag{1}$$

where $\mathbb{1}_{\{A\}}$ is the typical indicator function that takes the value of 1 if the generic event A is true, and 0 otherwise.

B. The MAB-game Model

Under real-time pricing, since each consumer's electric bill depends on how the other consumers respond to price signals, this is a classic situation of a non-cooperative game in the game theory literature. However, this is not a simple static game, as the decision of choosing which period to consume the flexible load needs to be made on a daily basis. Hence, this is an instance of dynamic games. In addition, this is also a game of incomplete information; that is, consumers do not know the explicit payoff functions of the other consumers; nor do they know how many players are in the game. In game theory literature, the standard equilibrium concept for dynamic games of incomplete information is Perfect Bayesian Nash equilibrium (PBNE) [2]. A PBNE consists the collection of each player's strategy profile, which is a function that maps the entire history of the game to each player's feasible set of actions, under the assumption that each player updates their beliefs of other players' payoff functions based on the Bayes' rule. As pointed out in [3], the requirements for PBNE are too strong to be practical: first, each player needs to choose a strategy profile that yields the best expected payoff (given other players choosing their corresponding PBNE strategy) over all possible histories of the game; second, all players need to update their beliefs' of other players' (unknown) payoff functions by the Bayes' rule through their observations in each time period. Not only such strategy profiles are not computable (as it would require to find the best mapping over the functional space of all possible mappings, leading to an infinite-dimension optimization problem), nor are electricity consumers in reality this sophisticated.

To avoid the technical difficulties associated with PBNE, we will have to relax the strong assumptions associated with the equilibrium concept. More specifically, we may want to relax the Bayes' updating assumption, and assume that the consumers do not necessarily require to find the best possible strategy, but a "good enough" strategy is sufficient. One way to quantify a "good enough" strategy is to use the concept of regrets; that is, to measure the cumulative differences between what would be the best response in a period and what the consumer chose based on a certain strategy (also referred to as a policy). Then the consumer may adopt a regret minimizing strategy for the sequential decision-making problem. With the market setting described in Subsection II-A, the regret-minimizing approach resembles the well-studied multi-armed bandit (MAB) problem. More specifically, on each day, each consumer decides which time period $t \in \{1, \ldots, T\}$ to consume their flexible load. This is like choosing an arm to play in a T-armed slot machine. Once the decision is made, the electric bill for that day is known in the end. However, before the end of the day, the consumers would not know the electric bill associated with choosing each time period t. Then, each consumer faces the trade-off between exploration – trying more arms that are not yet chosen, and exploitation – keep choosing the arm that gives the best reward (i.e., the lowest electric bill) so far. MAB problems have been well-

studied in the literature (see, for example, [5] and Chapter 6 in [8] for an overview). A key assumption in most of the MAB problems, however, is that the reward distribution of each arm, though unknown, is stationary. In the case of price-based DR, however, each consumer's reward (i.e., the electric bill) at each time period within a day may not be stationary, as the reward depends on the collective actions of all consumers. This lack of stationarity may be the major reason that there has been little work on MAB games, in which each player in the game solves an MAB problem. As mentioned in the introduction section, a recent breakthrough on MAB games in [3] has provided the theoretical foundations in studying the price-based DR as an MAB game. The specific settings and the algorithms of regret-minimization are provided below.

Decision epochs and arms. Each consumer $k_c \in \{1, \ldots, n\}$ makes a single decision at each day $d = \{1, 2, \ldots\}$, which is to determine within which time periods $t = \{1, \ldots, T\}$ to consume the flexible load. Each t is considered as an arm. Note that the assumption of a single decision per decision epoch is not necessarily very restrictive, as a decision epoch is somewhat arbitrary. For example, if consumers need to make more decisions in a day, we can just reduce the decision epoch to, for example, every 4 hours. We will report the effects on the simulation results with different temporal resolution of decision epochs in our future work.

Consumers' types. The types here represent the characteristics of electricity consumption of each consumer, such as the type of load, the consumption level, etc. For simplicity, we only consider the the consumption level of each consumer's flexible load, which are randomly generated according the description in Subsection II-A, and denoted as θ^{k_c}, for $k_c \in \{1, \ldots, n\}$.

States. The state of consumer k_c in day d, denoted by $z_d^{k_c}$, is a simplification of the history of the MAB game. The same as in [3], $z_d^{k_c}$ contains $2T$ elements, with T being defined before as the number of time periods in a day. The first T elements record the number of times that each arm $t \in \{1, \ldots, T\}$ has been chosen by consumer k_c; while the second T elements denote the average rewards (from $d = 0$ to the current day d) associated with each arm t. In addition, we let $\mathcal{Z}_d^{k_c}$ be the set of all possible states for consumer k_c at day d; hence, $\mathcal{Z}_d^{k_c} \subset \mathbb{Z}_+^{2T}$, with $\mathbb{Z}_+$ being the set of nonnegative integers.

Policies (or strategies).[3] Let $\Xi = \{\xi = (\xi_1, \ldots, \xi_T) : \sum_{t=1}^{T} \xi_t = 1\} \in [0, 1]^T$ be the set of probabilities. Then in the T-armed bandit problem faced by each consumer k_c, a policy, denoted by $\sigma_{k_c} : \mathcal{Z}_d^{k_c} \to \Xi$, is a function that maps from the current state variable space to the probability set Ξ.

For consumers employing a policy as defined above, the actual arm that a consumer k_c will choose in day d is then a random variable, denoted by $x_d^{k_c}(z_d^{k_c})$, as the value of the random variable depends on the current state of consumer k_c. The range of $x_d^{k_c}(z_d^{k_c})$ is the number of arms to choose from; that is, $x_d^{k_c}(z_d^{k_c}) \in \{1, \ldots, T\}$. Since the policy $\sigma_{k_c}(z_d^{k_c})$ is a vector valued function, we use $\sigma_{k_c}(z_d^{k_c}, t)$ to denote the probability of consumer k_c choosing the arm t. Then the probability distribution of $x_d^{k_c}(z_d^{k_c})$ for day d is

$$Prob(x_d^{k_c}(z_d^{k_c}) = t) = \sigma_{k_c}(z_d^{k_c}, t), \ \forall t, \ k_c, \text{ and } z_d^{k_c} \in \mathcal{Z}_d^c. \tag{2}$$

Population profile. Since each consumer's payoff (or electric bills) also depends on what other consumers do, we define the concept of population profile as the histogram of the arm choices of all consumers, denoted by $f_d(t)$. More specifically, for $d = 0, 1, \ldots$,

$$f_d(t) = \frac{1}{n} \sum_{k_c=1}^{n} \mathbb{1}_{\{x_d^{k_c}(z_d^{k_c})=t\}}, \ \forall t \in \{1, \ldots, T\}. \tag{3}$$

We use $\mathbf{f}$ to denote the dynamics of $\{f_0, f_1, f_2, \ldots, \}$. Since $x_d^{k_c}$'s are random variables; so are $f_d(t)$'s. Whether $f_d(t)$ follows a stationary distribution (after a certain number of days) is a key point in the MAB game, and will be discussed further when we discuss convergence of the MAB game to a steady state.

Rewards. We define the reward (or utility) for consumer k_c to choose arm t in day d, denoted as $U_d^{k_c}$, to be the negative of the corresponding electric bill in day d, which equals the negative of energy consumption level θ^{k_c}, multiplying real-time price at time t, $P_{dt}^{RT}(f(t))$. The real time price is determined through the economic dispatch process performed by the ISO, as illustrated in Fig. 1, when the actual aggregated demand at t (i.e., $f_d(t)$) is realized. More specifically,

$$U_d^{k_c}(\theta^{k_c}, t, f(t)) := -\theta^{k_c} P_{dt}^{RT}(f(t)). \tag{4}$$

Regeneration. A novel and key idea in an MAB game is regeneration, as proposed in [3]. More specifically, we assume that at each day d, each consumer has a probability β to be regrenerated, meaning that its state variable z_d will be re-initialized to all 0's, and its type, i.e., the energy consumption level, is re-drawn from a given distribution. This means that each consumer has a random life time following a geometric distribution. As pointed out in [3], this regeneration process accounts for the situation where there are always new customers joining the price-based DR program; while some customers in the program may opt out. A more important role for the regeneration is to ensure that even when the system reaches a steady-state, the consumers continue to learn. (Otherwise, they would just choose a fixed arm t without exploring other arms.)

Regrets and regret-minimizing policies. For a consumer k_c, let $\bar{z}_d^{k_c} : d \in \{0, 1, \ldots, D-1\}$, denote the states visited by k_c under a fixed policy σ_{k_c} up to day $D-1$. If we assume that the population profile $\mathbf{f}$ is stationary (again, this will be discussed more later), then the expected reward corresponding to pulling arm t for consumer k_c will remain the same across d's. As a result, we can ignore the index d in the reward expression in (4) and define the largest expected value of U^{k_c} over all $t \in \{1, \ldots, T\}$, which is denoted as $\mu^{k_c^*}$. Then the regret of

[3]Herein we use the words policy and strategy interchangeably, as 'strategy' is more commonly used in the game-theory literature, while 'policy' is more widely used in dynamic programming and machine learning community.

k_c after D rounds of decision-making is defined as follows:

$$R_D^{k_c} := D\mu^{k_c{}^*} - \frac{1}{D}\left[\mathbb{E}\sum_{d=0}^{D-1} U_d^{k_c}[\theta^{k_c}, \sigma_{k_c}(\bar{z}_d^{k_c}, t), f(\sigma_{k_c}(\bar{z}_d^{k_c}, t))]\right] \tag{5}$$

Regret-minimizing policies for a (single-agent) MAB problem have been well-studied. One popular policy can be obtained through the so-called UCB (Upper Confidence Bound) algorithm [1]. The algorithm is simple: at each decision epoch d, a consumer chooses the arm $\hat{t}$, with

$$\hat{t} \in \text{argmax}_{t\in\{1,\ldots,T\}}\left\{\overline{U}^{k_c}(t) + \sqrt{\frac{2\ln(d)}{d_t}}\right\}, \tag{6}$$

where $\overline{U}^{k_c}(t)$ represents the average reward for consumer k_c when the arm t is chosen up to decision epoch d, and d_t is the total number that arm t has been chosen. The objective function in (6) reflects the trade-off between exploitation (choosing a large $\overline{U}^{k_c}(t)$) and exploration (i.e., when d_t is small, the second term in (6) is large, forcing arm t to be chosen more often). It has been shown in [3] that under the geometric regeneration, as described above, the UCB policy is ϵ optimal; that is, the regret $R_D^{k_c}$ is upper bounded by ϵ:

$$R_D^{k_c} < \epsilon, \text{ with } \epsilon = \sum_{D=1}^{\infty}(1-\beta)\beta^{D-1}\frac{\alpha\ln(D)}{D}, \tag{7}$$

where α is some constant.

The UCB algorithm is easily programmable in a control automation device, such as in a smart home control hub, with minimum data storage or computing requirement. Policies obtained from other regret-minimizing-based algorithms can also be applied here. As pointed out in [3], the convergence to a steady state of an MAB game does not depend on the specific policy chosen.

Convergence to steady states. Let $\phi \in \Phi$ denote a joint distribution over all consumers' state spaces and type spaces, where Φ is the space of all Borel probability measures on the joint state and type spaces of all consumers. In [3], a pair $(\phi, \mathbf{f})$ is defined as a mean field steady state (MFSS) of an MAB game if it satisfies two conditions: first, given a stationary population profile $\mathbf{f}$, which will influence the state transition, can yield a steady state distribution ϕ; second, based on the steady state distribution ϕ of consumers' states and types, the stationary population profile $\mathbf{f}$ can indeed emerge from the MAB game.

Strong theoretical results regarding MFSSs have been shown in [3], including existence of an MFSS under any policy σ, uniqueness, and asymptomatic convergence to a MFSS when the number of players in the MAB game approaches to infinity. The last property is especially useful in the context of demand response in energy markets, as the number of price-responsive consumers can be very large. Even if numerical simulations could not handle the large number of agents, the volatility-soothing effect in an MAB game (as to be shown in the numerical results in the next section) remains the same for a large number of consumers, which is reassuring from the ISO's perspective should real-time pricing to be implemented in the real world. Another nice feature of the MAB-game approach is that consumers do not need the DA price, as they only learn from the past real-time prices (and the corresponding rewards). Also this is a completely decentralized approach in the sense that the ISO does not have to do anything additionally or differently than how they operate the DA and the RT markets now.

Note that a MFSS is in general not a PBNE to the corresponding dynamic game, as there may exist certain histories of the game under which a consumer k_c may have the incentive to deviate from its MFSS policy σ_{k_c} in order to maximize its discounted expected payoffs (this is so since regret-minimization may not be the same as discounted expected payoff optimization). However, the strong theoretical results associated with MFSS assure its applicability in real world. This is so because if all consumers have certain control automation devices coded with a regret-minimization algorithm, then a MFSS will emerge. In addition, if each consumers' regrets can be upper bounded or even approach to 0 as $d \to \infty$, depending on the algorithms used, the results should be acceptable to most of the consumers, who otherwise may not be able to glean the benefits of real-time pricing anyway.

The theoretical proofs in [3], however, are not directly applicable to the specific MAB game described here, as the reward associated with pulling each arm here is more complex than that in [3], where the reward is simply a Bernoulli random variable. We will extend the proofs to our MAB game in our immediate future research. Even without the theoretical results, numerical results from the MAB-game-based simulation are very encouraging, which are presented below.

C. The Naive-Response Model

In the naive-response model, consumers with flexible load will respond to the day-ahead prices to determine when to consume electricity in the next day. From the ISO's perspective, the key is how to forecast the demand in the next day. A more sophisticated approach is to recognize the closed-loop relationship between price forecasts and the actual demand; that is, the ISO will anticipate how the consumers would respond to a set of day-ahead price forecasts. Such an approach will require the ISO to employ methods from dynamic programming and optimal control, as studied in [14]. As our focus here is to study what the market outcomes would be when consumers' respond to real-time pricing under the *current* market operations, we do not consider the closed-loop approach of the ISO. Instead, we assume that the ISO forecasts the next day demand based on the past N days of realized demand. More specifically, the ISO demand forecast for day $d+1$, denoted as L_{d+1}^{DA}, is as follows:

$$\begin{aligned} L_{d+1}^{DA}(t) = \; & \frac{1}{N}\sum_{p=0}^{N}\sum_{k_c=1}^{n}\theta^{k_c}\mathbb{1}_{\{x_{d-p}^{k_c}=t\}} + BD_t, \\ & t = 1,\ldots,T,\ d \geq N\ . \end{aligned} \tag{8}$$

For the initial N days, the ISO just uses the average of all the past demand data; and for $d = 0$, the ISO is assumed to use the expected value of θ^{k_c}. (Note that while being a fixed number, θ^{k_c} is randomly drawn from a distribution in our simulation.)

With the day-ahead demand forecasts in (8), the day-ahead market clearing process produces the day-ahead prices, denoted as $P^{DA}_{d+1}(t)$, $t = 1, \ldots, T$. Based on the day-ahead prices, in the naive-response model, consumers choose the lowest $P^{DA}_{d+1}(t)$ to commit their flexible load θ^{k_c} in day $d+1$; that is, $x^{k_c}_{d+1} = \check{t}$, where $\check{t} \in \min_{t=1,\ldots,T}\{P^{DA}_{d+1}(t)\}$.

Similarly, in the case of prosumers with DG resources, each prosumer will choose the highest price period in the next day to generate θ^{k_p}; that is, $x^{k_p}_{d+1} = \hat{t}$, where $\hat{t} \in \max_{t=1,\ldots,T} P^{DA}_{d+1}(t)$.

Since all consumers receive the same day-ahead prices, and the decision rules are the same across the consumers, it is obvious that the real-time prices will exhibit both large deviations from the day-ahead prices and high volatility, when the percentage of flexible load is sufficiently high. Such points have been confirmed by our simulation results, which are presented and discussed in Section III.

D. Adaptive Mechanism

While the naive-response model may be too primitive in realizing demand response, we consider another decentralized approach proposed in [9], referred to as an adaptive mechanism. More specifically, a consumer c gradually adapts its decision towards the optimal selection $x^{c,*}$ (which is computed as the selection of the period t in a day d with the lowest average DA price) as follows: $x^c_{d+1} = x^c_d + \gamma(x^{c,*} - x^c_d)$, where γ is the adaptive rate. We can see that when $\gamma = 1$, the adaptive-response is exactly the naive-response model.

III. Test Case and Numerical Results

In this section, we present the test case and the simulation results corresponding to the three decentralized approaches to implement demand response as described in the previous section.

A. Input Data

System network. A simplified power system corresponding to the ISO New England (ISONE) wholesale power market has been developed in [4], and is used here for our testing purpose. Such a system consists of 8 zones, as illustrated in 2, in which green spots represent bus nodes and red spots represent generators. There are a total of 76 fossil fuel-fired generators, representing different generation technologies. The detailed data of the generators, including their generation costs and capacities, can be found in [4]. In addition, there are 12 transmission lines in the system. For our testing purpose (in stead of matching real-world data), we set all transmission lines' capacity to be $1000MW$.

At each zone, we assume that there are two types of loads: fixed and flexible. The fixed load does respond to price signals, and their values are summarized in the 24-hour base load

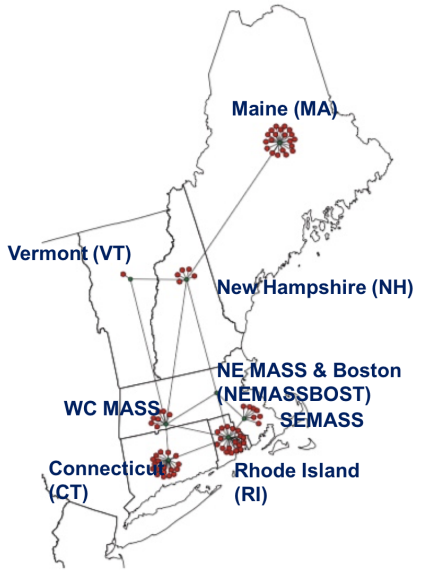

Figure 2: Transmission network for the 8-Zone ISO New England Test System [4].

profiles in Fig. 3. For flexible loads, we consider 200 heterogeneous consumers at each zone. For each consumer c, its type (aka its flexible load) is sampled from a Beta distribution of factor (2, 2), multiplied by a scaling factor.[4] Different consumers have different scaling factors, representing different levels of energy consumption (such as different household sizes for residential consumers). In addition, we assume that the types are identically and independently distributed. For all three models, we employ the regeneration scheme as introduced in Subsection II-B. The regeneration rate is set at 0.1, which means that each consumer's flexible demand amount is re-sampled from the Beta distribution every 10 days on average.

Decision epochs and temporal resolution. We simulate 200 decision epochs (i.e., 200 days) in our numerical studies. The number of decision epochs is just an arbitrary number set to be large enough for the MAB-game model to converge to a steady state. For the temporal resolution $t \in \mathcal{T}$ within each day d, while it is usually measured in hours or half-hours; here, as a starting point, we consider $T = 6$ for a day, and each period t consists of four consecutive hours. The main reason for this coarser resolution is to reflect the fact that certain loads, such as electric vehicle (EV) charging, require multiple hours to complete a task. At this point, we do not want to consider time-linking constraints in our models, and the 6-period partition of a day should be sufficient for common household loads to complete one cycle of their tasks.

Wind generation. In addition to the 76 thermal power plants, we also consider (aggregated) wind power plants in each zone.

[4]Note that there is no particular reason for the specific distribution chosen here. The key is that consumers' types can be random in this framework.

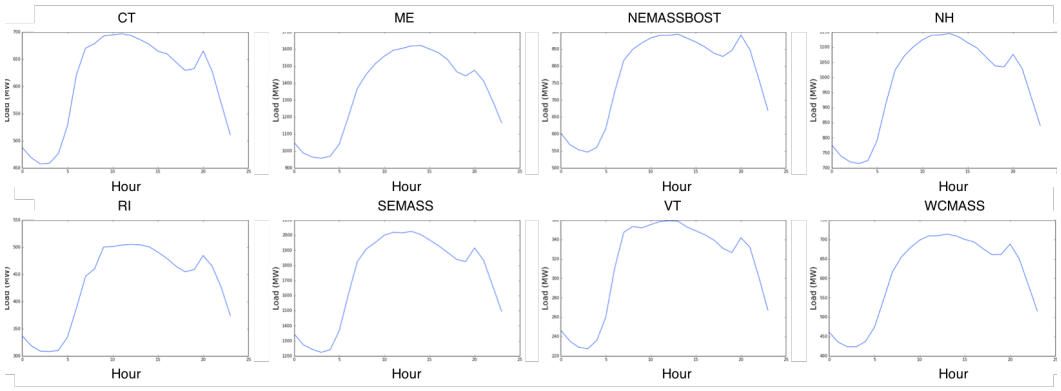

Figure 3: 24-hour Base loads for 8 zones in ISO New England Test System.

At each zone, the wind generation capacity is assumed to be of 30% of the average hourly load. The capacity factor (the ratio between actual generation outputs and total capacity within a time period) of wind plants in each zone of each hour is sampled from a Beta distribution of factor(2, $\frac{BD_h}{Capacity}$), where BD_h is the base load for hour h and $Capacity$ is the wind generation capacity. This setup is to reflect the fact that wind output may be the highest during off-peak hours, but the lowest during on-peak hours. While such wind simulation is overly simplistic, more sophisticated simulation approach to reflect both autocorrelation (such as in [7]) and spatial correlation can be easily added within the MAB game framework.

B. Simulation results

For the MAB-game model, in addition to the UCB algorithm described above, we also consider another policy: the ϵ-greedy algorithm. More specifically, at each day d, a consumer c chooses the arm t with the highest average reward so far with probability $1-\epsilon$, and randomly chooses another arm with probability ϵ. In our simulations, each consumer has the same probability of using either the UCB algorithm or the ϵ-greedy algorithm. The reason for this mix-up is to demonstrate the robustness (and practicability) of the MAB game approach; namely, the agents in an MAB game do not have to all use the same policy.

For the adaptive-response model, we select two adaptive rates: 0.3 and 0.7, for comparison purpose.

For each model, we perform four simulations with different seeds of the random variables. We use four different colors to plot the resulting price paths from each simulation of each model.

First we present the realized real-time electricity prices of the load pocket area Boston, which is the spot in Fig. 2 with no generators connected (NEMASSBOST in the test system). Other bus nodes have very similar results. Fig. 4 shows hourly average real-time electricity prices of the Boston zone.

It can be seen from Fig. 4 that the MAB-game framework quickly converges to a steady state in all four simulations. Not only the volatility is significantly smaller than other models,

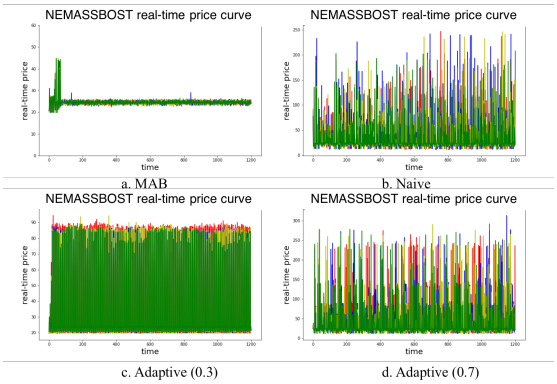

Figure 4: Hourly average real-time electricity prices of Boston.

this approach also has the peak-shaving/valley-filling effect; that is, the realized prices of all periods are very similar. This is a much desired result, as a flatter price (and load) curve means that the system is more predicable, and hence, would be more reliable. To quantitatively compare real-time price volatility across different models, we adopt the measure presented in [10], which is referred to as the log-scaled incremental mean volatility (IMV). More specifically,

$$IMV := \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} |log(P_{t+1}^{RT}) - log(P_t^{RT})|. \quad (9)$$

The IMV results of different models are presented in Fig. 5, which clearly demonstrate the advantage of the MAB game approach.

To study the effects of the three different approaches at system level, Fig. 6 presents daily average system costs of economic dispatch and transmission congestion. Arguably the most interesting results are the congest-cost reduction from the MAB game approach. The results are remarkable in the sense that the congestion cost reduction is achieved purely through

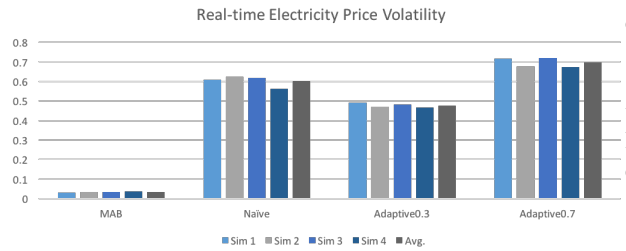

Figure 5: IMV of period's hourly average real-time electricity prices of Boston.

consumers' learning of the past prices (and the corresponding electric bills,) without consumers knowing anything about the system topology or the power generators.

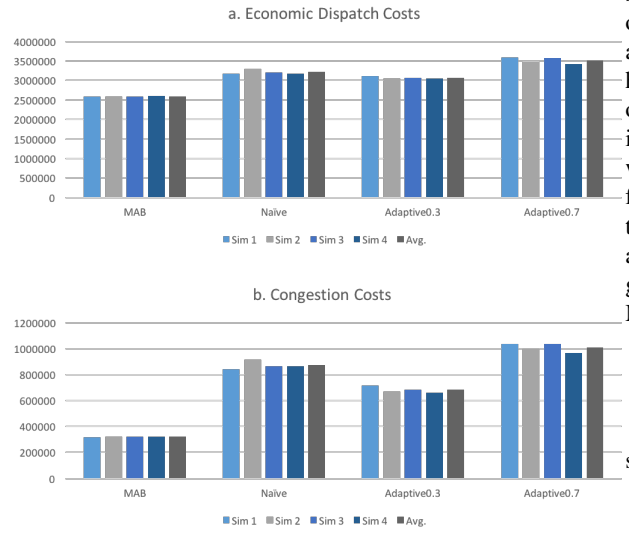

Figure 6: Daily average system costs. (a) economics dispatch costs; (b) congestion costs.

IV. Conclusions and Future Work

In this paper, we studied three decentralized approaches to implement real-time pricing and demand response in an energy market: the naive-response model, the adaptive mechanism and the MAB-game model. The resulting real-time prices from the naive-response model exhibit large variations on a daily basis, confirming the concerns raised in [10]. This large variation is fundamentally due to operating a closed-loop system in an open-loop fashion. Based on simulation results, we see that such an issue can be overcome by introducing learning-based algorithms to consumers, which will bring randomization into their decision making, and hence, avoiding the problem of having all consumers move to the same direction at the same time. While the adaptive mechanism is designed along this line, we see through simulations that the MAB game approach can achieve much greater benefits from a system perspective.

The MAB-game introduced in this paper, however, is only a starting point, and it has several limitations. First and foremost, while its feature of not relying on any price forecasts (and only learns through the history) may be considered as a strength, it can also be viewed as a weakness, especially when the power system is experiencing some emergency situations, such as the sudden loss of generation assets/transmission lines. Demand response is expected to be able to provide emergency response in such situations. However, this is not possible within the current MAB game framework. We are investigating approaches for consumers to incorporate price forecasts (or any emergency signals sent from ISOs) in their MAB algorithm. Second, the current MAB game model does not have explicit modeling of thermal loads (e.g., HVAC) yet. In reality, such load resources may be the major source of demand flexibility. We have already started working towards this direction and obtained some positive preliminary results, which will be reported in our follow-up papers. Third, the theoretical results in [3] are obtained without exogenous uncertainty. In this context, however, uncertainties (such as renewable outputs, forced outage of physical assets) are prevalent. To extend the results in [3] to the case of exogenous uncertainty (faced by all agents) would be a significant endeavor. Last and probably the most fundamental issue is if there can be any theoretical results on the gap of the social welfare between the ideal, centralized approach and the MAB game approach, hence to be able to gauge the efficiency (in terms of resource allocation) of the MAB game approach.

Acknowledgment

The author, Andrew Liu, would like to acknowledge the support of this work by the NSF grant ECCS-1509536.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, May 2002.

[2] D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, 1991.

[3] R. Gummadi, R. Johari, S. Schmit, and J. Y. Yu. Mean field analysis of multi-armed bandit games. Available at SSRN: https://ssrn.com/abstract=2045842orhttp://dx.doi.org/10.2139/ssrn.2045842, last revised: August 11, 2016.

[4] D. Krishnamurthy. psst: An open-source power system simulation toolbox in python. In *North American Power Symposium (NAPS), 2016*, pages 1–6. IEEE, 2016.

[5] A. Mahajan and D. Teneketzis. Multi-armed bandit problems. In A. O. Hero, D. Castanon, D. Cochran, and K. Kastella, editors, *Foundations and applications of sensor management*, pages 121 – 151. Springer Science & Business Media, 2007.

[6] J. Minkel. The 2003 northeast blackout – five years later. *Scientific American*, 13, 2008.

[7] A. Papavasiliou, S. S. Oren, and R. P. O'Neill. Reserve requirements for wind power integration: A scenario-based stochastic programming framework. *IEEE Transactions on Power Systems*, 26(4):2197–2206, 2011.

[8] W. B. Powell and I. O. Ryzhov. *Optimal learning*, volume 841. John Wiley & Sons, 2012.

[9] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. Jennings. Agent-based control for decentralised demand side management in the smart grid. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 5–12. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[10] M. Roozbehani, M. A. Dahleh, and S. K. Mitter. Volatility of power grids under real-time pricing. *IEEE Transactions on Power Systems*, 27(4):1926–1940, November 2012.

[11] F. C. Schweppe. Power systems '2000': hierarchical control strategies: Multilevel controls and home minis will enable utilities to buy and sell power at real time rates determined by supply and demand. *IEEE Spectrum*, 15(7):42–47, 1978.

[12] US Department of Energy. Benefits of demand response in electricity markets and recommendations for achieving them. https://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/DOE_Benefits_of_Demand_Response_in_Electricity_Markets_and_Recommendations_for_Achieving_Them_Report_to_Congress.pdf. Last accessed: 1/23/2017.

[13] D. Wogan. Electric utilities can now adjust your nest thermostat to shift energy demand. https://blogs.scientificamerican.com/plugged-in/electric-utilities-can-now-adjust-your-nest-thermostat-to-shift-energy-demand/. Last accessed: 1/23/2017.

[14] J. Xiao. *Grid integration and smart grid implementation of emerging technologies in electric power systems through approximate dynamic programming*. PhD thesis, Purdue University, 2013.

[15] Z. Zhao and A. L. Liu. Intelligent demand response for electricity consumers: A multi-armed bandit game approach. In *Intelligent System Application to Power Systems (ISAP), 2017 19th International Conference on*, pages 1–6. IEEE, 2017.