DECENTRALIZED PRICE-DRIVEN DEMAND RESPONSE IN SMART ENERGY GRID

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Zibo Zhao

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Andrew Lu Liu, Chair

School of Industrial Engineering

Dr. Harsha Honnappa

School of Industrial Engineering

Dr. Roshanak Nateghi

School of Industrial Engineering

Dr. Xiaojun Lin

School of Electrical and Computer Engineering

Dr. Harsha Honnappa

School of Electrical and Computer Engineering

Approved by:

Dr. Abhijit Deshmukh

Head of Industrial Engineering

ACKNOWLEDGMENTS

I would like to acknowledge everyone who played a role in my academic accomplishments. First of all, I would like to thank my Ph.D. advisor, Dr. Andrew Lu Liu, for his guidance, insight, patience, and trust, and for always being inspiring and supportive. I would also like to thank him for the freedom they gave me to pursue my own interests.

I would also like to thank my committee members, Dr. Xiaojun Lin, Dr. Dionysios Aliprantis, Dr. Harsha Honnappa and Dr. Roshanak Nateghi for sharing their expertise and giving me advice and encouragement.

Also, all of my family and a lot of friends, who supported me with love and understanding. Most of all, to my parents and parents-in-law. I could never have finished this dissertation and my Ph.D. study without their help and encouragement.

Finally, my dear wife, Ao Peng, and my lovely daughter, Evelyn Zhao. Thank you for all the supports and joy you give me during my Ph.D. study.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AMI	advanced metering infrastructure
CI	confidence interval
DA	day-ahead
DER	distributed energy resource
DG	distributed generation
DR	demand response
DSO	distribution system operator
ED	economics dispatch
EMC	energy management control
EV	electric vehicle
FIT	feed-in tariff
IMV	incremental mean volatility
ISO	independent system operator
LAR	low approximate regret
LIMV	log-scaled incremental mean volatility
LMP	locational marginal price
MAB	multi-armed bandit
MCP	market clearing price
MFSS	mean field steady state
NREL	National Renewable Energy Laboratory
OPF	optimal power flow
PBE	perfect Bayesian equilibrium
PEV	plug-in electric vehicle
PoA	price of anarchy

RT	real-time
RTO	regional transmission organization
RTP	real-time pricing
SAM	system advisor model
TOU	time-of-use
UC	unit commitment
UCB	upper confidence bound

ABSTRACT

Zhao, Zibo Ph.D., Purdue University, December 2019. Decentralized Price-Driven Demand Response in Smart Energy Grid . Major Professor: Andrew Lu Liu.

Real-time pricing (RTP) of electricity for consumers has long been argued to be crucial for realizing the many envisioned benefits of demand flexibility in a smart grid. However, many details of how to actually implement a RTP scheme are still under debate. Since most of the organized wholesale electricity markets in the US implement a two-settlement mechanism, with day-ahead electricity price forecasts guiding financial and 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 in real-time. However, if such an idea is not controlled properly, the inherent closed-loop operation may lead consumers to all respond in the same fashion, causing large swings of real-time demand and prices, which may jeopardize system stability and increase consumers' financial risks.

To overcome the potential uncertainties and undesired demand peak caused by "selfish" behaviors by individual consumers under RTP, in this research, we develop a fully decentralized price-driven demand response (DR) approach under game-theoretical frameworks. In game theory, agents usually make decisions based on their belief about competitors' states, which needs to maintain a large amount of knowledge and thus can be intractable and implausible for a large population. Instead, we propose using regret-based learning in games by focusing on each agent's own history and utility received. We study two learning mechanisms: bandit learning with incomplete information feedback, and low regret learning with full information feedback. With the learning in games, we establish performance guarantees for each

individual agent (i.e., regret minimization) and the overall system (i.e., bounds on price of anarchy).

In addition to the game-theoretical framework for price-driven demand response, we also apply such a framework for peer-to-peer energy trading auctions. The market-based approach can better incentivize the development of distributed energy resources (DERs) on demand side. However, the complexity of double-sided auctions in an energy market and agents' bounded rationality may invalidate many well-established theories in auction design, and consequently, hinder market development. To address these issues, we propose an automated bidding framework based on multi-armed bandit learning through repeated auctions, and is aimed to minimize each bidder's cumulative regret. We also use such a framework to compare market outcomes of three different auction designs.

1. INTRODUCTION

1.1 Motivation and Literature Review

In traditional power systems, electricity demand is considered to be highly inflexible and thus more likely to be predictable [1, 2], as consumers have been so used to the idea of consuming electricity whenever they want to and their behavior patterns are usually quite fixed. Since the reliable operation of a power system requires the supply and demand to be balanced at all time, the inflexible demand has added great pressures on the system to maintain enough redundancies in both generation and transmission capacities. Such redundancies are very costly, as they are very 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 [3]).

On the supply side, current trends of developing power systems indicate that renewable resources are becoming more and more involved in electricity supply [4, 5]. A NREL report [1] declares that the amount of renewable supply from technologies that are commercially available today could potentially meet 80% of the US electricity demand on an hourly basis in the year 2050. The main goal of power system operations has been to dispatch the most economic set of fossil-fuel generations to match the net-demand, i.e., the demand minus the uncontrollable renewable supply. However, relying on fossil-fuel generation alone to balance renewable variability will not only be very expensive, but also environmentally unfriendly. Active demand flexibility has been considered crucial for future power systems so that the electricity consumption can be shaped to match the variable and uncertain renewable supply [1].

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 [6–8]. The direct benefits of flexible demand are huge, ranging from saving a tremendous amount of money for consumers and making the power system more robust [9–12]. There are also more subtle benefits, such as the potential environmental benefits of using the flexible demand to better match the output from renewable resources [1], such as wind and solar, and hence reducing air pollutant emissions from fossil fuel plants. However, technologies alone are not enough to realize flexible demand and its potential benefits. 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 [13]. The former approach, as the name suggests, is to have the system operators or utilities directly manage their load [14–17], and thus such flexible demand resources can participate in wholesale markets as generation supply [18]. 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 [19].

While central load control may be the most efficient in theory, as the system operators have the entire system's information and can mange the load to achieve system-wide efficiency. Such approach may be effective for a few large-scale agents like industrial plants. However, in reality, the amount of resources that system operators and utilities can control is limited, partially due to software and computing power limitation, and thus the approach is not capable of controlling over a large small-scale distributed population (e.g. potentially millions). Moreover, in smart girds,

smart home agents have their flexible demand being controlled by some controller device with embedded algorithms other than being controlled by system operators or utilities. Though the scale of flexible demand is small for each individual smart home agent, the aggregation would affect the system more and more significantly as the fast growth of smart homes. In addition, the central approach may also raise issues on privacy, as some consumers do not feel comfortable of having someone else manage their household's electricity usage, which have been a severe issue for wider deployment.

The decentralized approach, on the other hand, depends on end-consumers to make their own electricity consumption decisions independently in a desirable fashion. However, autonomous distributed entities with demand flexibility will neither be interested nor effective in collaborating to improve the overall system's capability unless appropriate incentives or information sharing mechanisms are provided. Therefore, through certain incentives or price signals, it is hoped that the collective consumers' actions may bring the desired demand flexibility from the system's perspective [9, 20–23], such as balancing the variations in renewable resources. This is commonly referred to as demand response (DR). The demand-side response will be less costly and more environmentally friendly than simply relying on fossil-fuel generation alone to match renewable uncertainty. Within the general term of DR, there are many different forms. Using the classification in [24–26], 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 [25–27]. 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 predefined 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 [25, 26, 28–32]. The price information has to reflect to some de-

gree 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 for different time-varing retail prices [33], including time-of-use (TOU) pricing, critical-peak pricing (CPP), and real-time pricing (RTP). In the RTP mechanism, 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. The visionary late MIT professor Fred Schweppe envisioned an energy future with real-time electricity pricing and actively engaged demand response back in 1978 [34], as he recognized the many benefits associated with flexible demand. Moreover, Hogan [35], Borenstein [36], and others have long argued that RTP is the most efficient market approach to achieve the potential benefits of DR.

In this work, we focus on priced-based DR coupled with RTP; namely, end-consumers conduct DR with receiving the information on real-time electricity prices, and pay their electric bills based on such prices. We believe this is the framework that can bring the most benefits to the grid for utilizing the flexible demand of the large autonomous smart home population in future. However, price-driven DR will bring in new sources of variability and uncertainty from the demand side, which in turn affect the system dynamics and real-time electricity prices. The study in [37–39] raise the concern that if the DR behaviors are not controlled properly, the closed-loop interactions between real-time prices and price-based DR, as shown in Figure 1.1, may result in undesired outcomes, such as increasing price volatility and reducing system reliability. This is so since real-time electricity prices are only available after the actual supply and demand are realized. It is no point 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 since it not only increases the risk of system instability, but also increases the

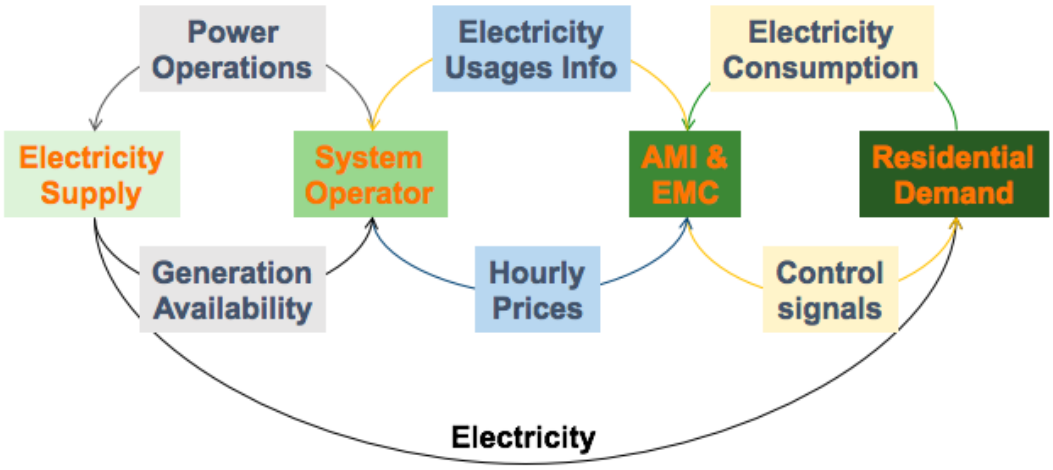

Figure 1.1.: Closed-loop interactions between real-time prices and end-consumers' behaviors.

price-uncertainty and financial risk faced by the flexible demand. Then the entities with flexible demand will be discouraged to conduct active demand response.

There have been a large amount of works on how an individual consumer should make decisions under real-time pricing, such as [40–45]. Though the optimization models proposed in the literature have been shown to be very helpful for a single smart home agent, or even a small number of agents, it could be a totally different story for a large population. This is so since the studies do not consider the closed-loop dynamics and assume that price signals are exogenous, meaning that the prices are not affected by the individual consumer's behavior. However, the collective actions of scalable population would generate undesired significant demand peaks due to overly-homogeneous patterns which, in turn, reduce the efficiency and security of the overall system.

Such a closed-loop system can be managed in a centralized approach. Robust control and optimization algorithms with complete information for centralized power system operations have been studied in the literature [9, 21, 46, 46–53]. The work in [54] solves the centralized problem through approximate dynamic programming. However, much fewer works exist to study the system-level impacts when a large

amount of consumers respond to RTP in a fully decentralized way. The existing works on implementing decentralized control under RTP are mainly of two groups. The first group is to solve a centralized control problem in a decentralized fashion, very similar to the popular alternating direction method of multipliers (ADMM) [55] for solving optimization problems arising from statistical learning. Such works include [56, 57]. While such a decentralized solution approach can reach system-wide optimality (under certain conditions), it would require frequent information exchange between a central controller (such as a system operator or a utility company) and the distributed resources. Also a system operator would update its economic dispatch algorithm to facilitate the convergence of the distributed optimization process to a system-wide optimal solution, and thus such works are not fully decentralized. The other group of works use heuristic learning on the consumers' side, such as in [58]. This bears certain similarities to the learning approaches proposed in this paper. The biggest distinction is that the learning algorithm in [58] is heuristic, without any theoretical guarantee on its performance (regardless what the performance measurement is).

Therefore, there is a void in literature on how to realize a fully decentralized price-based DR for a large "selfish" population with performance guarantee. Our study is to fill such a void by proposing fully decentralized online learning algorithms for distributed agents conducting price-driven[1] DR without causing extreme price volatility or jeopardizing system reliability under a game-theoretic framework.

1.2 Research Objectives and Contributions

In this study, we will first focus on demand-side participation of end-consumers (e.g. residential sectors) with demand flexibility. One of the research objectives in this dissertation therefore is to develop theoretic foundation and learning algorithms that allow electricity consumers with flexible demand to actively participate in price-driven

[1]In this dissertation, we use *price-based* and *price-driven* interchangeably.

demand response in an efficient manner. Since under RTP the utility (or cost) each agent receives is highly dependent on the collective actions of all other agents, other than studying about any individual agent's behavior, we develop game-theoretical frameworks that capture the decentralized nature of price-driven response due to the self-interest of market participants. Under the game-theoretical frameworks, each single agent makes decisions against the collective actions of all others for maximizing its own interest. We believe that the game approach is one of the contributions of this work due to its novelty. Moreover, we propose decentralized regret-based online learning algorithms for agents to automatically make decisions in games. With agents learning in games, we show that the performance of the overall system converge to efficient outcomes.

In this dissertation, two game-theoretical frameworks are presented. One is multi-armed bandit (MAB) games in which each agent solves its demand response problem through playing regret-minimizing bandit learning. The study in [59] has built theoretical foundations for our work in which it proves that as the population increase in MAB-game, the overall system converges to the mean filed steady state (MFSS) with stationary population profile. Such property is desired in the demand response field since the system demand has less volatility and thus higher stability. Moreover, we show that the MAB-game for DR by a large population has the smoothness property [60] by which the overall system's performance is bounded by its price of anarchy (POA). Therefore, the advantages of the MAB-game framework are as follows. First, it is a completely decentralized approach without needing any operational changes by system operators; nor does it need two-way information exchange between system operators and consumers. Yet desirable (though not necessarily optimal) system-level outcomes bounded by its POA, such as flatter load curves and less volatile real-time prices, can be achieved. Second, it allows consumers' heterogeneity and bounded rationality; meaning that the consumers are not required to optimize to determine future actions and do not need to use the same bandit strategies/algorithms in responding to price signals. Also, each individual agent's regret (for not playing the

underlying optimal actions) is bounded. Third, the approach is scalable. Even when numerical simulation may be impractical when the number of consumers increases, the MAB-game can be shown to converge to a mean-field equilibrium when the consumer number goes to infinity.

Instead of only utilizing the information observed with the played actions by bandit learning, the other game-theoretical framework we propose is for agents learning with full realized information feedback. As all agents play low approximate regret (LAR) algorithms, with the smoothness we show in this dissertation, the overall system can approximately converge to efficient outcomes, as established in [61]. Further, since in the LAR-game framework, agents make decisions based on full information feedback, we can show that the POA bound for the LAR-game is tighter than that for the MAB-game, and thus the LAR-game framework result in better system performance in general. Though the LAR-game approach is also fully decentralized, scalable, and have performance guarantee for both individual agent and the overall system, it does not enjoy the heterogeneity as MAB-game, and thus for now we need all agents to use the same learning algorithms for their price-based DR.

In addition to the novel demand response game-theoretical frameworks, we also propose a peer-to-peer (P2P) energy trading market in distribution systems based on the MAB-game for better incentivizing participation of demand-side with distributed generation. As more and more smart homes are equipped with distributed energy resources (DERs), such resources become a vital part of a smart grid. DERs can improve system reliability and resilience with their proximity to load, and promote sustainability, with the majority of DERs being solar and wind resources [62, 63]. To better incentivize investments in DERs, we propose P2P energy trading through double-side auctions instead of providing fixed feed-in-tariff and time-of-use rates for end-customers. Further we provide an algorithmic-framework based on the MAB-game that can be automated to aid consumers and prosumers (i.e. consumers with generation resources) to participate in repeated double auctions with bandit learning. Numerical simulations are conducted to study the market outcomes of three differ-

ent auction designs: a replicate of the wholesale market's uniform-price auction, a variant of Vickrey double-side auction [64], and maximum volume matching auction (which is pay-as-bid/receive-as-ask) [65]. Numerical results indicate the convergence of the market outcomes of the MAB-game to a steady-state in terms of total cleared quantities, total social welfare and total normalized reward.

The dissertation proceeds as follows. In the remainder of this chapter, technical background on two-settlement of electricity markets and real-time pricing will be provided for readers. Chapter 2 presents the game-theoretical framework of MAB games in which a large population of end-consumers make decisions regarding their flexible demand usage by regret-based bandit learning. The performance bounds for both individual agent and overall system are established. Chapter 3 presents the other game-theoretical framework LAR-game in which the large population use LAR learning with realized full information feedback to make DR decisions. The corresponding individual and system performance bounds are established as well. Meanwhile, both the theoretical and numerical comparisons between MAB-game and LAR-game are conducted in the chapter. At the end of the chapter, we use the test-bed of 8-zone Independent System Operator New England (ISO-NE) [66] to simulate the proposed gram-theoretical frameworks, and compare to the heuristic approach in [58] and a naive approach. In Chapter 4, we apply the MAB-game approach for a P2P energy trading market and simulate the approach under three different double-side auction designs. Chapter 5 concludes the dissertation and discusses potential future works.

1.3 Technical Background

1.3.1 Wholesale Electricity Markets

Generally, market participants in wholesale electricity markets can be grouped into demand side as buyers (e.g. utilities) and supply side as sellers (e.g. generation plants). There are also financial participants that can be either buyers or sellers in the

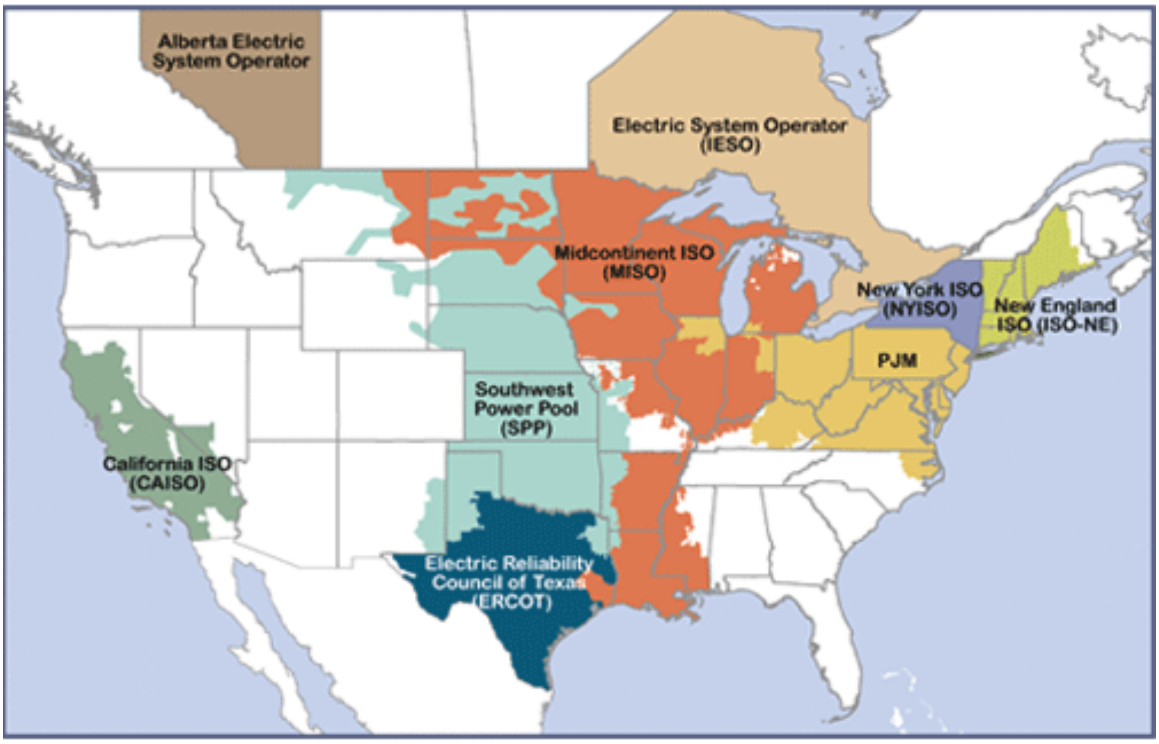

Figure 1.2.: Independent System Operators in North America [67].

markets. Independent System Operators (ISOs), typically non-profit, organizes and clears the markets through exercising final authority over the dispatch of generation. ISOs have to preserve reliability and facilitate efficiency, ensure non-discriminatory access, administer transmission tariffs, ensure the availability of ancillary services, and provide information about the status of the transmission system and available transmission capacity [68]. The distribution of current ISOs in North America is shown in Figure 1.2. In the white areas with no ISO, there are no organized wholesale electricity markets and the supply-demand balance is maintained by vertically integrated utilities.

In reality, The market clear work by ISOs are complex since demand-supply balance, security, physical, and many other constraints have to be taken into account. The main idea is to maximize the total social welfare within all the constraints. Simply speaking, market participants submit offers and bids into the auctions for each

time slot. In an auction, sellers offer the amount they can generate and ask for a price (usually at their marginal generation costs), and buyers bid the amount they would demand and how much they are willing to pay for it. When market closes, demand bids are sorted downward by prices while supply asks are sorted upward by prices. The intersection of the two curves clears the market as shown in Figure 1.3. All the accepted bids and offers take the uniform price P^* and the auction clears Q^* units of energy. Note that we focus on consumers' responses to real-time electricity prices, and hence do not consider active demand bidding in day-ahead markets.

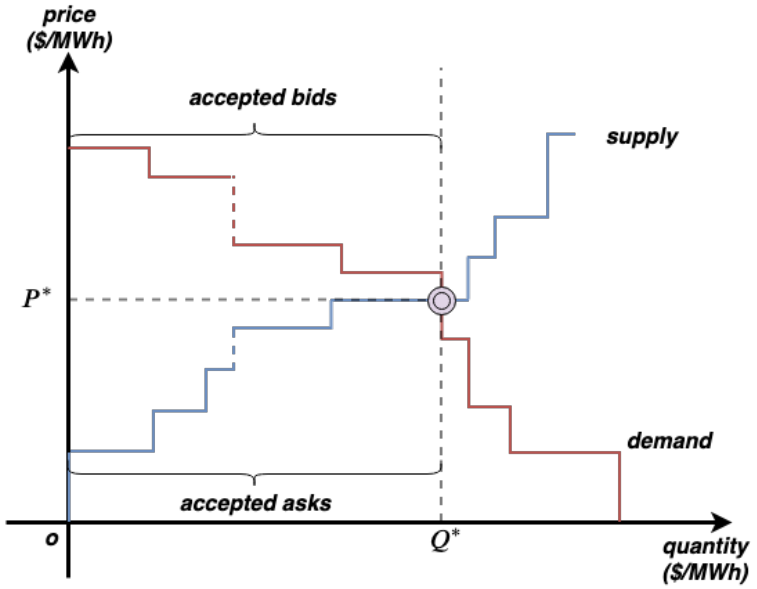

Figure 1.3.: Double-side auction with uniform clearing price.

1.3.2 Two-settlement Process

ISOs usually use a two-settlement fashion, unit commitment (UC) and economic dispatch (ED), to dispatch energy resources as shown in Figure 1.4. In a day-ahead (DA) market, an ISO solicits supply bids from power generators to meet the demand forecasts of each time period in the next day. Such a market acts as the basis for the market transactions and is used for generation schedules for the next day. The

market clearing prices for the DA market are referred to as DA prices. In real-time (RT) on the actual operating day, the ISO matches any supply and demand deviations with additional generation resources. Such additional balancing produces the RT electricity prices. Then for each time slot h, a market participant's total payment or revenue consists of DA and RT settlement as below:

$$\text{Settlement}_h = p_h^{DA} \cdot q_h^{DA} + p_h^{RT} \cdot (q_h^{RT} - q_h^{DA}), \tag{1.1}$$

where Settlement_h denotes the total pay/receive settlement, (p_h^{DA}, q_h^{DA}) and (p_h^{RT}, q_h^{RT}) denote the clearing price and quantity for DA and RT, respectively.

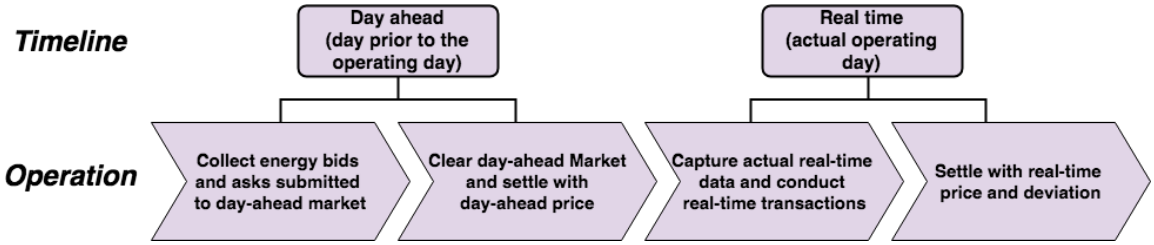

Figure 1.4.: Two settlement for day-ahead market and real-time market.

In day ahead, unit commitment is implemented with constraints like generator start-up/shut-down limitations and minimum up/down time requirement. An ISO finds an optimal (i.e., cost minimizing) schedule through solving large-scale linear programming (or convex quadratic) problems. Meanwhile, locational marginal prices (LMPs) are achieved as shadow prices associated with transmission flow balacning constraints. The interested readers are referred to [69, 70] for more details. In this dissertation, since we study demand response under RTP, a simplified direct current (DC) model is adopted. The UC model considering time-linking constraints is as below in (1.2):

$$
\begin{aligned}
\min_{g,u,v,x} \quad & \sum_{t=1}^{T}\sum_{z=1}^{Z}\sum_{i=1}^{N_z}[c_{z,i}(g_{z,i,t}) + u_{z,i,t}\cdot SU_{z,i} + v_{z,i,t}\cdot SD_{z,i}] \\
\text{s.t.} \quad & y_{z,t} = \sum_{i=1}^{N_Z} g_{z,i,t} - d_{z,t},\ \forall z,t \text{ (KCL)} \\
& \sum_{z=1}^{Z} y_{z,t} = 0,\ \forall t,\ \text{(supply = demand)} \\
& -T_l \le PTDF_{l,z}\cdot y_{z,t} \le T_l,\ \forall l,t,\ \text{(KVL)} \\
& K_{z,i}^{min} \le g_{z,i,t} \le K_{z,i}^{max},\ \forall z,i,t,\ \text{(capacity constraint)} \\
& -RD_{z,i} \le g_{z,i,t} - g_{z,i,t-1} \le RU_{z,i},\ \forall z,i,t,\ \text{(ramping constraint)} \\
& u_{z,i,t} \ge x_{z,i,t} - x_{z,i,t-1},\ \forall z,i,t,\ \text{(start-up)} \\
& v_{z,i,t} \ge x_{z,i,t-1} - x_{z,i,t},\ \forall z,i,t,\ \text{(shut-down)} \\
& x_{z,i,t} - x_{z,i,t-1} \le x_{z,i,\tau}, \tau = t,\ldots,t+MU_{z,i},\ \text{(minimum-up time)} \\
& x_{z,i,t-1} - x_{z,i,t} \le 1 - x_{z,i,\tau}, \tau = t,\ldots,t+MD_{z,i},\ \text{(minimum-down time)}
\end{aligned}
\tag{1.2}
$$

The corresponding ED model without time-linking constraints is as below:

$$
\begin{aligned}
\min_{g} \quad & \sum_{z=1}^{Z}\sum_{i=1}^{N_z} c_{z,i}(g_{z,i}) \\
\text{s.t.} \quad & y_z = \sum_{i=1}^{N_Z} g_{z,i} - d_z,\ \forall z,\ \text{(KCL)} \\
& \sum_{z=1}^{Z} y_z = 0,\ \text{(supply = demand)} \\
& -T_l \le PTDF_{l,z}\cdot y_z \le T_l,\ \forall l,\ \text{(KVL)} \\
& K_{z,i}^{min} \le g_{z,i} \le K_{z,i}^{max},\ \forall z,i,\ \text{(capacity constraint)}
\end{aligned}
\tag{1.3}
$$

The descriptions for the notations used in (1.2) and (1.3) are summarized in Table 1.1.

Table 1.1.: Notations in (1.2) and (1.3).

Notation	Description
z, i, t, l	index of node, generator, time, and transmission line, respectively.
$d_{z,t}, y_{z,t}$	demand and net demand of node z at time t.
$g_{z,i,t}$	output of node z's generator i at time t.
$c_{z,i}$	cost function of node z's generator i.
$u_{z,i,t}, v_{z,i,t}$	start-up and shut-down variable, respectively.
$SU_{z,i}, SD_{z,i}$	start-up and shut-down cost, respectively.
$PTDF_{l,z}$	Power Transmission Distribution Factor between l and z.
T_l	transmission capcaity of line l.
$K_{z,i}^{min}, K_{z,i}^{max}$	generator capcaity limits.
$RD_{z,i}, RU_{z,i}$	generator's ramping down/up constraint.
$MD_{z,i}, MU_{z,i}$	generator's minimum down/up time constraint.
$x_{z,i,t}$,	generator's running indicator variable .

1.3.3 Real-time Pricing

We differentiate two settings in which the closed-loop dynamics between flexible demand and pricing can take place. In the first setting, the exact price or incentive signals for the next time-interval are announced to entities with demand flexibility before they adjust their demand for the time-interval. This setting is applicable to a regulated utility (an example is Georgia Power [28]). Further, in the case of a deregulated utility, this setting is also applicable to a third-party electricity supplier who, under contractual terms with the consumers, must announce price or incentive signals for each time-interval before-hand [28]. In contrast, in the second setting, the price of a time-interval is produced ex-post, i.e., after the demand has been adjusted and realized. This setting is applicable to a customer directly participating in a real-time ISO market, where payments are settled based on ex-post prices that are computed after

the interval ends. The setting also works for the situation where a customer receives price from utility/suppliers that is indexed to the ex-post ISO market prices [28],[2] as shown in Figure 1.5. In our research, we focus on the second setting, and will develop a framework with learning algorithms for market participants with flexible demand to learn how to respond to the aggregate actions of other participants, which are only reflected through the ex-post market price.

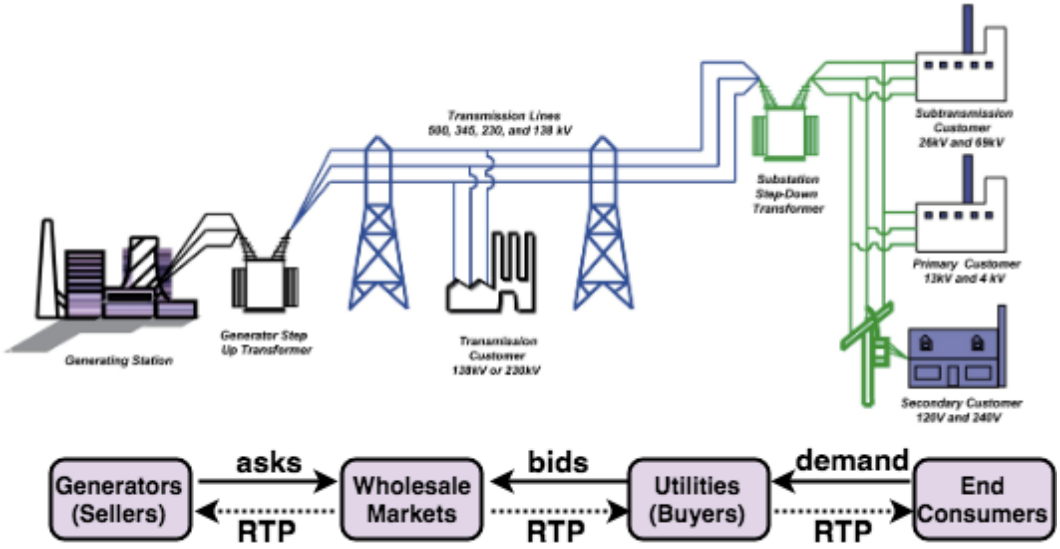

Figure 1.5.: Power system under RTP (for both wholesale market and retail end-consumers).

1.3.4 A real-world case on RTP: ComEd

ComEd is a utility company that provides its customers with RTP service for years, named as Basic Electric Service–Hourly (BESH) Energy Pricing. The customers who join the BESH program have access to day-ahead hourly market price and real-time

[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.

hourly market price and will receive Hourly Pricing alerts in addition to other useful information [71]. Specifically, the day-ahead hourly market price is the PJM day-ahead hourly market price which provides an indication of what the real-time hourly market prices could be for the following day. As an BESH Energy Pricing program participant, it is billed on the real-time hourly market price, not the day-ahead market price. ComEd simply passes along the PJM real-time hourly market prices with no mark-up. The real-time hourly market price is determined by the average of the twelve 5-minute prices from that hour, and thus the averaged real-time hourly price is not realized until after the hour is passed. With real-time hourly market prices, it is possible to have negative electricity price for short periods of time. This typically occurs in the middle of the night and under certain circumstances when electricity supply is far greater than demand. In the wholesale market, some types of electricity generators cannot reduce their output due to ramping constraints, and as a result some generators may provide electricity to the market at negative prices. In this case, customers under RTP are actually being paid to consume electricity during negative period hours. Delivery charges always apply to customers. More details about the ComEd BESH program can be found at its website. [3]

In Figure 1.6, we present the ComEd's real data of day-ahead and real-time prices for the week 09/01/2019 - 09/07/2019. We can see, without careful design, real-time prices could deviate from day-ahead forecasts seriously. Especially at 3pm on 09/07/2019, the day-ahead price is 2.8¢/KWh while the real-time price is 64.8¢/KWh. If such deviations keep happening, customers could have serious financial loss by making consumption decisions based on day-ahead prices, and thus could be discouraged to join the RTP program.

[3]ComEd hourly pricing program: `https://hourlypricing.comed.com/about/`

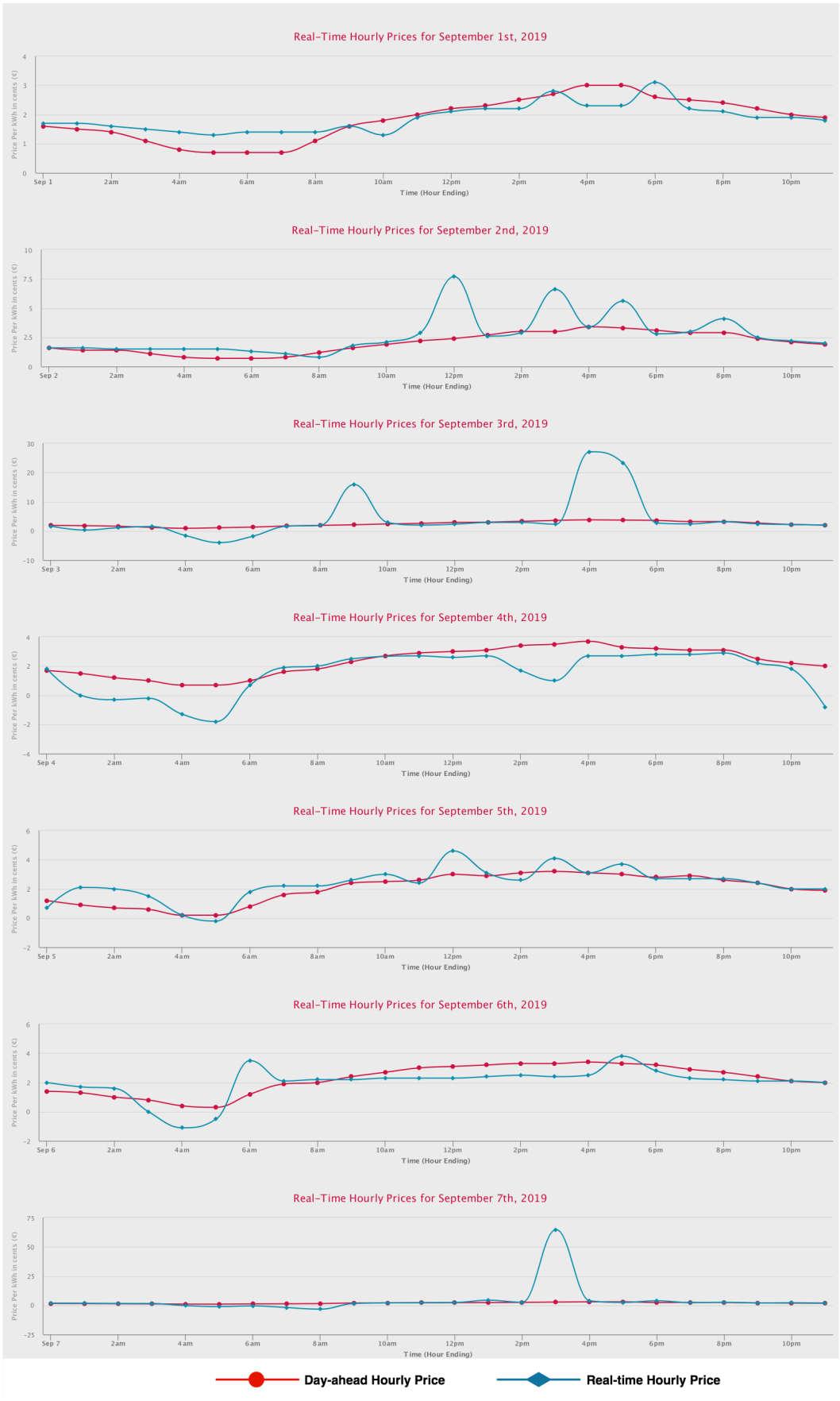

Figure 1.6.: ComEd hourly real-time and day-ahead prices (¢/KWh) for 09/01/2019 - 09/07/2019 [71].

2. DEMAND RESPONSE UNDER MAB GAMES

In this chapter, we focus on the MAB-game framework for implementing intelligent demand response. Under the framework, each agent solves a multi-armed bandit problem for its own demand response decision-making. The essence of this approach is that as each agent plays bandit learning with assuming stationary reward distributions, the overall system will approximately converge a mean field steady state (MFSS) with stationary population profile. Such property can resolve the demand uncertainties generated by end-consumers' "selfish" behaviors. Further, we show that with regret-minimizing[1] bandit learning by agents, the overall system's outcomes approximately converge to the price of anarchy (POA) of the utility-maximization games.

The rest of the chapter is organized as follows. In Section 2.1 , we describe the smart home model with flexible demand and its decision temporal resolution. Before introducing the MAB-game approach, non-theoretical approaches as baseline are presented in Section 2.2. Then in Section 2.3, we propose the MAB-game framework in details. Section 2.4 gives the system performance bound and corresponding proof.

2.1 Smart Home under Real-time Pricing

2.1.1 Smart Home Model

In this section we present a smart home model of an agent that manages its energy demands usage by maximizing its payoffs (or minimizing electricity bills) under real-time pricing. Commonly, demands can be classified into base demands and flexible demands, in which the former does not respond to price signals while the latter does with flexible usage schedule. A smart home is equipped with an controller

[1]In this dissertation, we use *regret-minimizing* and *no-regret* interchangeably.

that can automatically make decisions for the usage schedules of its flexible demands based on price information. Home appliances like washing machines, fridges, heating ventilation air conditioning (HVAC), and distributed energy resources like electric vehicles charging fall under the flexible category. In this work, we also consider smart homes equipped with distributed generation resources like rooftop solar panels and home batteries which can sell energy back to the grid with responding to price signals. Before proposing our decentralized control solutions to this, we describe a model of a smart home with non-thermal and thermal flexible demands and distributed generation resources. Here, we consider repetitive operating days denoted by $d \in \{1, 2, ...\}$. Within each day, there are time periods denoted by $t \in \{1, ..., T\}$. Each time period consists of H_t consecutive sub-periods without overlapping. The herein discussion and formulas concern a single period t. For example, a single 4-hour period t has 4 hourly time-slot arms with $H_t = 4$. The convergence rate of the MAB-game framework proposed in this chapter may depend on how many sub-periods that H_t contains. However, given a fixed H_t, the system outcomes are guaranteed to converge to the MFSS approximately as the population of agents increase.

Non-thermal Flexible Demands

We assume that for each flexible demand, there is a period that the agent would like to consume it. For example, people would like to charge their EVs after getting home. For now, we only consider EV charging as the flexible load. However, our framework can be extended to the case for a home to have different smart devices playing regret-minimization algorithms independently. Therefore, instead of modeling a set of specific appliances' demands, we consider in each time period t the total non-thermal flexible demands of a smart home's EV charging as a fixed amount $l^{f,nth}$, where the superscript f denotes flexible and nth denotes non-thermal. Then

the controller needs to decide in which sub-period to consume the load each day to minimize its average electric bills across D days under RTP, as follows:

$$\min_{x_d^{nth}} \quad \frac{1}{D}\sum_{d=1}^{D}\sum_{h=1}^{H} l^{f,nth}\mathbb{1}_{\{x_d^{nth}==h\}}p_{d,h}^{RT} \tag{2.1}$$

In the objective (2.1), $x_d^{nth} \in \{1, ..., H\}$ denotes the agent's decision for when to consume the load $l^{f,nth}$ in period t on day d, and $p_{d,h}^{RT}$ represents the real-time electricity price. The indicator function $\mathbb{1}_{\{x_d^{nth}==h\}}$ means that the smart home agent choose hour h to act over its non-thermal flexible load on day d. The D here goes to any arbitrary large number as the agent continues its demand response. Under stationary load population profile, there exist underlying cheapest time-slots for the agent to act. Then the agent desires to bound the average regret of D days for not choosing the cheapest time-slots through bandit learning.

Thermal Flexible demands

Herein, using a similar approach as in [41, 72], we simulate the warm-up demand of a HVAC system of a smart home with a linear model of indoor and outdoor temperatures as follows (which can be easily extended to the cool-down case):

$$l_{d,h}^{f,th} = \frac{1}{\beta}[TE_{h+1}^{in} - \alpha TE_h^{in} - (1-\alpha)TE_h^{out}], \tag{2.2}$$

where $l_{d,h}^{f,th}$ is the demand of a HVAC system in KWh, TE_h^{out} is the outdoor temperature in °F, TE_h^{in} and TE_{h+1}^{in} are the indoor temperature for the current and next time period in °F, respectively. The coefficient α is the system inertia[2] and β is the heating efficiency of a HVAC system in °F/KWh. The energy consumption from the warming up process is more significant than maintaining the required temperature. Therefore, in this work, we do not consider that the agents would respond to price signals with adjusting their temperature settings of HVAC systems to sacrifice comforts for reducing costs. Instead, a smart home controller manages thermostats by

[2]System inertia of a HVAC system is defined as the ability to oppose the outdoor temperature, and the α is in $[0, 1]$.

deciding when to warm up the house to a pre-set temperature such that the costs of doing so are minimized as follows:

$$\frac{1}{D}\sum_{d=1}^{D}\sum_{h=1}^{H} l_{d,h}^{f,th}(x_d^{th})p_{d,h}^{RT}, \tag{2.3}$$

where $l_{d,h}^{f,th}(x_d^{th})$ is calculated as follows:

$$\text{s.t.} \quad l_{d,h}^{f,th}(x_d^{th}) = \begin{cases} l_h^{f,th.LO} & x_d^{th} > h \\ l_h^{f,th,WM} & x_d^{th} = h \\ l_h^{f,th,HI} & x_d^{th} < h \end{cases} . \tag{2.4}$$

In (2.4), $\tilde{x}_{d,t}$ denotes the agent's decision for when to warm up the house. $l_h^{f,th,LO}$and $l_h^{f,th,HI}$ are the load of HVAC for maintaining the indoor temperature at the low and high temperature setting, respectively, and $l_h^{f,th,WM}$ is the load of HVAC for warming up the house. The superscripts f and th denote flexible thermal load. LO, HI and WM denote the scenarios of maintaining at low temperature setting, maintaining at high temperature setting, and warming up the house from low temperature to high temperature, respectively. The load of HVAC $l_h^{f,th,LO}$, $l_h^{f,th,HI}$ and $l_h^{f,th,WM}$ can be calculated by Equation (2.2) with the indoor setting and outdoor temperatures of the hour h and $h+1$. For example, when an agent leaves home for work, the indoor temperature is set at a low level and the HVAC consumes $l_h^{f,th,LO}$ in the hours that the agent is away. In the period before the agent gets home from work, the controller decides when to warm up the house to be at the high temperature setting before the agent gets home and the HVAC consumes $l_h^{f,th,WM}$ for the warming up hour. After warming up, the HVAC consumes $l_h^{f,th,HI}$ for maintaining the indoor temperature at the setting level.

2.1.2 Real-time Demand and Price

Besides the flexible demand described above, there exist inflexible base demands which are not affected by agents' behaviors and do not change across days. The

system's base demand in hour h is denoted by $\mathcal{L}_h^b$. We consider that the power system contains a set of n consumers (i.e. agents), where the consumers are indexed by $i \in \{1, ..., n\}$. The action x_i that each agent takes on day d consists of $(x_{i,d}^{nth}, x_{i,d}^{th})$ which contains the actions over non-thermal and thermal flexible loads, respectively. After each agent decides the schedule of their flexible demands, its flexible demand consumption for time slot h is as:

$$l_{i,d,h}^{f,x_i} = l_i^{f,nth} \cdot \mathbb{1}_{\{x_{i,d}^{nth}==h\}} + l_{i,d,h}^{f,th}(x_{i,d}^{th}), \tag{2.5}$$

where $l_i^{f,nth} \cdot \mathbb{1}_{\{x_{i,d}^{nth}==h\}}$ gives the non-thermal flexible load and $l_{i,d,h}^{f,th}(x_{i,d}^{th})$ gives the thermal flexible load which can be achieved by (2.2) and (2.4). Then the real-time aggregated demand is the total consumption including base demand

$$\mathcal{L}_{d,h}^{RT} = \mathcal{L}_h^b + \sum_{i=1}^{n} l_{i,d,h}^{f,x_i}, \tag{2.6}$$

which is for $\forall h \in \{1, ..., H_t\}$ and $\forall t \in \{1, ..., T\}$ and $\forall d \in \{1, 2...\}$. Though the decisions by each individual agent cannot affect the real-time prices, the collection of all agents behaviors can affect real-time prices $p_{d,h}^{RT}$, and the impact to real-time prices depend on the scale of flexible load.

2.2 Non-game Control Strategies

2.2.1 Naive-response Model

In the naive-response model, agents respond to day-ahead prices to determine when to act for each period in the next day. From the ISO's perspective, the key of day-ahead operations is how to forecast the demand for the next operating day. A more sophisticated approach is to recognize the closed-loop dynamics 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 [54]. As our focus here is to study what the market outcomes would be when agents' 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 W days of realized real-time demand. More specifically, the ISO demand forecast for day $d+1$, denoted as $\mathcal{L}^{DA}_{d,h}$ for $\forall h$, is as follows:

$$\mathcal{L}^{DA}_{d+1,h} = \frac{1}{W}\sum_{w=1}^{W} \mathcal{L}^{RT}_{d+1-w,h}. \tag{2.7}$$

With the day-ahead demand forecasts in (2.7), the day-ahead market clearing process produces the day-ahead prices, denoted as $p^{DA}_{d+1,h}$. Based on the day-ahead prices, in period t, agents choose the period with the lowest prices to act. Since all agents receive the same day-ahead price signals, and the decision rules are the same across the agents, then a large population of agents could move in the same direction. Then the resulting real-time prices will exhibit both large deviations from the day-ahead prices and high volatility, when the percentage of distributed flexible demand is sufficiently high for a large population. Our simulation results in Chapter 3's Section 3.2 have confirmed the conjecture.

2.2.2 Adaptive-response Model

While the naive-response model may be too primitive in realizing demand response, we consider another decentralized approach in [58], referred to as an adaptive mechanism. More specifically, for each period t, agent i gradually adapts its decision towards the optimal selection $\tilde{x}^*_i$ based on the day-ahead price information, as follows: $x_{i,d+1} = x_{i,d} + \alpha(\tilde{x}^*_{i,d+1} - x_{i,d})$, where $\alpha \in [0,1]$ is the adaptive rate. With applying different adaptive rate, such approach relieves the effects of large population moving in the same direction. However, the choice of the rate is totally heuristic based on agents' experience and preference. We can see that when $\alpha = 1$, the adaptive-response is exactly the naive-response model.

2.3 MAB-game Model

Under a real-time pricing mechanism, each individual agent's electric bill depends on the collection of behaviors of other agents. Therefore, this is a classic situation of a non-cooperative game in the game theory literature. However, this is not a simple static game, as when to consume the flexible load of each period (or sell the energy from distributed generations) needs to be decided on a daily basis, which makes it an instance of a dynamic game. Moreover, since agents do not know how many competitors are in the game, nor do they know the explicit payoff functions of the others, this is also an incomplete information game. In such a dynamic game with many agents, the standard equilibrium concept is Perfect Bayesian Nash Equilibrium (PBNE), which requires modeling of each agent's beliefs [73]. Specifically, a PBNE requires that agents play optimally after any history of the games by maintaining beliefs over their competitors' payoff functions based on the Bayes' rule, which is implausible for a large number of agents. In addition, as each agent's strategy profile is a function which maps the entire history to its feasible set of actions, choosing an optimal strategy profile requires exceedingly complex state information [59]., which is impractical for electricity consumers in reality.

Instead of finding the best possible strategy associated with PBNE, we can find an approximately optimal strategy by relaxing the Bayes' updating requirement. Here, we quantify an approximately optimal strategy with the concept of regrets that measure the cumulative differences between what the best response would be and what the agent chose based on a strategy across the games. Therefore, agents would like to adopt a regret minimization strategy to make decisions in the games based on the knowledge learned from history. The smart home model under real-time pricing in Section 2.1 resembles the well-studied multi-armed bandit (MAB) problem, and can use a regret-minimizing approach for the sequential decision-making based on past realized payoffs under real-time pricing. When all agents are solving their own MAB

problem with interactions, it forms a MAB-game. In the following Section 2.3.1, we discuss the MAB-game approach in details.

2.3.1 Game Settings

As presented in Section 2.1.1, in each period t of an operating day, each agent decides in which sub-period $h \in \{1, ..., H_t\}$ to consumer its flexible load. The decision-making resembles choosing an arm to play in a H_t-armed MAB problem. Before the end of each period, the agents would not know the real-time price and the electric bill associated with choosing each sub-period h. Once all agents' decisions are made, each agent's electric bill of the period is known in the end. Thus, for each period across repetitive operating days, each agent faces the trade-off between exploration versus exploitation; that is, an agent would like to trying more different choices to explore all the possibilities or staying with the choice that gives the best payoff so far to exploit it. MAB problems have been well studied, and we refer the interested readers to [74], [75], [76] and Chapter 6 in [77] for an overview.

In many classic bandit scenarios, a stationary environment is assumed in which reward on each arm has a stationary distribution. For MAB games, however, agents' collective actions would result in non-stationarity of rewards returned. A recent breakthrough on MAB games in [59] has provided the theoretical foundations in studying the dynamic interactions among agents in price-based DR as an MAB game. Each period $t \in \{1, ..., T\}$ across operating days is considered as a series of repetitive H_t-armed MAB games. The specific settings within each period t are presented below.

Decision epochs and arms.

Each agent makes its decision x_i for period t to determine within which sub-period to act. Therefore, each sub-period $h \in \{1, ..., H_t\}$ is considered as an arm for the H_t-armed MAB games.

States.

For agent i, the state for period t in day d, denoted by $z_{i,d}$, is a simplification of the history of the H_t-armed MAB games. The same is in [59], $z_{i,d}$ contains $2H_t$ elements, with H_t being defined before as the number of arms (i.e. time slots). The first H_t elements record the number of times that each arm has been chosen by the agent; while the second H_t denote the average rewards associated with each arm[3]. In addition, we let $\mathcal{Z}_{i,d}$ be the set of all possible states for agent i at day d.

Policies.

For period t, let $\Xi = \{\xi = (\xi_1, ..., \xi_{H_t}) : \sum_{h=1}^{H_t} \xi_h = 1\} \in [0,1]^{H_t}$ be the set of randomized actions over the arms. Then in the H_t-armed bandit problem, the policy used by each agent i is a function that maps from the current state space to the action space, denoted by $\sigma_i : \mathcal{Z}_{i,d} \to \Xi$. The choice made by policy σ_i is a random variable that depends on the agent's current state, denoted by $x_{i,d}(z_{i,d}) \in \{1, ..., H_t\}$. We use $\sigma_i(z_{i,d}, h)$ to denote the probability of agent i choosing arm h. Then the probability distribution of $x_{i,d}(z_{i,d})$ is as follows:

$$\mathbb{P}(x_{i,d}(z_{i,d}) = h) = \sigma_i(z_{i,d}, h), \forall\ z_{i,d} \in \mathcal{Z}_{i,d}. \tag{2.8}$$

Rewards.

We define the reward as the negative of each agent's electric bill. For each period t, after real-prices, $p_{d,h}^{RT}$ for $h \in \{1, ..., H_t\}$, are realized, each agent's reward can be determined according to its choice as follows:

$$U_d^k(x_{i,d}) = -\sum_{h=1}^{H_t} [l_{i,h}^b + l_{i,h}^f(x_{i,d})] p_{d,h}^{RT} \tag{2.9}$$

[3]The dimensions of the state can be beyond $2H_t$, such as $3H_t, 4H_t$ and so on, for recording more information associated with each arm. For example, for algorithm UCB1-Normal [74], H_t elements are used for recording the average squared rewards associated with each arm.

Dynamic population profile.

Since each agent's electric bill depends on the population-wide behaviors, we define the concept of population profile as the histogram of system-wide real-time net demand of each arm, denoted by $f_{d,h}$:

$$f_{d,h} = \frac{\mathcal{L}_{d,h}^{RT}}{\sum_{h=1}^{H_t} \mathcal{L}_{d,h}^{RT}} \tag{2.10}$$

Besides the decisions of agents, the *regeneration* of agents contributes to the dynamics of population profile. We assume that for each period t at each day d, agent i has a turnover probability $1-\beta$ to regenerate. When an agent is regenerated, its policy is changed arbitrarily, its type (i.e. non-thermal flexible load amount or HVAC settings) are changed, and its state variables z_i are re-initialized to 0's. An important role for the *regeneration* mechanism is to ensure that even when the system reaches a steady-state, the agents continue to learn without stopping exploring arms.

2.3.2 Control Strategies

Regrets and regret-minimizing policies.

If an agent assumes that the population profile $\boldsymbol{f} = (f_1, \ldots, f_{H_t})$ over arms is stationary across days, then the expected reward corresponding to choosing arm h for the agent will remain the same across days. As a result, we can ignore the index d in the reward expression in Equation 2.9 and define u^* as the highest expected value of u_i over $h \in \{1, ..., H_t\}$, where $u_i \in [0, 1]$ is the normalized reward as defined in Definition 2.3. As a result, the average regret of agent i for a fixed number of D days in the H_t-armed MAB games under σ_i is as follows:

$$R_i(D) := u^* - \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u_i(\sigma_{i,d}, \boldsymbol{f})]. \tag{2.11}$$

We consider the number of arms, H_t, to be fixed in our scenario. For the algorithms based on Upper Confidence Bound (UCB) [74] or Thompson Sampling [78], we have

logarithmic regret bounds, i.e. $R_i(D) < \frac{\Gamma(D)}{D}$, where $\Gamma(D) = \alpha \ln(D)$ for some constant α. Therefore, when agent i uses a regret-minimizing policy under stationary population profile, it plays approximately optimally. As pointed out in [59], the convergence to a steady state with stationary population profile does not depend on the specific policy chosen. Since D is the agent's regeneration life which is geometric with parameter β, we take expectation in Equation 2.11. In this case the policy is referred to as γ-*optimal* with $\gamma = \sum_{D=1}^{\infty}(1-\beta)\beta^{D-1}\frac{\Gamma(D)}{D}$ [59]:

$$u^* - \mathbb{E}_D[\frac{1}{D}\sum_{d=1}^{D}\sum_{h=1}^{H_t} u_i(h, \boldsymbol{f})\sigma_i(z_{i,d}, h)] < \gamma. \tag{2.12}$$

Convergence to steady state.

Let $\phi \in \Phi$ denote a joint distribution over all agents' state space, where Φ is the space of all Borel probability measures on the joint state space of all agents. We let $\mathbf{f}$ denote the dynamics of population profile $\{\boldsymbol{f}_1, \boldsymbol{f}_2, ...\}$ across days, where the number subscripts are the days. In [59], 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 $\mathbf{f}$ (i.e. population profile $boldsymbol f_i$'s are the same across days), which will influence the state transition, can yield a steady state distribution ϕ; second, based on the joint steady state distribution ϕ of all agents' states, the stationary population profile $\mathbf{f}$ can indeed emerge from the MAB-game. Strong theoretical results regarding MFSS have been shown in [59], including existence, uniqueness, and asymptomatic convergence to a MFSS when the population in the MAB-game approaches to infinity. The last property is especially useful in the context of decentralized demand response in energy markets, as the number of price-responsive agents can be very large. In the case of non-stationary population profile with finite agents, the average regret over D days is defined as follows:

$$\tilde{R}_i(D) := \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u_d^* - \sum_{d=1}^{D} u_{i,d}(\sigma_i, \boldsymbol{f}_d^n(\boldsymbol{\sigma}))], \tag{2.13}$$

where $\boldsymbol{f}_d^n(\boldsymbol{\sigma})$ denotes the population profile on day d under the n-agent system with policies $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$. The proposition below gives the bound of the regret in Equation 2.13.

Proposition 2.1 *When* $|u_i(f_h) - u_i(f'_h)| \leq L|f_h - f'_h|$ *for* $\forall h \in (1, \ldots, H)$ *and* $\beta(1+L) < 1$ *hold, there exists* $\epsilon > 0$ *such that*

$$\tilde{R}_i(D) \leq \gamma + \frac{2H\epsilon(1-\beta)}{\beta(1-\beta(1+L))} \tag{2.14}$$

where $\epsilon \to 0$ *as total population demand* $\mathcal{L}$ *approaches to infinity with an infinite number of agents playing regret-minimizing algorithms.*

Proof Since the proposition is for an individual agent in a sequence of games in period t across days, we ignore the index t. Now, we let $u_d^* := u_i(h_d^*, \boldsymbol{f}_d^n)$, where f_d^n represents the population profile in day d of an n-agent system, and h_d^* is the choice of sub-period that gives the best reward in day d. Meanwhile, we let u^* in (2.11) be $u_i(h^o, \boldsymbol{f})$, where f represents the stationary population profile of infinite agents and h^o is the underlying best choice. In addition, we let h_d^n be the choice of agent i under policy σ_i in day d under the n-agent system, and h_d be the choice under the infinite-agent system. Then we can write (2.13) as follows:

$$\begin{aligned}
\tilde{R}_i(D) &= \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h_d^*, \boldsymbol{f_d^n}) - \sum_{d=1}^{D} u(h_d^n, \boldsymbol{f_d^n})] \\
&= \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} \Big(u(h_d^*, \boldsymbol{f_d^n}) - u(h^o, \boldsymbol{f}) + u(h^o, \boldsymbol{f}) - u(h_d^n, \boldsymbol{f_d^n})\Big)]
\end{aligned} \tag{2.15}$$

where the first two items in (2.15) can be further written as:

$$\begin{aligned}
& u(h_d^*, \boldsymbol{f_d^n}) - u(h^o, \boldsymbol{f}) \\
&= u(h_d^*, \boldsymbol{f_d^n}) - u(h_d^*, \boldsymbol{f}) + u(h_d^*, \boldsymbol{f}) - u(h^o, \boldsymbol{f}) \\
&\leq L\sum_{h=1}^{H} |f_{d,h}^n - f_{d,h}| + u(h_d^*, \boldsymbol{f}) - u(h^o, \boldsymbol{f}) \quad , \\
&= L\sum_{h=1}^{H} |f_{d,h}^n - f_{d,h}|
\end{aligned} \tag{2.16}$$

where the second last step is by the condition $|u_i(f_h) - u_i(f'_h)| \leq L|f_h - f'_h|$. In the last step, $u(h^*_d, \boldsymbol{f})$ and $u(h^o, \boldsymbol{f})$ are both under the infinite-agent system. In our context, when there are infinite agents in the system and population profile are stationary, it must be that the demand population in each time-slot h are equal to each other. Otherwise, agents will learn the time-slots with lower demand and commit to it, which will make the demands equal again after a while of learning. Since the population demand are equal under stationary profile, the resulted rewards will be the same. Thus we can have $u(h^*_d, \boldsymbol{f}) - u(h^o, \boldsymbol{f}) = 0$.

In [59], it has been proved that for the given population n, there exists ϵ such that:

$$\mathbb{E}[\sup_h |f^n_{d,h} - f_{d,h}|] \leq \frac{\epsilon(1-\beta)}{L\beta(1-\beta(1+L))}, \tag{2.17}$$

where $\epsilon \to 0$ as $n \to \infty$. After we plug (2.17) into (2.15), we will have:

$$\tilde{R}_i(D) \leq \frac{H\epsilon(1-\beta)}{\beta(1-\beta(1+L))} + \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h^o, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f^n_d})]. \tag{2.18}$$

Now, use the same trick as in (2.16) to the last two terms in (2.15):

$$\begin{aligned}
&\frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h^o, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f^n_d})] \\
&= \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h^o, \boldsymbol{f}) - u(h_d, \boldsymbol{f}) + u(h_d, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f^n_d})] \\
&\leq \gamma + \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h_d, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f^n_d})] \quad \text{(by (2.11))} \\
&= \gamma + \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h_d, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f}) + u(h^n_d, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f^n_d})] \\
&= \gamma + \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D} u(h_d, \boldsymbol{f}) - u(h^n_d, \boldsymbol{f})] \\
&\leq \gamma + \frac{H\epsilon(1-\beta)}{\beta(1-\beta(1+L))}
\end{aligned} \tag{2.19}$$

Plugging (2.19) into (2.18) completes the proof. ■

In Proposition 2.1, we can see γ goes to 0 as D goes to infinity, the individual regret under non-stationary population profile will converge to a constant level. Further, ϵ

goes to 0 as n goes to infinity, the regret bound in the proposition will converge to the regret bound γ for stationary population profile.

In our context, the number of agents can be very large. If all agents' smart home have automation control devices coded with a regret-minimizing algorithm, combined with the strong results associated with MFSS, each agent will achieve a bounded regret as in (2.14). As the number of agents approaches infinity, each agent's regret bound converges to the result as in (2.12) and the system converges to its MFSS. 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 an agent i may have the incentive to deviate from its MFSS policy σ_i in order to maximize its discounted expected payoffs (this is so since regret-minimization may not be the same as discounted expected payoff optimization). As the the system converges to MFSS, the population profile will stabilize eventually, which gives stable load across days under real-time pricing.

2.4 Smoothness of Utility-maximization Game

When all agents' individual regrets are bounded, we want to know if the total social welfare of the system will converge to some efficient outcomes compared to the outcomes from centralized control. The discussion in this section is for consumers with non-thermal flexible load only as a starting point. Herein, the formulas concern a single game period $t \in \{1, \ldots, T\}$ across days[4]. To bound the total social welfare subject to decentralized "selfish" behaviors of the agents, we define the centralized optimum of the system as follows:

$$\mathrm{OPT}_{\mathcal{U}} = \max_{\boldsymbol{\sigma}} \sum_{i} u_i(\sigma_i, \boldsymbol{f}(\boldsymbol{\sigma})). \tag{2.20}$$

In a large class of games, termed *smooth games* by Roughgarden [79], no-regret learning dynamics converge to approximately optimal social welfare. In the following,

[4]As mentioned before, a day can be divided into T periods, i.e. T independent sequences of games here.

we introduce the concept of *smooth games*, and further prove that the MAB games' smoothness.

Definition 2.1 *Smooth game [79]: A utility maximization game is called* (λ, μ)*-smooth if for all policy profiles* $\boldsymbol{\sigma}$ *and* $\boldsymbol{\sigma}^*$*:*

$$\sum_i u_i(\sigma_i^*, \boldsymbol{f}(\boldsymbol{\sigma_{-i}})) \geq \lambda \sum_i u_i(\sigma_i^*, \boldsymbol{f}(\boldsymbol{\sigma^*})) - \mu \sum_i u_i(\sigma_i, \boldsymbol{f}(\boldsymbol{\sigma})). \tag{2.21}$$

It is relaxed-smooth if the inequality holds for any policy σ *and an optimal policy profile* σ^* *that achieves the* $OPT_{\mathcal{U}}$ *in (2.20).*

The price of anarchy (PoA) [80] measures the sub-optimality caused by agents' "selfish" behaviors. The PoA of a game is defined as the ratio between the worst outcome of a Nash Equilibrium and the optimal outcome. Therefore, a bound (a lower bound for utility maximization or an upper for cost minimization) on the PoA applies to every equilibrium and obviates the need to predict asingle outcome of "selfish" behaviors [79]. The PoA for utility-maximization games are defined as follows:

Definition 2.2 *PoA for utility-maximization games is defined as the ratio between the minimum social utility of a Nash Equilibrium and the maximum social utility among all action profiles, i.e.* $min_{\boldsymbol{\sigma} \in NE} \mathcal{U}(\boldsymbol{\sigma})/OPT_{\mathcal{U}}$*, where* $\mathcal{U}(\boldsymbol{\sigma})$ *is the total social utility of all agents.*

For utility maximization smooth games, the price of anarchy is at most $(1+\mu)/\lambda$, meaning that Nash equilibria of the game, as well as regret minimizing learning outcomes, in the limit have total social welfare at least $\frac{\lambda}{1+\mu}$OPT. Consider the utility maximization game for a large population of consumers in a stage game with H_t arms. To construct the smoothness property, we assume affine price functions $p_h^{RT}(\mathcal{L}_h^{RT}(\boldsymbol{\sigma})) = a_h \mathcal{L}_h^{RT}(\boldsymbol{\sigma}) + b_h$, where $a_h, b_h \geq 0$ for each time slot h and $\mathcal{L}_h^{RT}(\boldsymbol{\sigma})$ is the total real-time demand on time slot h. Using the same calculation in (2.9), the cost for agent i under the policy profile $\boldsymbol{\sigma}$ is as below:

$$cost_i(\boldsymbol{\sigma}) = \sum_{h=1}^{H} [a_h \mathcal{L}_h^{RT}(\boldsymbol{\sigma}) + b_h][l_{i,h}^b + l_{i,h}^f(x_{i,d})]. \tag{2.22}$$

The cost in (2.22) can be normalized with being divided by the maximum individual cost among all consumer agents. With the normalized cost $c_i(\boldsymbol{\sigma})$, we define the normalized utility as below

Definition 2.3 *For policy profile of n agents $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$, the total normalized social cost is*

$$\mathcal{C}(\boldsymbol{\sigma}) = \sum_{i=1}^{n} c_i(\boldsymbol{\sigma}), \tag{2.23}$$

where $c_i(\boldsymbol{\sigma}) \in [0, 1]$ is the normalized cost of agent i. Then the corresponding normalized utility of agent i is $u_i(\boldsymbol{\sigma}) = 1 - c_i(\boldsymbol{\sigma})$ which is in $[0, 1]$, and the total normalized utility is

$$\begin{aligned} \mathcal{U}(\boldsymbol{\sigma}) &= \sum_{i=1}^{n} u_i(\boldsymbol{\sigma}) \\ &= n - \mathcal{C}(\boldsymbol{\sigma}). \end{aligned} \tag{2.24}$$

Minimum social cost $OPT_{\mathcal{C}}$: the system cost that suffers efficiency loss due to the independently selfish behavior of the agents is lower bounded by

$$OPT_{\mathcal{C}} = \min_{\boldsymbol{\sigma}} \sum_{i=1}^{n} c_i(\boldsymbol{\sigma}), \tag{2.25}$$

and corresponding maximum social utility $OPT_{\mathcal{U}}$ in (2.20) is

$$OPT_{\mathcal{U}} = n - OPT_{\mathcal{C}}. \tag{2.26}$$

Now, we let $\overline{l^f}$ denote the maximum individual flexible demand among agents (where superscript f denotes flexible), and $\underline{L^b} := \min_h \mathcal{L}^{BD}(h)$ denote the minimum total base demand among all time slots in $(1, \ldots, H_t)$. Let $\rho = \overline{l^f}/\underline{L^b}$. Due to the small scale of each single agent, we can safely assume that $\overline{l^f} << \underline{L^b}$ with a large population of n smart home agents, and thus we have $\overline{l^f} << \underline{L^b} \leq \mathcal{L}_h^{RT}(\boldsymbol{\sigma})$ and $\rho << 1$ for any time slot h under any policy profile $\boldsymbol{\sigma}$. Aggregated base demands increase as the population increases, and consequently, ρ approaches 0.

Lemma 2.2 *For $\forall \alpha, \beta \in \mathcal{Z}^{\geq 0}$, we have [81]*

$$\beta(\alpha + 1) \leq \frac{5}{3}\beta^2 + \frac{1}{3}\alpha^2. \tag{2.27}$$

For $\forall \alpha, \beta \in \mathcal{Z}^{\geq 0}$ and $\forall a, b \geq 0$, by (2.27), we have [79]

$$\beta[a(\alpha+1)+b] \leq \frac{5}{3}(a\beta+b)\beta + \frac{1}{3}(a\alpha+b)\alpha. \tag{2.28}$$

Lemma 2.3 *For a large population n, and any arbitrary policy profiles $\boldsymbol{\sigma}^*$ and $\boldsymbol{\sigma}$, we have*

$$\sum_{i=1}^{n} c_i(\sigma_i^*, \boldsymbol{\sigma_{-i}}) \leq (\frac{5}{3}+\rho)\cdot \mathcal{C}(\boldsymbol{\sigma}^*) + \frac{1}{3}\cdot \mathcal{C}(\boldsymbol{\sigma}). \tag{2.29}$$

Proof By individual cost in Eq. (2.22), we have total system cost under policy profile $\boldsymbol{\sigma}$ as below

$$\begin{aligned} Cost(\boldsymbol{\sigma}) &= \sum_{i=1}^{n} cost_i(\boldsymbol{\sigma}) \\ &= \sum_{i=1}^{n}\sum_{h=1}^{H}(a_h\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+b_h)(l_{i,h}^b + l_{i,h}^f(x_{i,d})) \\ &= \sum_{h=1}^{H}(a_h\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+b_h)\sum_{i=1}^{n}(l_{i,h}^b + l_{i,h}^f(x_{i,d})) \\ &= \sum_{h=1}^{H}(a_h\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}) \end{aligned} \tag{2.30}$$

The total demand on each time slot h under policy profile $(\sigma_i^*, \boldsymbol{\sigma_{-i}})$ is at most $\overline{l^f}$ more than that under $\boldsymbol{\sigma}$, and this time slot h contributes to precisely real-time demand terms $\mathcal{L}_h^{RT}(\boldsymbol{\sigma})$ under $(\sigma_i^*, \boldsymbol{\sigma_{-i}})$. Then for any policy profile $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^*$, we have

$$\begin{aligned} \sum_{i=1}^{n} cost_i(\sigma_i^*, \boldsymbol{\sigma_{-i}}) &\leq \sum_{h=1}^{H}(a_h(\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+\overline{l^f})+b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}) \\ &= \sum_{h=1}^{H}(a_h(\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+1+\overline{l^f}-1)+b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*) \\ &= \sum_{h=1}^{H}[(a_h(\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+1)+b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*) + a_h(\overline{l^f}-1)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*)] \\ &\leq \sum_{h=1}^{H}\frac{5}{3}(a_h\mathcal{L}_h^{RT}(\boldsymbol{\sigma})+b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}) + \sum_{h=1}^{H}\frac{1}{3}(a_h\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*)+b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{h=1}^{H} a_h(\overline{l^f} - 1)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*) \quad (\textit{by Lemma 2.2}) \\
&\leq \frac{5}{3}Cost(\boldsymbol{\sigma}^*) + \frac{1}{3}Cost(\boldsymbol{\sigma}) + \sum_{h=1}^{H} a_h \rho(\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*))^2 \\
&\leq \frac{5}{3}Cost(\boldsymbol{\sigma}^*) + \frac{1}{3}Cost(\boldsymbol{\sigma}) + \rho \sum_{h=1}^{H} (a_h \mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*) + b_h)\mathcal{L}_h^{RT}(\boldsymbol{\sigma}^*) \\
&= (\frac{5}{3} + \rho)Cost(\boldsymbol{\sigma}^*) + \frac{1}{3}Cost(\boldsymbol{\sigma}). \qquad (2.31)
\end{aligned}
$$

Then normalizing on both sides completes the proof. ■

Proposition 2.4 *Consider an optimal policy profile $\boldsymbol{\sigma}^*$ that results in $OPT_{\mathcal{C}}$ and $OPT_{\mathcal{U}}$. Let γ denote the ratio of n to $OPT_{\mathcal{C}}$, i.e. $\gamma = \frac{n}{OPT_{\mathcal{C}}} \geq 1$. For any η that satisfies*

$$\eta \geq \frac{\rho + 1}{\gamma - 1}, \qquad (2.32)$$

by Lemma 2.3 the utility maximization game has relaxed-smoothness of $(\lambda = \frac{2}{3} - \eta, \mu = -\frac{1}{3})$ for any arbitrary policy profile $\boldsymbol{\sigma}$ as below

$$\sum_{i=1}^{n} u_i(\sigma_i^*, \boldsymbol{\sigma_{-i}}) \geq (\frac{2}{3} - \eta) \cdot OPT_{\mathcal{U}} + \frac{1}{3} \cdot \mathcal{U}(\boldsymbol{\sigma}). \qquad (2.33)$$

Proof Consider any policy profile $\boldsymbol{\sigma}$ and an optimal profile $\boldsymbol{\sigma}^*$ that achieves $OPT_{\mathcal{C}}$ and $OPT_{\mathcal{U}}$. Then we can have

$$
\begin{aligned}
\sum_{i=1}^{n} u_i(\sigma_i^*, \boldsymbol{\sigma_{-i}}) &= \sum_{i=1}^{n} [1 - c_i(\sigma_i^*, \boldsymbol{\sigma_{-i}})] \quad (\text{by Definition 2.3}) \\
&= n - \sum_{i=1}^{n} c_i(\sigma_i^*, \boldsymbol{\sigma_{-i}}) \\
&\geq n - (\frac{5}{3} + \rho) \cdot \mathcal{C}(\boldsymbol{\sigma}^*) - \frac{1}{3} \cdot \mathcal{C}(\boldsymbol{\sigma}) \quad (\text{by Lemma 2.3}) \\
&= (\frac{5}{3} + \rho) \cdot [n - \mathcal{C}(\boldsymbol{\sigma}^*)] + \frac{1}{3} \cdot [n - \mathcal{C}(\boldsymbol{\sigma})] + [1 - (\frac{5}{3} + \rho) - \frac{1}{3}] \cdot n \\
&= (\frac{5}{3} + \rho) \cdot \mathcal{U}(\boldsymbol{\sigma}^*) + \frac{1}{3} \cdot \mathcal{U}(\boldsymbol{\sigma}) - (\rho + 1) \cdot n \quad (\text{by Eq. (2.24)}).
\end{aligned} \qquad (2.34)
$$

By Eq. (2.32), we can have

$$(\rho + 1 + \eta) \cdot \mathcal{C}(\boldsymbol{\sigma}^*) - \eta \cdot n \leq 0. \tag{2.35}$$

With adding Eq. (2.35) to the last line in Eq. (2.34), we have

$$\begin{aligned}
\sum_{i=1}^{n} u_i(\sigma_i^*, \boldsymbol{\sigma_{-i}}) &\geq (\frac{5}{3} + \rho) \cdot \mathcal{U}(\boldsymbol{\sigma}^*) + \frac{1}{3} \cdot \mathcal{U}(\boldsymbol{\sigma}) - (\rho + 1) \cdot n + (\rho + 1 + \eta) \cdot \mathcal{C}(\boldsymbol{\sigma}^*) - \eta \cdot n \\
&= (\frac{5}{3} + \rho) \cdot \mathcal{U}(\boldsymbol{\sigma}^*) + \frac{1}{3} \cdot \mathcal{U}(\boldsymbol{\sigma}) - (\rho + 1 + \eta) \cdot (n - \mathcal{C}(\boldsymbol{\sigma}^*)) \\
&= (\frac{5}{3} + \rho) \cdot \mathcal{U}(\boldsymbol{\sigma}^*) + \frac{1}{3} \cdot \mathcal{U}(\boldsymbol{\sigma}) - (\rho + 1 + \eta) \cdot \mathcal{U}(\boldsymbol{\sigma}^*) \ \text{(by Eq. (2.24))} \\
&= (\frac{2}{3} - \eta) \cdot \mathcal{U}(\boldsymbol{\sigma}^*) + \frac{1}{3} \cdot \mathcal{U}(\boldsymbol{\sigma}).
\end{aligned} \tag{2.36}$$

Substituting $\mathcal{U}(\boldsymbol{\sigma}^*)$ in (2.34) by $OPT_{\mathcal{U}}$ completes the proof. ■

By Proposition 2.4, the utility-maximization game is a $(\lambda, \mu) - smooth$ game with $\lambda = \frac{2}{3} - \eta$ and $\mu = -\frac{1}{3}$, and the corresponding POA is $\frac{\lambda}{1+\mu} = 1 - \frac{3}{2} \cdot \eta$. When η reaches at the lower bound in Eq. (2.32), the tightest POA bound is achieved as $1 - \frac{3}{2} \cdot \frac{\rho+1}{\gamma-1}$.

One thing that needs attention here is the POA bound is meaningful (i.e. $\frac{\lambda}{1+\mu} > 0$) only when $\gamma > \frac{5+3\cdot\rho}{2}$, where $\rho << 1$. A very simple example here. Consider $\gamma = 2$, which means $OPT_{\mathcal{C}} = \frac{1}{2} \cdot n$, then $1 - \frac{3}{2} \cdot \frac{\rho+1}{\gamma-1} = 1 - \frac{3}{2} \cdot (1 + \rho) < 0$. In another example considering the extreme case in which base demands are 0 and all n agents have the same amount of flexible demand. Then for a game of H time slots, the worst action profile is that the whole demand population choose the same time slot, while $OPT_{\mathcal{C}}$ can be achieved when there are $\frac{1}{H}$ of the demand population on each time slot[5]. In this case, the price (affine and linear) ratio is very close to H to 1, and thus $OPT_{\mathcal{C}}$ is close to $\frac{n}{H}$. Then $\gamma = H$ and POA is close to $1 - \frac{3}{2} \cdot \frac{1}{H-1}$. For example, when $H = 4$, POA is very close to $\frac{1}{2}$.

Therefore, when the system's $OPT_{\mathcal{C}}$ is much smaller than n, the ratio γ is much larger than 1. Then the resulting η will be small and the POA bound achieved by the game's utility-maximization smoothness will be tight.

[5]In this case, we assume that the supply curve does not change across time slots $(1, \ldots, H_t)$

Proposition 2.5 *If all n agents use $\gamma-$ optimal regret-minimizing policies with $\Gamma(D) = \alpha ln(D)$ in a (λ, μ)-smooth MAB-game, with parameter $\tilde{\epsilon} = 2H\epsilon(1-\beta)/\beta$, then for the action profile drawn from $\boldsymbol{\sigma}$,*

$$\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{U}_d(\boldsymbol{\sigma})] \geq \frac{\lambda}{1+\mu}OPT - \frac{n}{1+\mu}(\gamma+\tilde{\epsilon}) \tag{2.37}$$

where $\sum_d \mathbb{E}[\mathcal{U}_d(\boldsymbol{\sigma})] = \sum_d \sum_i \mathbb{E}[u_{i,d}(\sigma_i, \boldsymbol{f}(\boldsymbol{\sigma}))]$, $OPT = \frac{1}{D}\sum_d OPT_{\mathcal{U},d}$, and $(\lambda, \mu) = (\frac{2}{3}-\eta, -\frac{1}{3})$.

Proof By adding up the bounded regrets in Proposition 2.1, we have:

$$\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{U}_d(\boldsymbol{\sigma})] \geq \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D}\sum_{i=1}^{n}u_d^*] - n(\gamma+\tilde{\epsilon}) \tag{2.38}$$

Then we conduct summation across D days over the inequality in Definition 3.3, we can have:

$$\mu\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{U}_d(\boldsymbol{\sigma})] \geq \lambda\mathrm{OPT} - \frac{1}{D}\mathbb{E}[\sum_{d=1}^{D}\sum_{i=1}^{n}u_{i,d}(\sigma_i^*, \boldsymbol{f}(\boldsymbol{\sigma_{-i}}))] \tag{2.39}$$

Add (2.38) to (2.39), we have:

$$(1+\mu)\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{U}_d(\boldsymbol{\sigma})] \geq \lambda\mathrm{OPT} - n(\gamma+\tilde{\epsilon}) + \Delta \tag{2.40}$$

where

$$\Delta = \frac{1}{D}\sum_{d=1}^{D}\Big(\mathbb{E}[\sum_{d=1}^{D}\sum_{i=1}^{n}u_d^*] - \mathbb{E}[\sum_{d=1}^{D}\sum_{i=1}^{n}u_{i,d}(\sigma_i^*, \boldsymbol{f}(\boldsymbol{\sigma_{-i}}))]\Big) \tag{2.41}$$

Since u_d^* is the best choice in each day, we always have $u_d^* \geq u_{i,d}(\sigma_i^*, \boldsymbol{f}(\boldsymbol{\sigma_{-i}}))$. Therefore, $\Delta \geq 0$, which lead (2.40) to as follows:

$$\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{U}_d(\boldsymbol{\sigma})] \geq \frac{\lambda}{1+\mu}\mathrm{OPT} - \frac{n}{1+\mu}(\gamma+\tilde{\epsilon}) \tag{2.42}$$

■

In (2.37), given any population n, as the games repeatedly go on, γ goes to 0. As discussed before, due to the convexity of generation costs, $OPT_{\mathcal{U}}$ is achieved when

demands are equal among time slots. Therefore, each individual agent's normalized utility is a constant under $OPT_{\mathcal{U}}$ regardless of the population. Since $OPT_{\mathcal{U}}$ is the total normalized utility (i.e. $\sum_i u_i$), $OPT_{\mathcal{U}}$ is $O(n)$, and thus OPT is $O(n)$. Further, $\tilde{\epsilon} \to 0$ as $n \to \infty$. Therefore, as the population increases, OPT increases in order $O(n)$ and $\tilde{\epsilon}$ goes to 0, which means OPT grows faster than $n \cdot \tilde{\epsilon}$, and thus the system dynamics approximately converge to efficient outcomes guaranteed by the price of anarchy.

3. DEMAND RESPONSE WITH LOW APPROXIMATE REGRET LEARNING IN GAMES

In this chapter, we propose a game-theoretic framework in which all smart home agents decide their flexible demand consumption strategy in a decentralized way by playing a learning algorithm that guarantees small regret. Such learning utilize full information feedback instead of only looking at the actions played. In a (λ, μ) cost-minimization smooth game, the price of anarchy (PoA) is $\lambda/(1-\mu)$ [79]. While the environment is constantly changing due to each agent's choice of strategy, such decentralized no-regret dynamics are guaranteed to converge to the system's PoA if all agents play low approximate regret learning algorithms, and the speed at which the game outcomes converge to the approximate optimality is governed by individual agents' regret bounds [61]. Through the learning in games, the regrets of individual agents and the total social cost of the overall system are both bounded.

The rest of the chapter is structured as follows. In Section 3.1, we will describe smart home agents with flexible loads conducting learning under RTP scheme for having approximate no-regret costs. Then in Section 3.2, numerical simulations are presented by comparing the learning approach proposed in this work with the decentralized demand management approaches described in Chapter 2.

3.1 Low Approximate Regret Learning in Games

3.1.1 Learning Dynamics

Each agent has an action space $\boldsymbol{S_i}$ with cardinality H[1], i.e. $|\boldsymbol{S_i}| = H$, and a cost function $\boldsymbol{cost_i} : \boldsymbol{S_1} \times \cdots \times \boldsymbol{S_n} \to [0, 1]$ that maps an action profile $\boldsymbol{s} = (s_1, \ldots, s_n)$

[1]The action space of each agent doesn't have to be of the same. Herein, for the simplicity of illustration, we only assume that the agents' action space have the same cardinality.

to the normalized cost of agent i. Further, we let $\boldsymbol{w} = (\boldsymbol{w_1}, \ldots, \boldsymbol{w_n})$ denote a list of probability distribution over all agents' action space, where $\boldsymbol{w_i} \in \Delta(\boldsymbol{S_i})$ and $w_{i,x}$ is the probability of action $x \in \boldsymbol{S_i}$. Time period T is repeated for D days, on each day d each agent i picks a probability distribution $\boldsymbol{w_i^d}$ over actions and draws their action s_i^d from this distribution, i.e. choose a time slot to consume their flexible demand.

After each day, each agent i receives realized feedback in terms of real-time ex post prices of period T, and observes the costs they would have received had they chosen any action $x \in \boldsymbol{S_i}$ given the realized actions taken by the other agents. The underlying assumption here is that the shifting of any single agent's choice does not affect the real-time prices due to its small scale. Specifically, we let $c_{i,x}^d = cost_i(x, \boldsymbol{s_{-i}^d})$, where $\boldsymbol{s_{-i}^d}$ is the set of realized strategies of all but the i^{th} agent on day d, and let $\boldsymbol{c_i^d} = (c_{i,x}^d)_{x \in \boldsymbol{S_i}}$. Accordingly, the expected cost of agent i in period T on day d conditioned on the other agents' realized actions is the inner product $\langle \boldsymbol{w_i^d}, \boldsymbol{c_i^d} \rangle$.

Learning algorithms that satisfy the property *Low Approximate Regret (LAR)* [61], defined in Definition 3.1, can give the agents the cumulative cost multiplicatively (i.e. multiplied by a constant) approximate to the cost of the best action (i..e. the minimum cost) they could have chosen in hindsight.

Definition 3.1 *Low Approximate Regret (LAR) [61]: for some parameter $\epsilon > 0$ and function $R(H, D)$, a learning algorithm for agent i satisfies the LAR property if for all action distributions $\boldsymbol{w^*} \in \Delta(\boldsymbol{S_i})$,*

$$(1-\epsilon)\sum_{d=1}^{D}\langle \boldsymbol{w_i^d}, \boldsymbol{c_i^d} \rangle \leq \sum_{d=1}^{D}\langle \boldsymbol{w^*}, \boldsymbol{c_i^d} \rangle + \frac{R(H,D)}{\epsilon}. \tag{3.1}$$

A learning algorithm has LAR property against shifting experts if for all action distribution sequences $\boldsymbol{w^{*,d}} \in \Delta(\boldsymbol{S_i})$ for $d \in \{1, \ldots, D\}$, it satisfies the inequality below,

$$(1-\epsilon)\sum_{d=1}^{D}\langle \boldsymbol{w_i^d}, \boldsymbol{c_i^d} \rangle \leq \sum_{d=1}^{D}\langle \boldsymbol{w^{*,d}}, \boldsymbol{c_i^d} \rangle + (1+K)\frac{R(H,D)}{\epsilon}, \tag{3.2}$$

where $K = \sum_{d=2}^{D} ||\boldsymbol{w^{*,d}} - \boldsymbol{w^{*,d-1}}||_1$ is the number of action distribution shifts.

The arbitrariness of the distribution $\boldsymbol{w}^*$ in Definition 3.1 includes the best action in hindsight. For instance of LAR algorithms, the *Hedge* algorithm [82] achieves LAR with $R(H, D) = log(H)$ by using any fixed $\epsilon > 0$ as learning rate, *Optimistic Hedge* [83] satisfies LAR with $R(H, D) = 8log(H)$, and *Noisy Hedge* [61] with learning rate ϵ has $R(H, D) = 2log(H \cdot D)$ for achieving LAR against shifting experts. The learning procedures of *Noisy Hedge* are presented in Algorithm 1. In our context, agents can use such algorithms to choose a time slot to consume its flexible demand in time period T on each day, and have LAR compared to the realized cheapest time slot.

Algorithm 1 Noisy Hedge

Initialization:

1. Fix $\theta \in [0, 1]$, $\epsilon > 0$.
2. Let $\boldsymbol{\pi}$ be the uniform distribution over $[H]$.
3. Let $\boldsymbol{w_1} = \boldsymbol{\pi}$.

Learning:

For $d = 1, 2, \ldots$:

1. $\tilde{w}^h_{d+1} = w^h_d \cdot e^{-\epsilon c^h_d}$;
2. $g^h_{d+1} = \tilde{w}^h_{d+1} / \sum_{j\in[H]} \tilde{w}^j_{d+1}$;
3. $\boldsymbol{w}_{d+1} = (1 - \theta) \cdot \boldsymbol{g}_{d+1} + \theta \cdot \boldsymbol{\pi}$.

3.1.2 Cost-minimization Smooth Game

Traditional learning dynamics has been shown to converge to approximately optimal social welfare in smooth games [79]. In [61], we see fast convergence of learning dynamics to approximate PoA of the system when LAR is coupled with smooth game. For electricity price p that is an affine function of total demand $\mathcal{L}$, i.e. $p_h(\mathcal{L}_h) = a_h\mathcal{L}_h + b_h$ with $a_h, b_h \geq 0$ for $h \in \{1, \ldots, H\}$, the cost-minimization game has smoothness. Use the definition in Chapter 2, we have

Definition 3.2 *Social cost: for an action profile* $\boldsymbol{s} = (s_1, \ldots, s_n)$*, the social cost is*

$$\mathcal{C}(\boldsymbol{s}) = \sum_{i=1}^{n} cost_i(\boldsymbol{s}). \tag{3.3}$$

Minimum social cost $OPT_{\mathcal{C}}$*: the system cost that suffers efficiency loss due to the independently selfish behavior of the agents is bounded by*

$$OPT_{\mathcal{C}} = \min_{\boldsymbol{s}} \sum_{i=1}^{n} cost_i(\boldsymbol{s}). \tag{3.4}$$

Definition 3.3 *Cost-minimization smooth game [79]: a cost-minimization game is* $(\lambda, \mu) - smooth$*, with* $\lambda > 0$ *and* $\mu < 1$*, if for every two action profiles* $\boldsymbol{s}$ *and* $\boldsymbol{s^*}$*,*

$$\sum_{i=1}^{n} cost_i(s_i^*, \boldsymbol{s_{-i}}) \leq \lambda \cdot \mathcal{C}(\boldsymbol{s^*}) + \mu \cdot \mathcal{C}(\boldsymbol{s}). \tag{3.5}$$

Definition 3.4 *PoA for cost-minimization games is defined as the ratio between the maximum social cost of a Nash Equilibrium and the minimum social cost among all action profiles, i.e.* $max_{\boldsymbol{s} \in NE} \mathcal{C}(\boldsymbol{s})/OPT_{\mathcal{C}}$*, where* $\mathcal{C}(\boldsymbol{s})$ *is the total social cost of all agents.*

In a $(\lambda, \mu) - smooth$ cost-minimization game, the cost imposed on any agent by the actions of the others is bounded, and the PoA is at most $\lambda/(1-\mu)$, i.e. each of its Nash equilibria and no-regret learning outcomes in the limit have social cost at most $\frac{\lambda}{1-\mu} OPT$ [61,79]. In Proposition 3.1 below, we show that the cost-minimization game formed by a large population of smart home agents is $(5/3 + \rho, 1/3) - smooth$, where $\rho << 1$ is the ratio of the maximum flexible demand among agents to the minimum total base demand among time slots in period T, i.e. $\rho = \max_i l_i^f / \min_h L_h^b$ with $\mathcal{L}_h^b = \sum_{i=1}^{n} l_{i,h}^b$, exactly the same as defined in Chapter 2.

Proposition 3.1 *By Lemma 2.3, the cost-minimization game formed by a set of n smart home agents has smoothness of* $(5/3 + \rho, 1/3) - smooth$ *with affine real-time price functions, where n is a reasonable large number and* $\rho << 1$*.*

Proposition 3.1 implies an upper bound of $(5+3\rho)/2$ on the PoA of Nash equilibria, as well as no-regret learning outcomes, of the cost-minimization game formed by smart home agents. As the number of agents n increases, the ratio ρ goes to 0, and thus the PoA's upper bound goes to 5/2. The individual cost can be normalized to $[0, 1]$ for agents learning with LAR algorithms described in Section 3.1.1. In the following, we show that as every agent conducts learning with LAR algorithms, the cost-minimization game converges to the PoA bound approximately.

3.1.3 Learning with Full Information Feedback

In period T across D days, as agents select a time slot to consume their flexible demand, they can always observe the real-time ex post prices of all time slots after each day, even for the time slots they had not chosen. Therefore, we consider agents receive full information feedback regarding real-time ex post prices which reflect the demand situations of time slots because of the RTP scheme. Further, as mentioned in Section 3.1.1, the form of feedback is referred to as realized feedback since for agent i it only depends on the realized actions $\boldsymbol{s}_{-i}^{\boldsymbol{d}}$ sampled by the opponents from their distributions $\boldsymbol{w}_{-i}^{\boldsymbol{d}}$. Therefore, at the end of each round (or day), agents observe their own entire cost vector $\boldsymbol{c}_{\boldsymbol{i}}^{\boldsymbol{d}} = cost_i(x, \boldsymbol{s}_{-i}^{\boldsymbol{d}})_{x \in \boldsymbol{S_i}}$ through the real-time price ex post prices, but are not aware of other agents' costs in the game. Based on the realized full information feedback on round d (day d in our context) and a time-invariant cost-minimization action profile $\boldsymbol{s^*}$ that achieves OPT, we define the hypothetical additional cost for agent i had it used the action s_i^* instead of s_i^d as below

$$r_i(\boldsymbol{s^d}) = cost_i(s_i^*, \boldsymbol{s}_{-i}^{\boldsymbol{d}}) - cost_i(\boldsymbol{s^d}), \tag{3.6}$$

where $r_i(\boldsymbol{s^d})$ can be positive or negative for an arbitrary action profile $\boldsymbol{s^d}$, and be non-negative when $\boldsymbol{s^d}$ is a Nash equilibrium. In a $(\lambda, \mu) - smooth$ game as defined in Definition 3.3, using the smoothness property in (3.5) to $r_i(\boldsymbol{s^d})$ in (3.6), we have

$$\mathcal{C}(\boldsymbol{s^d}) \leq \frac{\lambda}{1-\mu} \cdot \mathcal{C}(\boldsymbol{s^*}) - \frac{1}{1-\mu} \sum_{i=1}^{n} r_i(\boldsymbol{s^d}). \tag{3.7}$$

Consider a sequence of action profiles $\boldsymbol{s^d}$ for $d = 1, \ldots, D$, averaging (3.7) over D rounds (or D days in our context) and rearranging the double summation terms give us

$$\frac{1}{D} \sum_{d=1}^{D} \mathcal{C}(\boldsymbol{s^d}) \leq \frac{\lambda}{1-\mu} \cdot \mathcal{C}(\boldsymbol{s^*}) - \frac{1}{1-\mu} \sum_{i=1}^{n} (\frac{1}{D} \sum_{d=1}^{D} r_i(\boldsymbol{s^d})). \tag{3.8}$$

If every agent i experiences vanishing average external regret in the outcome sequence $\boldsymbol{s^d}$ for $d = 1, \ldots, D$, meaning that the additional cost $[\sum_{d=1}^{D} r_i(\boldsymbol{s^d})]/D$ is bounded above by a $o(1)$ term that goes to 0 as $D \to \infty$, then we can have [79]

$$\frac{1}{D} \sum_{d=1}^{D} \mathcal{C}(\boldsymbol{s^d}) \leq [\frac{\lambda}{1-\mu} + o(1)] \cdot \mathcal{C}(\boldsymbol{s^*}). \tag{3.9}$$

The inequality above indicates the convergence to PoA of the smooth game when every agent i has vanishing average external regret. The *LAR* defined in Definition 3.1 relaxes the regret bound and thus relaxes the quality of approximation from the bound guaranteed by smoothness [61]. Simple "off-the-shell" *LAR* algorithms that can obtain fast convergence are ubiquitous, such as *Hedge* and its varaints mentioned above, and they only ask for feedback based on realized outcomes instead of expected outcomes. If all agents use *LAR* algorithms for deciding when to consume their flexible demand in the $(\lambda, \mu) - smooth$ game we have shown in Section 3.1.2, efficient outcomes established in [61] can be achieved by the overall system.

Proposition 3.2 *[61] If all agents use LAR algorithms satisfying inequality (3.1) with parameter $\epsilon > 0$ and function $R(H, D)$ in a $(\lambda, \mu) - smooth$ game, then for the action profiles $\boldsymbol{s^d}$ drawn on round d from the corresponding actions of the agents by the LAR algorithms, we have,*

$$\frac{1}{D} \sum_{d=1}^{D} \mathbb{E}[\mathcal{C}(\boldsymbol{s^d})] \leq \frac{\lambda}{1-\mu-\epsilon} OPT_{\mathcal{C}} + \frac{n}{D} \cdot \frac{1}{1-\mu-\epsilon} \cdot \frac{R(H, D)}{\epsilon}. \tag{3.10}$$

The proof for Proposition 3.2 is starightforward by using (3.1) and (3.5) with $\boldsymbol{s}^*$ resulted by $\boldsymbol{w}^*$. When $\epsilon << (1-\mu)$, the approximation factor $\lambda/(1-\mu-\epsilon)$ is very close to the PoA $\lambda/(1-\mu)$. Therefore, for $R(H,D)/D$ bounded above by $o(1)$, LAR learning dynamics quickly converges to outcomes with social cost arbitrarily close to the social cost guaranteed for Nash equiliria by the PoA. Moreover, agents can experience fast convergence with high probability with LAR learning.

Proposition 3.3 *[61] If all agents use LAR algorithms satisfying inequality (3.1) with parameter $\epsilon > 0$ and function $R(H,D)$ in a $(\lambda,\mu)-$smooth game, then for the action profiles $\boldsymbol{s}^{\boldsymbol{d}}$ drawn on round d from the corresponding actions of the agents by the LAR algorithms and $\gamma = 2\epsilon/(1+\epsilon)$, with probability at least $1-\delta$ for $\forall\delta > 0$, we have,*

$$\begin{aligned}\frac{1}{D}\sum_{d=1}^{D}\mathcal{C}(\boldsymbol{s}^{\boldsymbol{d}}) \leq \frac{\lambda}{1-\mu-\gamma}OPT_{\mathcal{C}} + \\ \frac{n}{D}\cdot\frac{1}{1-\mu-\gamma}\cdot[\frac{4R(H,D)}{\gamma}+\frac{12log(n\ log_2(D)/\delta)}{\gamma}].\end{aligned} \tag{3.11}$$

If $R(H,D) = O(log(H))$, then with probability at least $1-\delta$ for $\forall\delta > 0$, we have,

$$\begin{aligned}\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{C}(\boldsymbol{s}^{\boldsymbol{d}})] \leq \frac{\lambda}{1-\mu-\epsilon}OPT_{\mathcal{C}} + \\ \frac{n}{D}\cdot\frac{1}{1-\mu-\epsilon}\cdot[\frac{O(log(H))}{\epsilon}+\frac{O(log(n\ log_2(D)/\delta))}{\epsilon}].\end{aligned} \tag{3.12}$$

Smart home agents in the cost-minimization smooth game established in Section 3.1.2 can use online learning algorithms of order $O(log(H))$ like *Hedge* to achieve LAR when they decide when to consume its flexible demand, and the overall system can thus quickly converge to the PoA bound $(5+3\rho)/2$ approximately.

3.1.4 Dynamic Population by Regeneration

In previous sections, the repeated games are conducted among the exact same set of agents whose flexible demand in period T do not vary across D days. To mimic the dynamic flexible demand in reality more closely, we introduce the *regeneration*

mechanism [59] in this section. Same as the MAB-game model, at each round (or day) d, every agent i is regenerated with a probability β which is termed as *regeneration rate*. When an agent is regenerated, its flexible demand amount is changed and thus its cost function is updated accordingly. As described in Section 3.1.3, the agents receive realized full information feedback about their cost vector $\boldsymbol{c_i^d} = cost_i^d(x, \boldsymbol{s_{-i}^d})_{x \in \boldsymbol{S_i^d}}$ through real-time ex post prices. The *regeneration* mechanism can account for two situations, one is that an agent changes its flexible demand amount in period T, the other is that an existing agent leaves while a new agent joins the game.

With dynamic population, the benchmark $OPT_{\mathcal{C}}$ defined in (3.4) is not time-invariant anymore. Instead, there existing a sequence of shifting optimal experts $\boldsymbol{s^{*,d}}$ achieving minimum social cost $OPT_{\mathcal{C},d}$ due to the *regeneration rate* β. Therefore, agents need to have low regret against the shifting experts $s_i^{*,d}$ to guarantee low overall social cost using the smoothness property discussed in Section 3.1.2 and 3.1.3. The work in [61] and [84] show that if the following three conditions can be met, the PoA still can be achieved approximately. Concretely, if 1. there exists a relatively stable sequence of action profiles whose outcomes at each round approximate $OPT_{\mathcal{C},d}$ by a factor of η; 2. all agents using low adaptive regret algorithms [85]; and 3. the *regeneration rate* β is bounded above by a function of ϵ, then at least a $\eta\lambda/(1-\mu-\epsilon)$ fraction of the optimal outcome is guaranteed. Further, in [61], it has shown that in dynamic population smooth games, if every agent uses an online learning algorithm that achieves LAR against shifting experts as in (3.2), the overall system can converge to the fraction result $\eta\lambda/(1-\mu-\epsilon)$ approximately, which extend the efficient outcome in Proposition 3.2 to the proposition below.

Proposition 3.4 *[61] If all agents use LAR algorithms satisfying inequality (3.2) with parameter $\epsilon > 0$ and function $R(H, D)$ in a dynamic population $(\lambda, \mu) - smooth$*

game, then for the action profiles $\boldsymbol{s^d}$ *drawn on round* d *from the corresponding actions of the agents by the LAR algorithms, we have,*

$$\frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\mathcal{C}(\boldsymbol{s^d})] \leq \frac{1}{T}\cdot\frac{\lambda\cdot\eta}{1-\mu-\epsilon}\sum_{d=1}^{D}\mathbb{E}[OPT_{\mathcal{C},d}] + \frac{n+\mathbb{E}[\sum_{i=1}^{n}K_i]}{D}\cdot\frac{1}{1-\mu-\epsilon}\cdot\frac{R(H,D)}{\epsilon}, \tag{3.13}$$

where K_i *denotes the number of action changes of agent* i *in the stable sequence* $s_i^{*,1:D}$*, and* $\boldsymbol{s^{*,1:D}}$ *is near-optimal with* $\sum_{i=1}^{n} cost_i^d(\boldsymbol{s^{*,d}}) \leq \eta\cdot OPT_{\mathcal{C},d}$*.*

To approximate the bound by PoA more closely, in (3.13), algorithms with $R(H,D)$ of lower order in D can allow higher *regeneration rate* β with higher $\mathbb{E}[\sum_{i=1}^{n}K_i]$. As mentioned in Section 3.1.1, *Noisy Hedge* with learning rate ϵ can achieve LAR against shifting experts with $R(H,D) = 2log(H\cdot D)$, thus thus satisfies (3.13). In the next section, numerical simulations are conducted in which agents use *Noisy Hedge* with *regeneration.*

3.2 Numerical Simulations

3.2.1 Simulation Data

ISO-NE Test System.

An 8-Zone test system based on the power system and electricity wholesale market organized by New England Independent System Operator (ISO-NE) is developed in [66, 86], consisting of zones Maine (ME), Vermont (VT), New Hampshire (NH), Northeastern Massachusetts & Boston (NEMA & BOST), West-central Massachusetts (WCMA), Southeastern Massachusetts (SEMA), Connecticut (CT), and Rhode Island (RI). As shown in Figure 3.1, there are 76 thermal generation units (represented by red spots) distributed around 8 zones (represented by green spots). The 76 thermal generation units are selected for inclusion in the benchmark generation mix, each treated as an independent generator, which have a combined installed

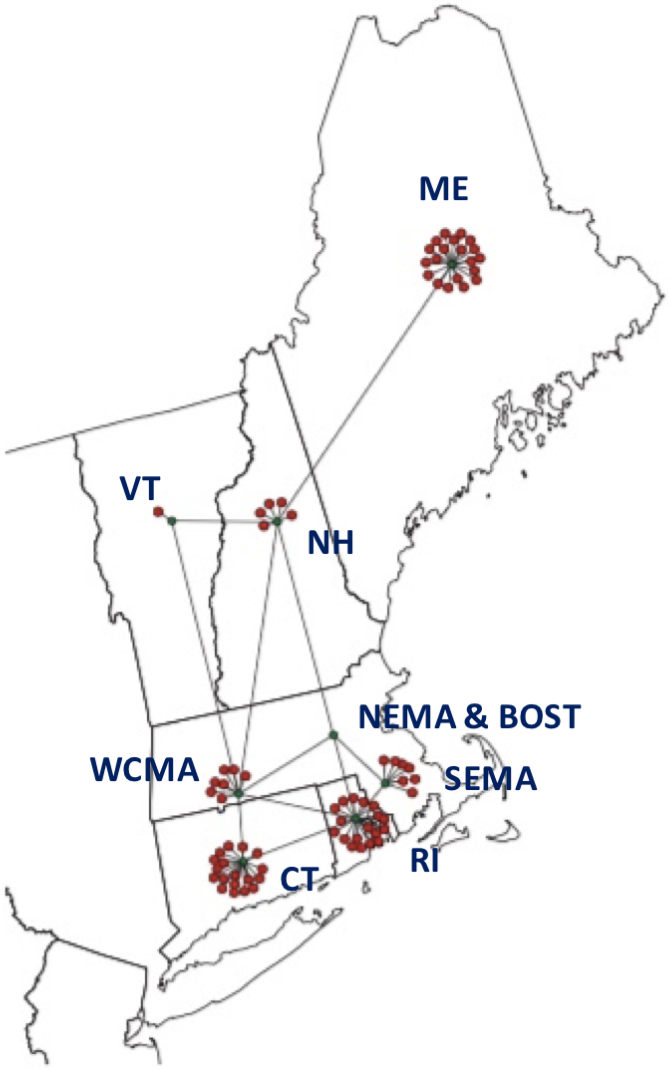

Figure 3.1.: Transmission grid for the 8-Zone ISO-NE Test System and generation units distribution [86].

generation capacity of $23,100MW$ and account for about 70% of the total actual ISO-NE capacity. Each generation unit j has fuel type among Coal (BIT), Coal (SUB), Fuel Oil, Natural Gas, and Nuclear, and has a quadratic dispatch cost function ($\$/h$) as

$$C_j(G_j) = a_j(G_j)^2 + b_j G_j + c_j, \tag{3.14}$$

where $G_j(MW)$ denotes generator j's power output. The cost coefficients a_j and b_j in (3.14) are all non-negative as shown in Table 3.1, and thus electricity price by marginal cost is an affine function of demand due to supply-demand balance. More detailed data of the generators can be found in [66].

Table 3.1.: Generation cost coefficients by fuel type [66].

Fuel type	a ($/MWˆ2h)	b ($/MWh)	c ($)
Coal	0.000116 - 0.001667	18.28 - 19.98	236 - 3043
Fuel Oil	0.0059 - 0.0342	150 - 233	0 - 10379
Natural Gas	0.002 - 0.008	21.13 - 57.03	0 - 3859
Nuclear	0.00015 - 0.00023	5-11	1000 - 1500

In addition, the 8 zones are connected by a transmission grid of 12 lines. For our testing purpose, instead of matching real-word data, we set all transmission lines' capacity to be $1000MW$. Moreover, at each zone there are hourly fixed base-demand that do not respond to price signals, as shown in Figure 3.2. Also, in each zone there are distributed smart home agents that control their flexible demand to respond to price signals.

Wind Generation.

Besides the thermal generation units provided by ISO-NE Test System, we also consider wind power generation whose marginal generation cost is treated as zero in our simulations. Since ready-made wind power data are not available for the simplified ISO-NE system, we build up a time series model to generate wind speed, and further to generate wind power output. Because of the highly nonlinear mapping of wind speed to wind power generation, we model wind speed instead of wind power at the most beginning [52]. With generated wind speed, we follow the approach in [52] and [87] of using an aggregated power curve to map wind speed to wind power output

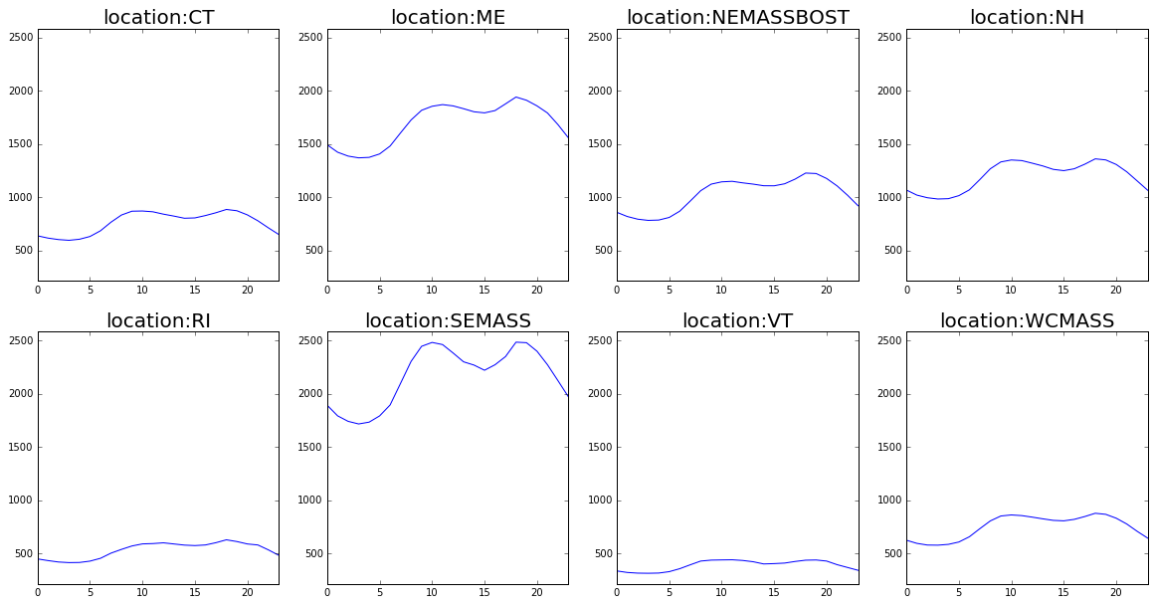

Figure 3.2.: Zonal base demands for 24 hours in 8-Zone ISO-NE Test System.

for an entire zone. The assumption here is that wind turbines have the same power curves and are powered by the same wind speed in each zone, respectively. In the data report [88] by the National Renewable Energy Laboratory (NREL), the states covered by ISO-NE have onshore wind generation capacities as in Table 3.2. Then we scale up a single wind turbine's power curve according to each zone's total capacity to be the aggregate power curve for the entire zone node without changing the turbine's speed specifications, and in our numerical simulations we adopt wind turbine GE 1.5-MW which has cut-in speed $3m/s$, rated speed $12m/s$ and cut-out speed $25m/s$ [89]. Specifically, we let $\{V_{k,h}; h = 1, 2, \dots\}$ denote the kth zonal wind speed series of $\boldsymbol{V_h}$, where h denotes the time dimension. We use vector autoregression (VAR) model of order p to represent the K-zone ($K = 8$ in our case) Gaussian autoregressive base process (denoted as $\text{VAR}_K(p)$) as below [90]:

$$\boldsymbol{V_h} = \sum_{j=1}^{p} \boldsymbol{\alpha_j} \boldsymbol{V_{h-j}} + \boldsymbol{\epsilon_h}, \tag{3.15}$$

where $\{\boldsymbol{\alpha_j}; j = 1, 2, \dots, p\}$ are fixed $K \times K$ autoregressive coefficient matrices and $\boldsymbol{\epsilon_h}$ is the underlying white noise with K-dimension. The noise term $\boldsymbol{\epsilon_h}$ has covari-

Table 3.2.: State totals of onshore wind generations in ISO-NE

	ME	MA	VT	CT	NH	RI	Total
Capacity (MW)	5863	2166	2019	919	2371	1039	14377

ance matrix $\mathbf{\Sigma_\epsilon}$ which should be selected appropriately to ensure that each $V_{k,h}$ is marginally standard normal. Such a model can capture both autocorrelation and spatial correlation of wind speed. Using the time series model in (3.15), we randomly generate the noise $\boldsymbol{\epsilon_h}$ to get wind speed data series. However, the model suffers the problem that negative values can be generated. To overcome the problem, we transform the wind speed generated in (3.15) and assume that the transformed wind speed data $V_{k,h}$ follow the inverse Gaussian distribution $IG_{k,h}$, respectively. That is, wind speed data in each zone have inverse Gaussian distributions with different parameters for different time periods due to diurnal effects. Similarly, we can adjust the parameters for different seasons for seasonal effects. The reason for choosing the inverse Gaussian distribution is that in [52] and [87], the inverse Gaussian distribution is found to provide the best fit to wind speed. Specifically, we transform the wind speed as below:

$$V'_{k,h} = F_{IG}^{-1}(N(V_{k,h})), \tag{3.16}$$

where $V'_{k,h}$ is the transformed wind speed which follow the inverse Gaussian distribution, $F_{IG}^{-1}(\cdot)$ is the inverse of the cumulative distribution function (CDF) of the inverse Gaussian distribution and $N(\cdot)$ is the CDF of the normal distribution. Once we obtain the transformed wind speed data through transformation in (3.16), we feed it into the zonal aggregate power curves to get wind power outputs.

Two-settlement Mechanism.

We consider that ISO-NE conducts economic dispatch optimization in a two-settlement mechanism on an hourly basis. As we focus on the decentralized DSM control strategies by consumers, we do not consider any demand-side active bidding into the wholesale market. In the day-ahead market at each day, the ISO solicits supply bids from power generators to match the demand forecasts of each hour in the next day, and results in day-ahead prices. Since day-ahead forecast is out of the research scope, herein we simply use rolling average of last 10 days real-time net demand as forecast for each hour. At each zone, the net demand is the realized total demand minus the wind generation. In real-time, the ISO matches any supply and demand deviations with additional generation resources, and produces real-time prices. Under RTP, consumers are charged the real-time electricity prices.

Decision Epochs and Temporal Resolution.

In our numerical studies, we conducted 4 simulation epochs and in each simulation epoch there are 100 decision rounds (i.e. 100 days). The number of decision rounds is just an arbitrary number set to be large enough for the repeated games to display convergence. In addition, we consider the 4-hour peak period 17:00-21:00 as our T, and each hour is a time-slot for smart home agents to choose for consuming their flexible demand in the period. Therefore, we have $D = 100$ and $H = 4$. For the simulation purpose, we intentionally drag down the base demand at hour 20:00, as shown in Figure 3.3. Such intention is to see if our algorithms can shift more flexible demand to the hour with lower total base demand, and the system can have flatter total demand with better stability if so.

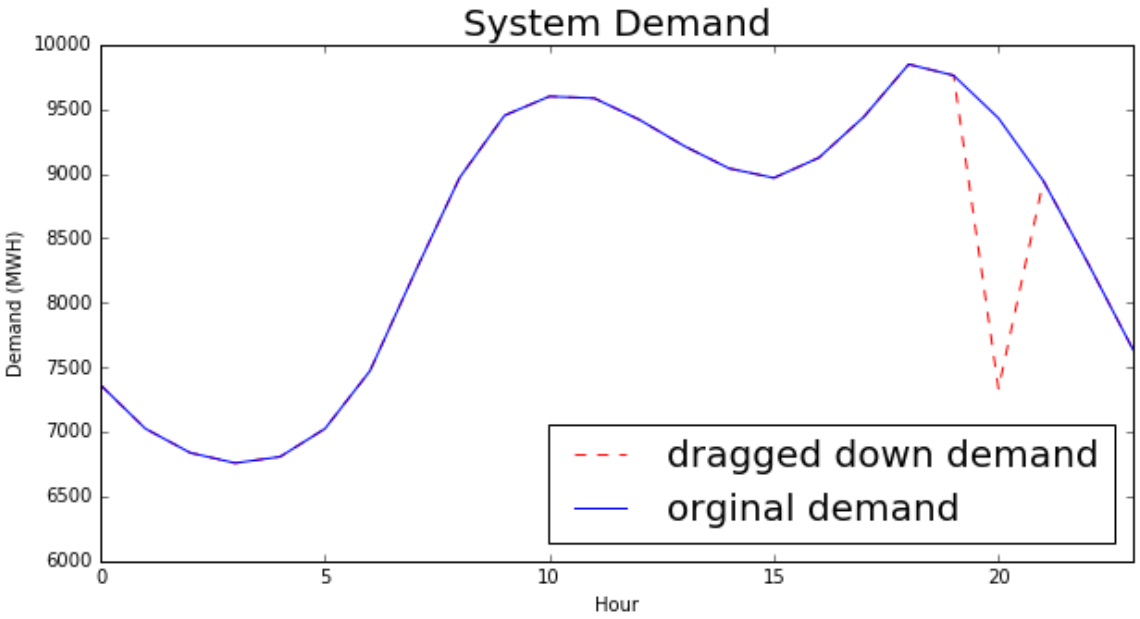

Figure 3.3.: Aggregate system demand: original v.s. dragged-down

Control Strategies by Smart Home Agent.

We consider 5000 heterogeneous smart home agents at each zone, i,e, 40000 agents in total. Each agent controls its flexible demand sampled from *beta*-distribution $Beta(2,2) \times 0.5MWh$ with *regeneration rate* 0.05. In our simulations, we compare four different decentralized DSM control strategies for smart home agents. Besides the LAR algorithm *Noisy Hedge* proposed in this work, Naive-response, Adaptive-response [91], and MAB-game approaches are simulated as follows. In *Naive-response,* agents make decision based on the day-ahead forecast. In *Adaptive-response*, agents sample their *adaptiverate* uniformly in $[0,1]$, respectively. In *MAB-game,* every agent chooses its own bandit learning algorithm from *UCB1, UCB1-Tuned, UCB1-Normal, UCB-2, ϵ-greedy*, and ϵ_n*-greedy* [74] for managing its flexible demand.

3.2.2 Simulation Results

In this section, we compare four decentralized flexible demand management approaches (*LAR* learning by *Noisy Hedge*, MAB-game, Adaptive-response, and Naive-response) under RTP scheme and the simulation framework described in Section 3.2.1. Intuitively, *Noisy Hedge* and MAB-game both conduct online learning, and thus would perform better than heuristic approaches Adaptive-response and Naive-response. Further, *Noisy Hedge* utilize full information feedback while MAB-game only let agents learn from the feedback resulted by the action taken in each round. Therefore, *Noisy Hedge* would be more advanced for using more information to make decisions. Adaptive-response would have better performance than Naive-response because agents have less chance of moving to the same time-slot. Our simulations results in the following validate the intuitions.

System Demand

In Figure 3.4, we present ISO-NE system demands of Hour 17:00-21:00 across 100 days. Four simulation epochs represented by different colors are conducted for each decentralized control approach, respectively. The black horizontal line in each subplot is the system base demand of the hour, and Hour 20 has much lower base demand than the other three hours. From Row 1 to Row 4, the most advanced approach *Noisy Hedge* to the least one Naive-response are presented, respectively. We can easily see that for more advanced control approach, the system has flatter demand curve and thus less demand volatility. *Noisy Hedge* and MAB-game enable the system to have its demand converge very fast while Adaptive-response and Naive-response keep the system demand fluctuating. The high demand volatility with extreme peaks would make dispatch work by ISO impractical and would physically endanger the whole power system.

If we further look at the system net demands (in Figure 3.5 through Figure 3.8, we can see with wind uncertainties, *Noisy Hedge* and MAB-game still generate

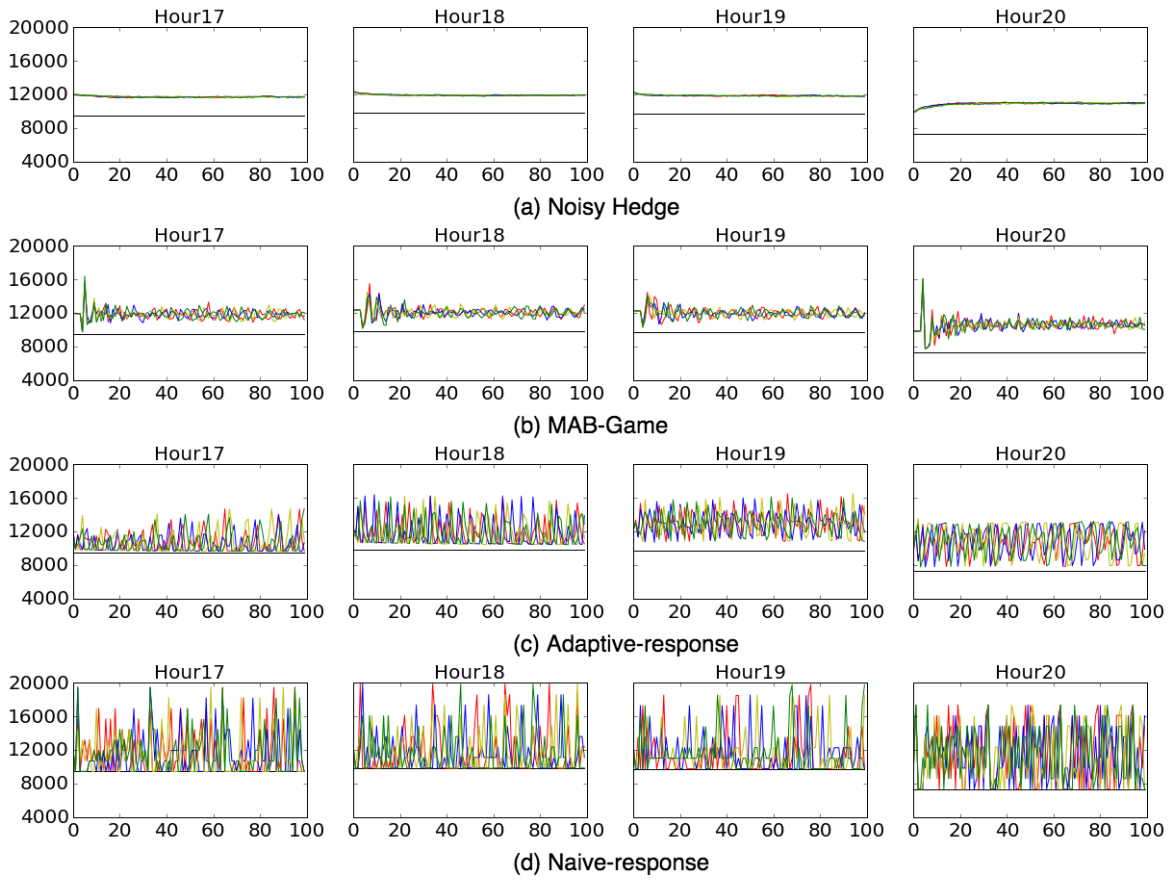

Figure 3.4.: ISO-NE real-time system demand (MWh) of Hour 17:00 - 21:00 across 100 days.

much flatter demand curves. Moreover, since *Noisy Hedge* and MAB-game solve the closed-loop issue using game-theoretic frameworks, the divergence between day-ahead forecast and real-time realized demands are much less, which is further reflected by the day-ahead and real-time prices.

LMP at NEMA & BOST

In Figure 3.9 - 3.12, we present the day-ahead and real-time locational marginal prices (LMP) of Zone NEMA & BOST. The selected zone is a load pocket without any thermal generation around, as shown in Figure 3.1. The other zones have very

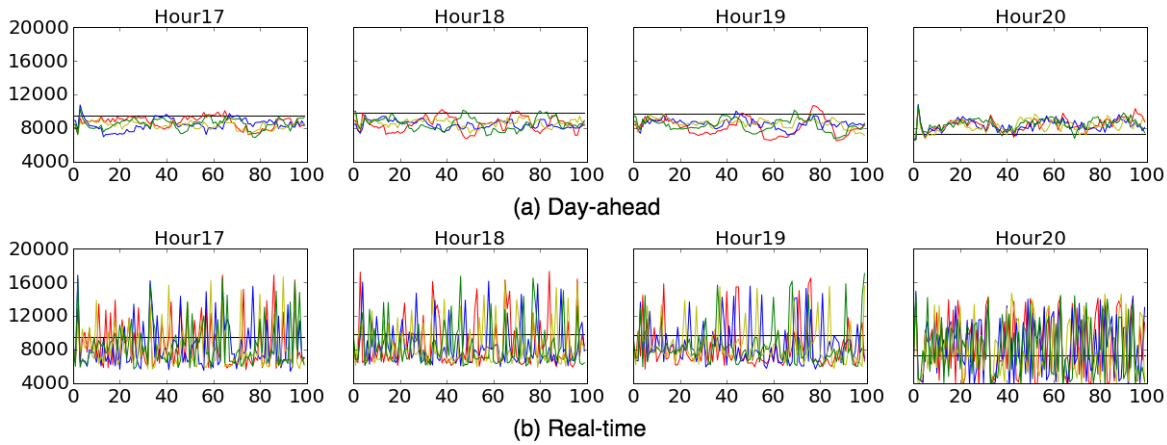

Figure 3.5.: ISO-NE system net demand (MWh) of Hour 17:00 - 21:00 across 100 days under Naive-response: (a) day-ahead; (b) real-time.

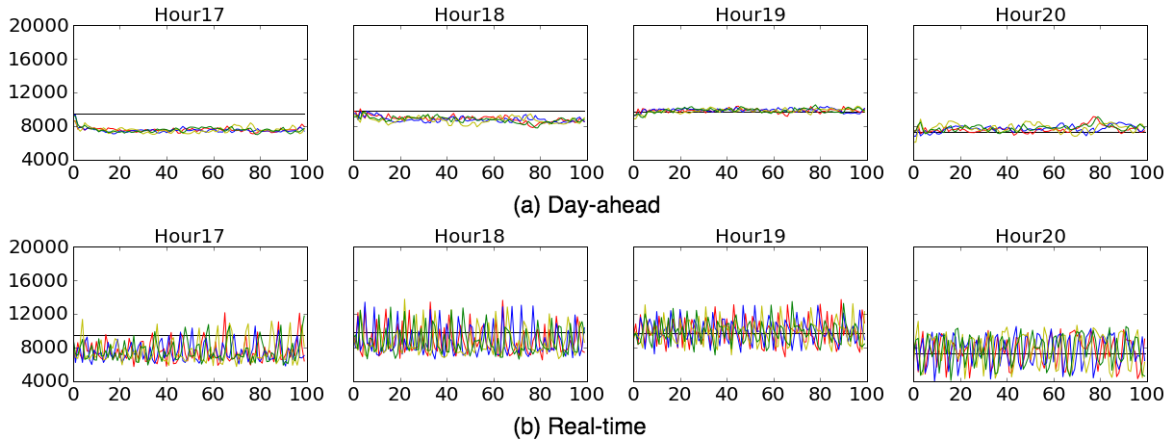

Figure 3.6.: ISO-NE system net demand (MWh) of Hour 17:00 - 21:00 across 100 days under Adaptive-response: (a) day-ahead; (b) real-time.

similar LMP patterns. We can see the LMPs are structured similarly to the net system demands shown above. Specifically, *Noisy Hedge* results in the most flatter LMP curve for each hour while Naive-response generates a number of price spikes. Such price spikes would bring about undesired financial risks to energy customers.

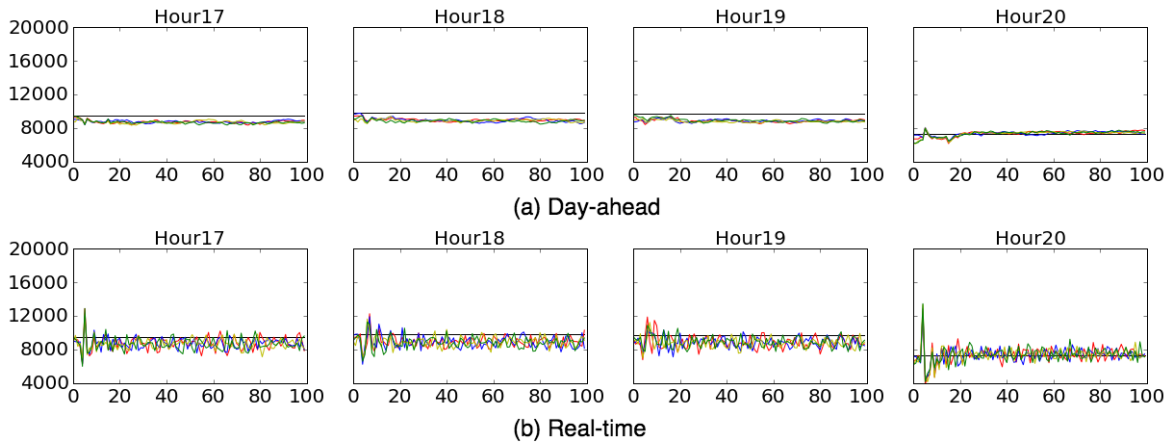

Figure 3.7.: ISO-NE system net demand (MWh) of Hour 17:00 - 21:00 across 100 days under MAB-Game: (a) day-ahead; (b) real-time.

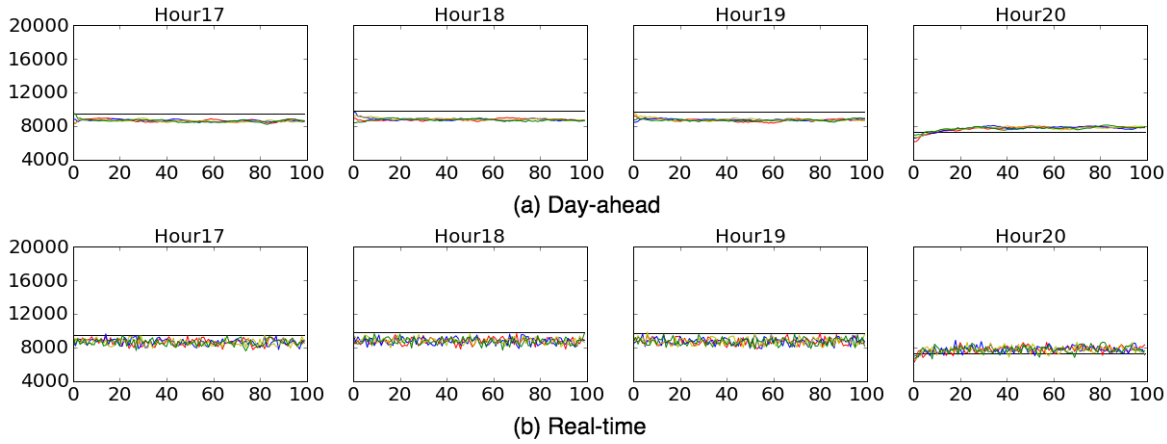

Figure 3.8.: ISO-NE system net demand (MWh) of Hour 17:00 - 21:00 across 100 days under Noisy Hedge: (a) day-ahead; (b) real-time.

In addition to the price volatility for each hour, we are also interested in the price volatility across hours. To quantitatively study the consecutive price volatility, we define the incremental mean volatility (IMV) as below:

$$IMV = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{H-1}\sum_{h=1}^{H}|p_{d,h+1}^{RT} - p_{d,h}^{RT}|, \tag{3.17}$$

and the corresponding log-scaled incremental mean volatility (LIMV) is as below:

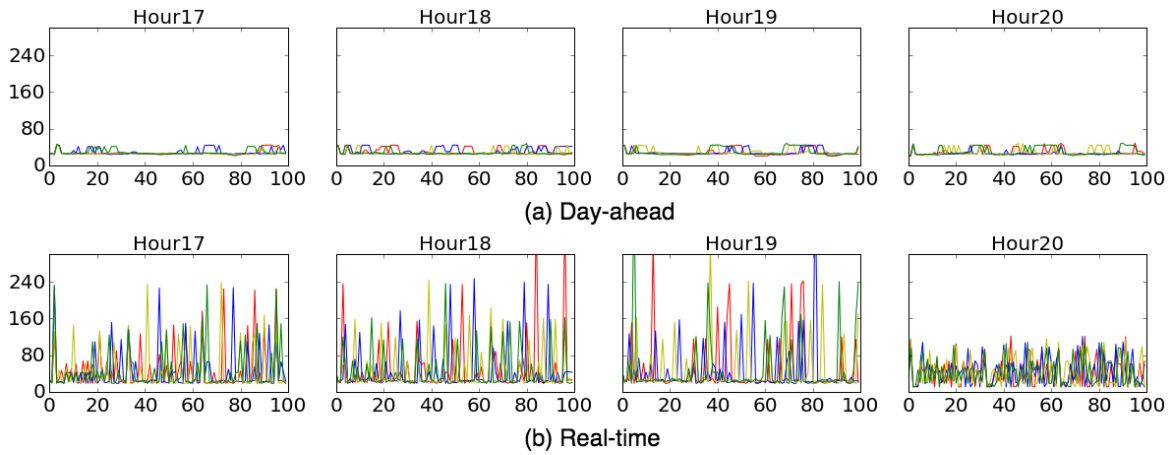

Figure 3.9.: Zone NEMA & BOST LMP ($\$/MWh$) of Hour 17:00 - 21:00 across 100 days under Naive-response: (a) day-ahead; (b) real-time.

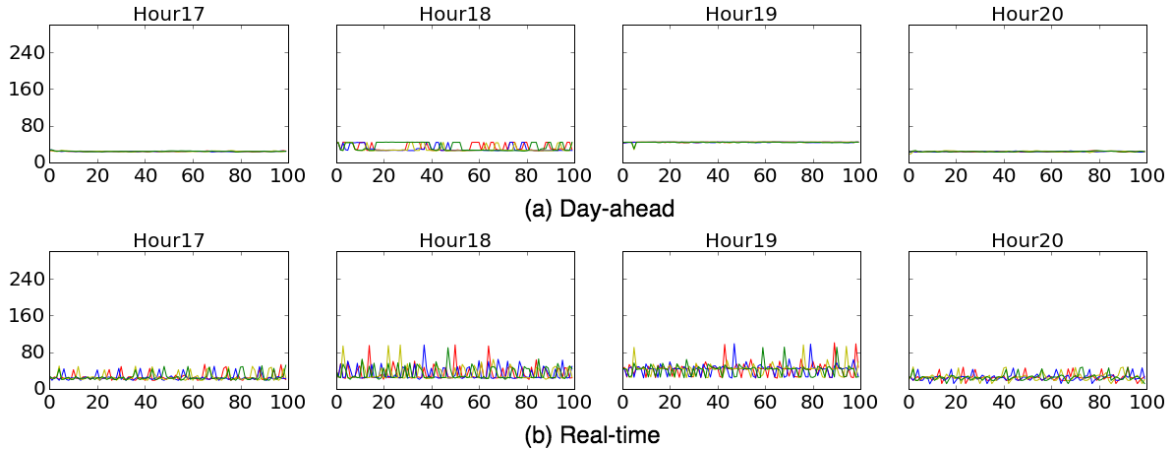

Figure 3.10.: Zone NEMA & BOST LMP ($\$/MWh$) of Hour 17:00 - 21:00 across 100 days under Adaptive-response: (a) day-ahead; (b) real-time.

$$LIMV = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{H-1}\sum_{h=1}^{H}|log(p^{RT}_{d,h+1}) - log(p^{RT}_{d,h})|. \tag{3.18}$$

The value of IMV indicates on average how much price difference there are between two consecutive hours. In Table 3.3, we present the both IMV and LIMV for each method. We can easily find that for consecutive hours, we also have less price volatility under more advanced approach. Therefore, advance game-theoretic approaches can

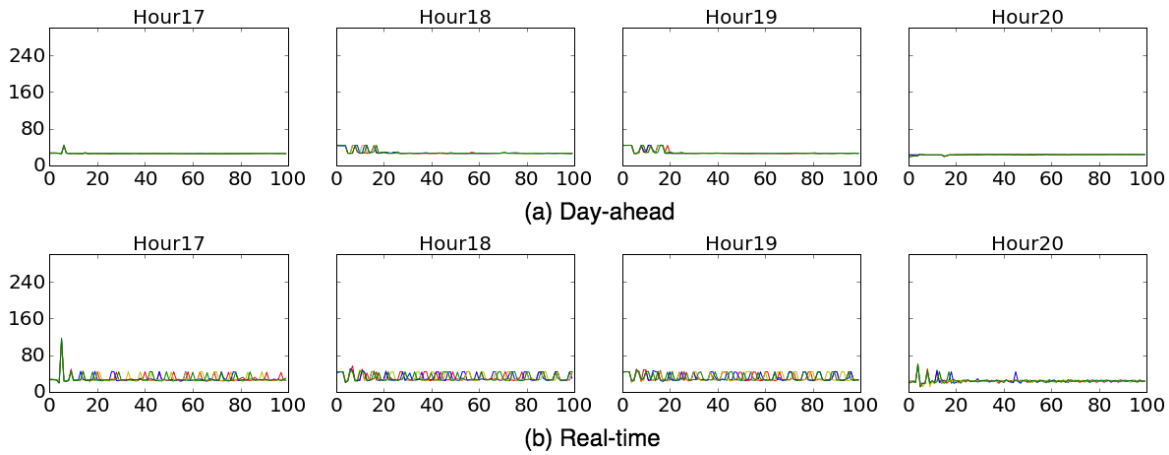

Figure 3.11.: Zone NEMA & BOST LMP ($\$/MWh$) of Hour 17:00 - 21:00 across 100 days under MAB-Game: (a) day-ahead; (b) real-time.

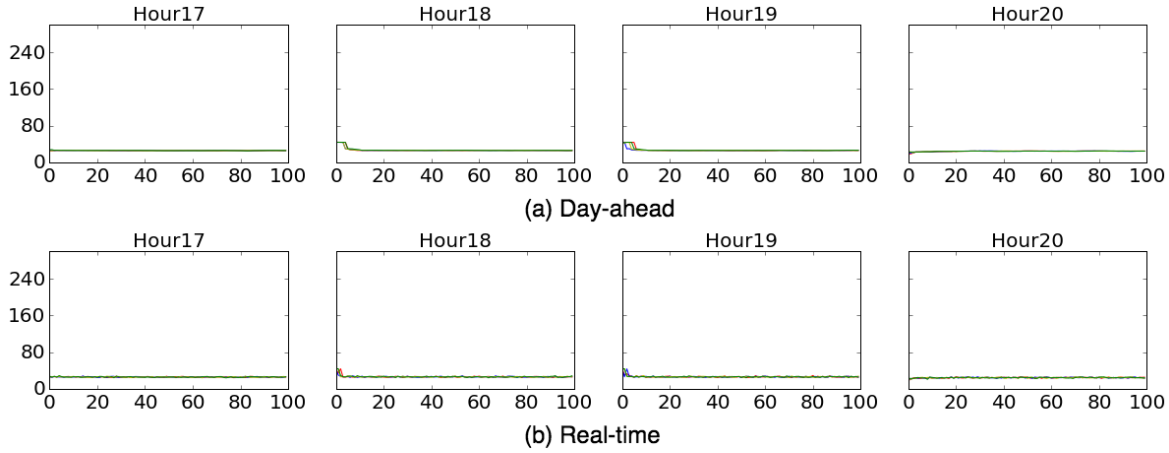

Figure 3.12.: Zone NEMA & BOST LMP ($\$/MWh$) of Hour 17:00 - 21:00 across 100 days under Noisy Hedge: (a) day-ahead; (b) real-time.

generate less volatile prices for both the same hour across days and consecutive hours. To further validate the results in Table 3.3, we run 20 simulation epochs and get 95% Confidence Interval (CI) for IMV of each approach. We have $[1.48, 1.51]$ for *Noisy Hedge*, $[8.55, 8.63]$ for MAB-game, $[16.57, 16.76]$ for Adaptive-response, and $[45.38, 45.64]$ for Naive-response.

Table 3.3.: Zone NEMA & BOST: real-time LMP volatility

	Noisy Hedge	MAB-game	Adaptive-response	Naive-response
IMV	1.49	8.60	16.71	45.50
LIMV	0.06	0.27	0.48	0.82

Besides price volatility, when real-time price deviates from day-time price largely, customers under RTP could have serious financial loss. We can see that the more advanced approach demonstrates less divergence between day-ahead and real-time prices. Numerically, we present average difference in absolute value $\$/MWh$ and percentage % (the percentages are calculated based on day-ahead prices) for each approach in each hour in Table 3.4 and 3.5, respectively.

Table 3.4.: Zone NEMA & BOST: average LMP divergence in absolute difference ($\$/MWh$) between day-ahead and real-time

	Hour 17	Hour 18	Hour 19	Hour 20
Noisy Hedge	\$0.62	\$1.08	\$1.20	\$0.74
MAB-game	\$4.27	\$5.98	\$6.30	\$2.51
Adaptive-response	\$4.62	\$12.05	\$8.36	\$5.06
Naïve-response	\$26.78	\$26.41	\$26.73	\$28.26

System-level Costs

Moreover, in Figure 3.4, we can find that by the online learning approaches, besides the resulting flat demand curve in each hour across days, agents consume more flexible demand in hour with lower system base demand, which could result in demand with lower volatility across hours as well. Thus the system would enjoy lower dispatch costs for meeting less volatile demand due to the convexity of generation costs. In Figure

Table 3.5.: Zone NEMA & BOST: average LMP divergence in percentage (%) between day-ahead and real-time.

	Hour 17	Hour 18	Hour 19	Hour 20
Noisy Hedge	2.42%	3.43%	3.89%	3.10%
MAB-game	16.31%	21.65%	22.24%	11.06%
Adaptive-response	19.38%	39.64%	19.43%	21.53%
Naïve-response	100.96%	98.46%	96.64%	106.56%

3.13 and Table 3.6, we present ISO-NE system economic dispatch costs for meeting demands associated with agents' decentralized demand response under RTP which have been averaged across days and simulation epochs for each hour, respectively. We can find that more advanced control approaches have less total economic dispatch costs, and Naive-response has about 30% more than *Noisy Hedge*.

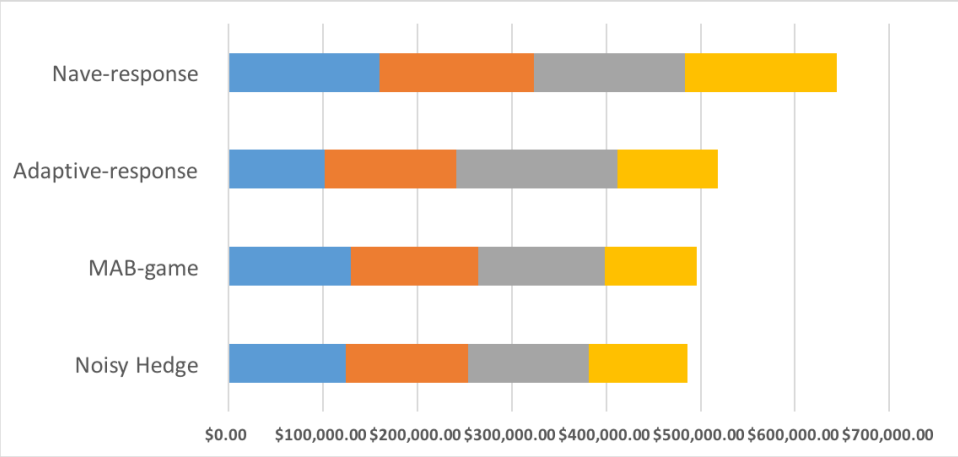

Figure 3.13.: ISO-NE average system economic dispatch costs of Hour 17:00 - 21:00.

Besides, when demand fluctuates with high volatility, the system network would suffer from congestion due to potential critical peaks. In Figure 3.14 and Table 3.6, we present the average congestion costs for each approach in each hour. We can see the similar trend as economic dispatch costs, i.e. more advanced control approaches

Table 3.6.: ISO-NE average system economic dispatch costs of Hour 17:00 - 21:00 in 4 simulation epochs.

	Hour 17	Hour 18	Hour 19	Hour 20	Total
Noisy Hedge	$124,845.07	$129,117.32	$128,118.72	$104,183.77	$486,264.88
MAB-game	$129,427.85	$135,434.51	$133,423.21	$97,809.11	$496,094.69
Adaptive-response	$102,066.08	$139,018.08	$170,608.74	$106,445.55	$518,138.44
Naïve-response	$160,606.45	$162,621.42	$159,921.95	$161,484.65	$644,634.47

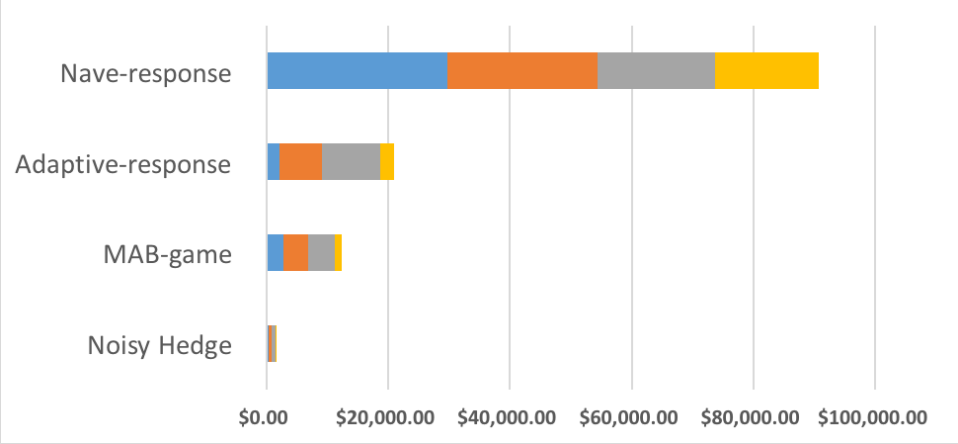

Figure 3.14.: ISO-NE average system congestion costs of Hour 17:00 - 21:00.

Table 3.7.: ISO-NE average system congestion costs of Hour 17:00 - 21:00 in 4 simulation epochs.

	Hour 17	Hour 18	Hour 19	Hour 20	Total
Noisy Hedge	$381.69	$478.71	$602.30	$130.38	$1,593.06
MAB-game	$2,820.75	$4,027.52	$4,362.46	$1,129.72	$12,340.45
Adaptive-response	$2,162.31	$7,026.47	$9,563.96	$2,268.60	$21,021.33
Naïve-response	$29,785.54	$24,580.85	$19,384.64	$16,975.09	$90,726.12

have less total congestion costs. Since *Noisy Hedge* has the flattest demand curve with nearly no demand peak, associated congestion costs are thus very low. In the opposite, Naive-response has extremely high congestion costs which are about 60 times the costs of *Noisy Hedge*.

With involving wind generation uncertainties, the results of the online LAR learning presented in this section still validate the system level convergence shown in Section 3.1, and its advantages against other decentralized approaches further motivate us to design our agent-based demand-side management mechanism more carefully under RTP and large population.

To further validate the results, we run 20 simulation epochs for each approach, respectively. The 95% Confidence Intervals (CI) for the system costs are summarized as below:

Table 3.8.: 95% CI for Economic Dispatch (ED) costs and Congestion (CG) costs.

	ED Cost	CG Cost
Noisy Hedge	[486263.28, 486266.21]	[1592.13, 1595.97]
MAB-game	[496093.26, 496098.61]	[12337.04, 12342.70]
Adaptive-response	[518134.47, 518140.77]	[21017.44, 21022.99]
Nave-response	[644630.81, 644638.54]	[90722.25, 90729.06]

4. MULTI-AGENT LEARNING IN DOUBLE AUCTIONS FOR P2P ENERGY TRADING

Distributed energy resources (DERs) are a vital part of a smart grid, as such resources can improve system reliability and resilience with their proximity to load, and promote sustainability, with the majority of DERs being solar and wind resources [62,63]. While we believe that RTP is crucial to realize the benefits of demand flexibility, including the distributed generations (DG), such pricing mechanisms for end-customers have not been widely adopted yet. Thus, currently to incentivize investments in DERs, there are two general approaches: non-market-based versus market-based. The most common and widely used policy in a non-market-based approach is feed-in-tariff (FIT) [92] (including net-metering). While effective in promoting DERs, it may create equity issues as consumers without DERs would face increased electricity rates to pay for the FIT. In a market-based approach, a marketplace exists for consumers and DER owners, also referred to as prosumers, to buy or sell energy, in a distribution system under the wholesale market. The actual rates that market participants pay/receive would fluctuate over time, reflecting the dynamic supply and demand conditions.

In a bilateral marketplace, a leading mechanism to match supply and demand is through a double-side auction. While auction designs have been well studied in the field of economics and game theory [64, 65, 93, 94], several special features of a peer-to-peer (P2P) energy market require special attention. To name a few, a P2P energy market inherently involves repeated auctions and exogenous uncertainties (e.g., wind/solar availability), making the analysis of market participants' bidding/asking strategies much more difficult. In addition, market participants are likely to have bounded rationality in the sense that they do not know their own valuation of energy production and consumption. Furthermore, their (implicit) valuations are likely

dependent, such as in a hot summer day, all buyers would value high of energy consumption for air conditioning. This feature alone would nullify the assumptions of most of the works in auction theory.

To address the theoretical difficulties, and to provide an algorithmic-framework that can be automated to aid consumers and prosumers to participate in a repeated double-auction, we propose a multi-agent, multi-armed bandit learning approach. We run multiple simulations to study the market outcomes of three different auction designs: a replicate of the wholesale market's uniform-price auction, a variant of Vickrey double-side auction [64], and maximum volume matching auction (which is pay-as-bid/receive-as-ask) [65]. Numerical results indicate the convergence of the market outcomes of a MAB game to a steady-state. Based on the simulations, from market participants' perspective, the uniform-price auction outperforms the other two as it can offer higher clear quantities, total social welfare and total normalized reward.

The rest of the dissertation is structured as follows. In Section 4.1, we will describe in details how market participants bid/ask through bandit learning in MAB-game auctions. Then in Section 4.2, three double-side auction mechanisms are presented for P2P energy trading market. Numerical simulations are presented in Section 4.3 by comparing learning results in three different auction mechanisms.

4.1 Learning under MAB-game Framework

Without a P2P energy market, DER owners can only sell their generated energy to the utility or distribution system operator (DSO) at some pre-defined fixed FIT. Similarly, consumers and prosumers can only buy energy from the utility under some agreed pricing contract. In this work, we consider time-of-use (TOU) pricing, widely applied by utilities in the U.S., for customers buying energy, i.e a fixed rate for each time period (e.g. hourly), respectively. While in a bilateral P2P marketplace, consumers and prosumers can trade with each other at rates accepted by both buy-

side and sell-side. Intuitively, a marketplace is desired by both sides if it can provide agents with some rate higher FIT and lower than TOU rate. Otherwise, agents can simply sign contract with the utility to buy/sell at TOU/FIT.

To incentivize the growth of DERs, a double-side auction can be organized for clearing bids/asks from market participants in each time period. Agents need to decide their unit price and quantities of energy for submitting bids/asks to the auctions. In this work we assume that with some smart devices using historical and weather data agents can accurately forecast how much energy themselves will consume or generate in very near future (e.g. in one hour), and thus quantities can be easily decided for the auctions. While agents' implicit valuations of energy are much more difficult to decide. To address the difficulties, we propose a MAB-game learning approach for a multi-agent system in which bidding/asking prices of agents are automatically chosen by bandit learning algorithms, respectively.

More specifically, consider a double-auction where supply and demand bids are submitted in each time period h (e.g., hourly) in each day d. Within each h, we assume that market participants only choose a price to ask/bid, not quantities of energy. We further discretize per-unit price bids (i.e., ¢/KWh) into K possible choices. When each agent decides which price to bid/ask, it is similar to choosing one slot machine, out of K such machines, to pull the arm. In this case, the agents are uncertain if they will win (bids cleared) or lose (bids not cleared), and in the case of winning, how much the payoff would be. This is similar to the classic multi-armed bandit (MAB) learning problem which has been well studied in a wide of litarture, such as [74, 77, 95, 96]. A key difference here, however, is that one agent's probability of winning and the payoff distributions (of each arm) depend on how other market participants bid/ask. A MAB-game [59] is formed when all agents apply bandit learning for deciding their bid/ask with incomplete information feedback. The central idea in a multi-agent MAB game is that each agent assumes that the winning probability, and the payoff distribution, though both unknown to the agent, are stationary, and hence can define cumulative regrets for the auction in period h up to day D.

For illustration purposes, the herein presented formulas concern a single trading-period h (e.g. 1 hour) across days. We consider a set of agents $\mathcal{A} = \mathcal{A}_b \cup \mathcal{A}_s$, where $\mathcal{A}_b$ and $\mathcal{A}_s$ are the sets of buyers and sellers, respectively. Further, we let P_{FIT} and P_{TOU} denote the FIT and TOU rate in ¢/KWh, respectively, and we only consider the situation where the FIT is lower than the TOU rate, i.e. $P_{FIT} < P_{TOU}$.

4.1.1 Discrete Price Arms

The majority of DERs are solar and wind resources, and thus we consider their generation marginal costs as zero despite of fixed installment and maintenance fees. Therefore, any rate higher than FIT would be attractive to DER owners. Similarly, energy buyers desire for any rate lower than TOU rate. Therefore, any rate (in ¢/KWh) in the range $[P_{FIT}, P_{TOU}]$ would profit both energy buyers and sellers, and any reasonable agent $i \in \mathcal{A}$ has a bidding/asking price space $\mathcal{P}_i \in \mathcal{Z}^{\geq 0}$ which contains both P_{FIT} and P_{TOU}.

Herein, each discrete unit price in space $\mathcal{P}_i$ is a price arm that can be picked up for the agent's bid/ask. How to choose a price arm is complicated due to the dynamics of auctions. For each individual agent, it prefers a lower/higher auction clear price if it is a buyer/seller. However, it is not necessary that an agent's bidding/asking price is the auction clear price which depends on the collection of bids and asks. Since agents are not bidding/asking based on their implicit valuations (which are not known by agents), under some auction designs, like uniform price double auction, some agents may take chance by bidding/asking some extreme high/low unit price to make their bids/asks more likely to be accepted by the auction while enjoy the more profitable clear price. In Section 4.1.3, we will discuss the performance bound (i.e regret bound) of picking up price arms in the auction games by bandit learning for agents.

4.1.2 Rewards

The reward each agent receives in the auction represents the normalized level of the actual sent/received payment, Λ_i, between the lower and upper benchmarks by P_{TOU} and P_{FIT} which are denoted by $\underline{\Lambda_i}$ and $\overline{\Lambda_i}$, respectively. Herein, we let q_i denote the demand/supply of agent i, and q_i is negative for a buyer and positive for a seller, i.e. $q_i < 0|_{i\in\mathcal{A}_b}$ and $q_i > 0|_{i\in\mathcal{A}_s}$. For a buyer agent, the lower and upper benchmarks refer to buying all of q_i at P_{TOU} and P_{FIT}, respectively. In the opposite, a seller agent has its lower and upper benchmarks with selling q_i at P_{FIT} and P_{TOU}. Therefore, we have

$$\underline{\Lambda_i} = q_i \cdot [P_{TOU} \cdot \mathbb{1}_{\{i\in\mathcal{A}_b\}} + P_{FIT} \cdot \mathbb{1}_{\{i\in\mathcal{A}_s\}}], \tag{4.1}$$

$$\overline{\Lambda_i} = q_i \cdot [P_{FIT} \cdot \mathbb{1}_{\{i\in\mathcal{A}_b\}} + P_{TOU} \cdot \mathbb{1}_{\{i\in\mathcal{A}_s\}}]. \tag{4.2}$$

The actual sent/received payment of each agent $\forall i \in \mathcal{A}$ consists of two parts for trading in the auction and with the utility, which are denoted by Λ_i^{au} and Λ_i^{ut}, respectively. Thus, we have

$$\Lambda_i = \Lambda_i^{au} + \Lambda_i^{ut}. \tag{4.3}$$

With attending the auction, market participants send/receive payments based on the clear result. Specifically, each agent's sent/received payment in the auction is calculated according to its clear price, p_i^{au}, and clear quantity, q_i^{au}, as below

$$\Lambda_i^{au} = p_i^{au} \cdot q_i^{au}. \tag{4.4}$$

In auctions like uniform price double auction, all agents have the same clear price. While in the maximum volume matching auction [65], the agents may have different clear price since they pay/receive at their bid/ask price.

However, it is not necessary that all agents are buying/selling in the P2P market since some agents' bids/asks may not be (fully) cleared by the market. In this case, for not wasting the (renewable) energy from DERs, prosumers are allowed to sell the unclear energy to the utility at P_{FIT}. Also, consumers always can buy their demand

not satisfied by the P2P market from the utility at P_{TOU}. Therefore, the sent/received payment to/from the utility for agent i is as below

$$\Lambda_i^{ut} = p_i^{ut} \cdot q_i^{ut}, \tag{4.5}$$

where $p_i^{ut} = P_{FIT}$ if $i \in \mathcal{A}_s$ and $p_i^{ut} = P_{TOU}$ if $i \in \mathcal{A}_b$, and q_i^{ut} denotes the unclear energy quantity.

Then we have $q_i = q_i^{au} + q_i^{ut}$. When the agent's auction clear price $p_i^{au} \in [P_{FIT}, P_{TOU}]$, we have $\Lambda_i \in [\underline{\Lambda_i}, \overline{\Lambda_i}]$ and thus we have the normalized reward $\pi_i \in [0, 1]$ calculated as below

$$\pi_i = (\Lambda_i - \underline{\Lambda_i})/(\overline{\Lambda_i} - \underline{\Lambda_i}). \tag{4.6}$$

In Eq. (4.6), we can see for $p_i^{au} = P_{FIT}$, a buyer agent has $\pi_i = 1$ while a seller agent has $\pi_i = 0$, and for $p_i^{au} = P_{TOU}$ we have the opposite values. However, in Section 4.1.1 we mentioned that the agent's bidding/asking price space $\mathcal{P}_i$ contains P_{FIT} and P_{TOU}, and thus the agent may bid/ask some price outside the range $[P_{FIT}, P_{TOU}]$. Though it is counter-intuitive, the auction clear price p_i^{au} could be outside $[P_{FIT}, P_{TOU}]$, even in the uniform-price double auction if a significant population are doing so. In the case $p_i^{au} < P_{FIT}$, we consider $\pi_i = 1|_{i \in \mathcal{A}_b}$ and $\pi_i = 0|_{i \in \mathcal{A}_s}$; for $p_i^{au} > P_{TOU}$, $\pi_i = 0|_{i \in \mathcal{A}_b}$ and $\pi_i = 1|_{i \in \mathcal{A}_s}$. Combined with Eq. (4.6), we have

$$\pi_i = \begin{cases} 1 \cdot \mathbb{1}_{\{i \in \mathcal{A}_b\}} + 0 \cdot \mathbb{1}_{\{i \in \mathcal{A}_s\}}, & \text{for } p_i^{au} < P_{FIT} \\ (\Lambda_i - \underline{\Lambda_i})/(\overline{\Lambda_i} - \underline{\Lambda_i}), & \text{for } P_{FIT} \le p_i^{au} \le P_{TOU} \\ 0 \cdot \mathbb{1}_{\{i \in \mathcal{A}_b\}} + 1 \cdot \mathbb{1}_{\{i \in \mathcal{A}_s\}}, & \text{for } p_i^{au} > P_{TOU} \end{cases}, \tag{4.7}$$

where $\underline{\Lambda_i}$, $\overline{\Lambda_i}$, and Λ_i can be achieved by Eq. (4.1), (4.2), and (4.3), respectively.

4.1.3 Pricing by Bandit Learning

As in Eq. (4.7), we can see the reward π_i of each agent highly depends on its clear price in auction which further depends on its bid/ask and the collection of other agents' bids/asks. The dynamic auction games result in nonstationary clear prices,

which makes the bidding/asking decision-making difficult for agents. In regular game theory literature, the standard equilibrium concept for dynamic games of incomplete information is Perfect Bayesian Nash equilibrium (PBNE) [73, 97]. In a PBNE, the collection of each agent's action profile maps the entire history of the games to each agent's feasible set of actions, under the assumption that each agent maintains their beliefs of other competitors' distribution of action space based on the Bayes' updating rule. For a large population, the assumption requirement is impractical and implausible for small-scale (in terms of computation power) agents in P2P energy trading auctions. This is where MAB-game comes in. Instead of tracking their competitors' tremendous states, agents only need to look at their own history in repeated games. A recent breakthrough on MAB-game in [59] has provided us with the theoretical foundations in studying the auction games with a large population in this work. A key point in MAB-game with many agents is that as every agent conducts its own stochastic *no-regret* bandit learning independently in repeated games, the finite system will approximately converge to the unique *mean field steady state* (MFSS) of the infinite population system. The population profile (i.e. the proportion of population on each arm) is stationary in the MFSS, and the approximation gets better as the finite population increase. Under the stationary population profile, efficient outcomes will be achieved since each individual agent can solve its MAB problem with stationary reward distributions as in classic MAB problem settings.

We let $\boldsymbol{f}$ denote the energy quantities' stationary population profile of the agent set $\mathcal{A}$, where $f(k)$ represent the distribution of buying and selling energy quantities on price arm k. With stationary population profile $\boldsymbol{f}$, each agent has its underlying optimal bid/ask price arms whose associated clear price results in the optimal reward as below

$$\pi_i^*(\boldsymbol{f}) = \max_{k \in \mathcal{P}_i} \mathbb{E}[\pi_i(\boldsymbol{f}, k)], \tag{4.8}$$

where $\pi_i(\boldsymbol{f}, k)$ denotes the reward of agent i for picking up price arm k under population profile $\boldsymbol{f}$.

Suppose that for the trading-period h across D days (i.e. D rounds in our context), agent i uses a policy σ which is an algorithm picking up the next price arm based on its learning history. The history is only about the agent's own sequence of played price arms and corresponding observed rewards, which largely reduces the knowledge dimension that the agent has to maintain. Though the underling optimal reward $\pi_i^*(\boldsymbol{f})$ is unknown to the agent, the policy σ enables the agent to learn about the distributions of rewards for each price arm. Let $N_\sigma(D,k)$ be the number of times price arm k has been picked up by the policy σ during all the D rounds. Then for agent i, we define its cumulative regret under the policy σ for every D rounds as below

$$\Delta_\sigma = \pi_i^*(\boldsymbol{f}) \cdot D - \sum_{k \in \mathcal{P}_i} \mathbb{E}[\pi_i(\boldsymbol{f}, k) \cdot N_\sigma(D,k)]. \tag{4.9}$$

The regret Δ_σ in Eq. (4.9) is the expected loss due to the fact that the policy does not necessarily always pick up the optimal price arm under the stationary population profile which is unknown to the agent. The policy σ is a *no-regret* bandit learning policy if the regret in Eq. (4.9) satisfies:

$$\frac{1}{D}\Delta_\sigma < R(D,K), \tag{4.10}$$

for some $o(1)$ function R in terms of D; where K is the cardinality of $\mathcal{P}_i$, i.e. $|\mathcal{P}_i| = K$. Then $R(D,K)$ gives an upper bound to the average regret under the policy σ. For the bandit learning algorithms based on UCB [74], such as UCB1, UCB-tuned and UCB2, we have logarithmic regret bounds that are $o(1)$ in terms of total rounds D: $R(D,K) = \alpha(K) \cdot \frac{1}{D} ln(D)$. Therefore, as the auction games go on, the agent's average regret goes to 0.

4.2 Double Auction Designs

In this section, we first define the individual monetary utility, corresponding total social welfare, and auctioneer's profit with a P2P energy market auction. Then we discuss about three different double-side auction designs that can be applied for the

market clear: the uniform-price auction, a variant of Vickrey double-side auction [64], and the maximum volume matching auction [65].

4.2.1 Social Welfare and Auctioneer's Profit

As mentioned above, agents are rarely aware of their private valuation of energy production and consumption. To define agents' individual monetary utility, we consider it as profit for energy sellers and costs reduction for buyers with participating the P2P market. Since for renewable DER owners, the marginal cost is almost zero, the total profit of energy seller $i \in \mathcal{A}_s$ is as below

$$u_i|_{i\in\mathcal{A}_s} = p_i^{au} \cdot q_i^{au} + P_{FIT} \cdot q_i^{ut}, \tag{4.11}$$

which has the same value as Λ_i in Eq. (4.3). For consumers, they have to pay at P_{TOU} without the P2P market, thus we have the cost reduction as

$$u_i|_{i\in\mathcal{A}_b} = (P_{TOU} - p_i^{au}) \cdot |q_i^{au}|. \tag{4.12}$$

In spite of the auctioneer's profit, the total social welfare of all agents, denoted by $U_{\mathcal{A}}$, is simply the aggregation of all agents' utility, i.e. $U_{\mathcal{A}} = \sum_{i\in\mathcal{A}} u_i$. For the auctioneer (which can be played by the utility or DSO), the total auction trading surplus it earns is the sum of bid-ask price difference for each energy unit traded in the auction, which is calculated as below

$$U_{\mathcal{M}} = \sum_{i\in\mathcal{A}_b} (p_i^{au} \cdot |q_i^{au}|) - \sum_{i\in\mathcal{A}_s} (p_i^{au} \cdot q_i^{au}), \tag{4.13}$$

where $U_{\mathcal{M}}$ denotes the auctioneer's profit.

4.2.2 Uniform-Price Double Auction

If price is plotted as a function of aggregate energy quantity following the convention in economics, then the energy demand and supply curves slope downward and upward, respectively, as shown in Figure 4.1. Graphically, the intersection (P^*, Q^*)

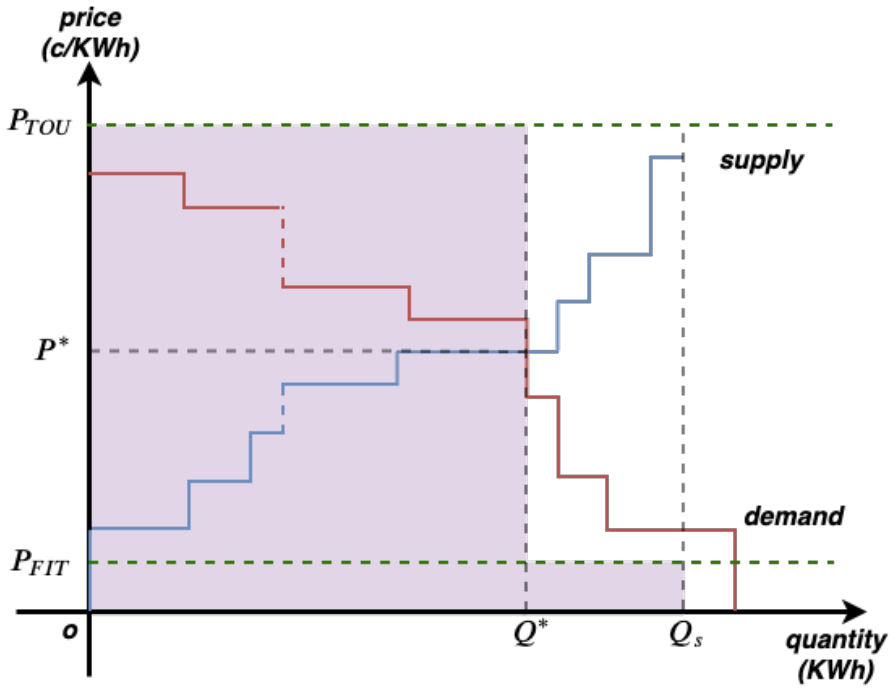

Figure 4.1.: A uniform-price double auction market.

of the supply and demand curves clears the market at which the quantity demanded is equal to the quantity supplied. The price P^* is the *equilibrium price*, and the corresponding energy quantity is the *equilibrium quantity*. As such, all agents pay/receive at the uniform price P^*, and the quantity Q^* in total is traded in the auction. Then the rest supply $Q_s - Q^*$ is sold to the utility at P_{FIT}, in which Q_s denotes the total energy supplied by DERs, i.e. $Q_s = \sum_{i \in \mathcal{A}_s} q_i$. Also, the unsatisfied demand is purchased from the utility at P_{TOU}. Therefore, in Figure 4.1, the shadow area in light purple represents the total social welfare $U_{\mathcal{A}}$, i.e.

$$U_{\mathcal{A}} = P_{TOU} \cdot Q^* + P_{FIT} \cdot (Q_s - Q^*). \tag{4.14}$$

Since $p_i^{au} = P^*$ for all agents $i \in \mathcal{A}$, and both $\sum_{i \in \mathcal{A}_b} |q_i^{au}|$ and $\sum_{i \in \mathcal{A}_s} q_i^{au}$ are equal to Q^*, by Eq. (4.13) the auctioneer earns zero profit in the auction, i.e. $U_{\mathcal{M}} = 0$.

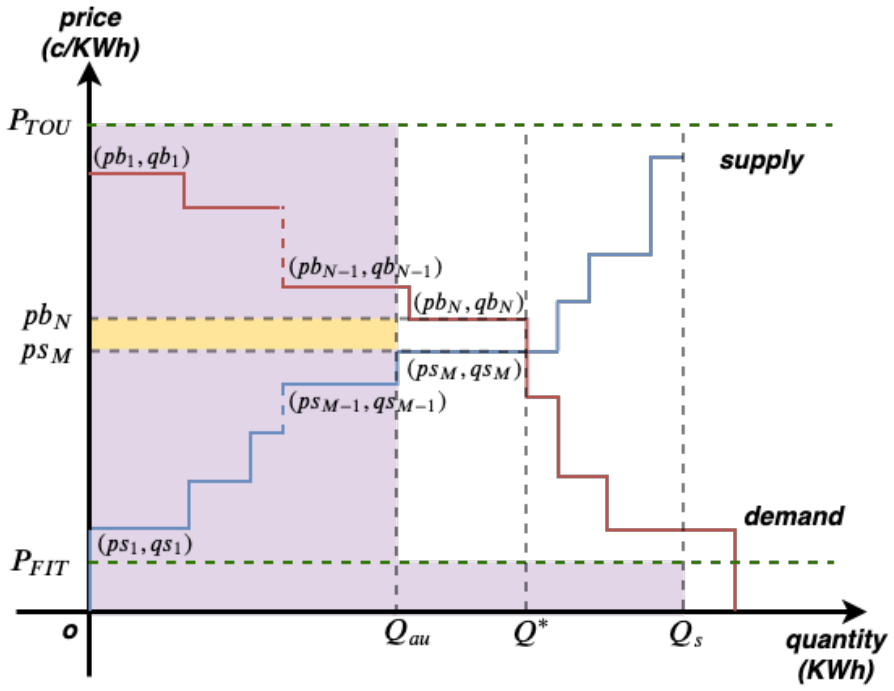

Figure 4.2.: A Vickrey-like double auction market (Case I).

4.2.3 Vickrey Variant Double Auction

Instead of paying/receiving at the uniform *equilibrium price*, we consider the Vickrey-like auction clear on both sides of the market., i.e. paying/receiving the price at Q^* on the demand/supply curve, respectively. The mechanism works as follows. Similar to the uniform-price auction, all bids/asks are sorted down/up by price, and we can have stair-wise demand/supply curves as shown in Figure 4.2, in which each stair represents the collective bids/asks at price arm pb_n/ps_m. At the critical intersection point (P^*, Q^*) where the aggregate demand and supply meet, there are collective bid (pb_N, qb_N) and ask (ps_M, qs_M). Then we consider two cases. Case I (as shown in Figure 4.2):

$$pb_N \geq ps_M \geq pb_{N+1}, \tag{4.15}$$

$$\sum_{m=1}^{M-1} qs_m \leq \sum_{n=1}^{N} qb_n \leq \sum_{m=1}^{M} qs_m, \tag{4.16}$$

and Case II:

$$ps_{M+1} \geq pb_N \geq ps_M, \tag{4.17}$$

$$\sum_{n=1}^{N-1} qb_n \leq \sum_{m=1}^{M} qs_n \leq \sum_{n=1}^{N} qb_m. \tag{4.18}$$

Herein, we describe the clear mechanism for Case I; Case II is similar.

Rule 1

If $\sum_{n=1}^{N-1} qb_n \geq \sum_{m=1}^{M-1} qs_m$, there is overdemand. All the asks with $m < M$ sell all their supply qs_m at price ps_M; all the asks with $m \geq M$ sell their supply at P_{FIT} to the utility. All the bids with $n < N$ buy at pb_N and each of them buys a volume equal to $qb_n - (\sum_{n=1}^{N-1} qb_n - \sum_{m=1}^{M-1} qs_m)/(N-1)$; all the unsuccessful bids buy at P_{TOU} from the utility.

Rule 2

If $\sum_{n=1}^{N-1} qb_n \leq \sum_{m=1}^{M-1} qs_m$, there is oversupply. All the bids with $n < N$ buy all their demand qb_n at price pb_N; all the bids with $n \geq N$ buy their demand at P_{TOU} from the utility. All the asks with $m < M$ sell at ps_M and each of them sells a volume equal to $qs_m - (\sum_{m=1}^{M-1} qs_m - \sum_{n=1}^{N-1} qb_n)/(M-1)$; all the unsuccessful asks sell at P_{FIT} to the utility.

According to the clear rules, the total trade volume in the auction is

$$Q_{au} = min(\sum_{n=1}^{N-1} qb_n, \sum_{m=1}^{M-1} qs_m). \tag{4.19}$$

Then the total social welfare for all agents can be calculated as below (which is represented by the light purple area in Figure 4.2)

$$U_{\mathcal{A}} = [(P_{TOU} - pb_N) + ps_M] \cdot Q_{au} + P_{FIT} \cdot (Q_s - Q_{au}). \tag{4.20}$$

The auctioneer's profit represented by the yellow shadow area in Figure 4.2 is as below

$$U_{\mathcal{M}} = (pb_N - ps_M) \cdot Q_{au}. \tag{4.21}$$

4.2.4 Maximum Volume Matching Double Auction

Other than chasing socail welfare for agents or profit for acutionner, the auction design proposed in [65] is for maximizing the traded volume given a set of bids and asks. The idea of market clear can be intuitively and graphically illustrated in Figure 4.3. Suppose the demand/supply curves are based on the bids/asks shown in Fig 4.1. The supply curve is flipped horizontally and then shifted right towards the demand curve until the two curves touch. The distance, denoted by Q_{au}, that it can move is the minimal horizontal distance between the flipped supply curve and the demand curve which is exactly the maximum trading volume of the auction can be achieved. Then for the energy quantity 0 through Q_{au}, the corresponding bids (pb_n, qb_n) on the demand curve and asks (ps_m, qs_m) on the shifted supply curve are matched, and then successfully matched buyers/sellers pay/receive at their bid/ask price, respectively. Let $\mathcal{S}_b$ and $\mathcal{S}_a$ denote the set of successful bids and asks, respectively. The supply amount $Q_s - Q_{au}$ of the unsuccessful asks is sold to the utility at P_{FIT}, and also the unsatisfied demand is bought at P_{TOU}.

According to the clear mechanism, the total social welfare of all agents is as below (represented by the light purple shadow area in Figure 4.3)

$$U_{\mathcal{A}} = \sum_{n \in \mathcal{S}_b} (P_{TOU} - pb_n) qb_n + \sum_{m \in \mathcal{S}_a} ps_m qs_m + P_{FIT}(Q_s - Q_{au}). \tag{4.22}$$

The auctioneer's profit is still the auction trading surplus (represented by the yellow shadow area in Figure 4.3) as below:

$$U_{\mathcal{M}} = \sum_{n \in \mathcal{S}_b} (pb_n \cdot qb_n) - \sum_{m \in \mathcal{S}_a} (ps_m \cdot qs_m). \tag{4.23}$$

4.3 Numerical Simulations

In this section, we present the simulation results with distributed bandit learning corresponding to the three double-side auction designs for P2P energy trading as described in the previous section.

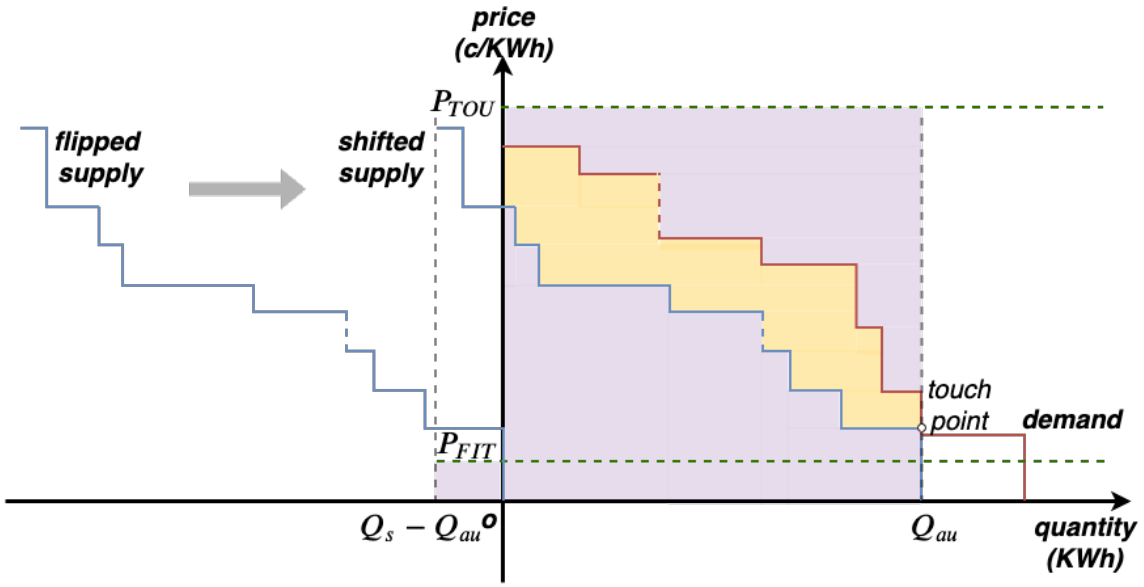

Figure 4.3.: A maximum volume matching auction market.

4.3.1 Input Data

Decision epochs and temporal resolution

As a starting point, we do not consider time-linking constraints in our models, and each trading window is independent of others in a day. The simulations presented herein concern a single one-hour trading period for the peak hour 17:00 - 18:00 across 300 days, i.e. $D = 300$.

TOU/FIT and decision space

We consider fixed TOU/FIT across days, and we let $P_{TOU} = 11$ ¢/KWh and $P_{FIT} = 5$ ¢/KWh. All agents has the same decision space $\mathcal{P}$ that contains all the discretized price arms through 0 ¢/KWh to 14 ¢/KWh, and thus P_{TOU}/P_{FIT} are included in $\mathcal{P}$.

Bandit learning algorithms for pricing

For picking up price arms to bid/ask in the auctions, each agent $i \in \mathcal{A}$ uniformaly chooses its bandit learning algorithms among UCB1, UCB-tuned, UCB2, and $\epsilon - greedy$. Interested readers can refer to [74] for the details of the algorithms.

Consumers and energy demand to buy

In the numerical test cases, we simulate 2000 distributed residential household consumers that participate in the auctions, i.e. $|\mathcal{A}_b| = 2000$. According to the Residential Residential Energy Consumption Survey (RECS) by U.S. Energy Information Administration (EIA) [98], a residential customer consumes about 30 KWh per day on average. Consider it is a peak hour, we naively let consumers repeatedly sample their energy demand quantities from a *Uniform* distribution $U(1.5, 2)$ in KWh, independently, for the hour across days, which is slightly higher than the average consumption level.

Prosumers and energy supply to sell

On the sell-side, we also consider 2000 prosumers with DERs, i.e. $|\mathcal{A}_s| = 2000$. For the DERs, we only consider two renewable resources, solar and wind, for small-scale distributed agents in this work. Due to the popularity of distributed residential solar panels (especially in western), we assume 4/5 of the prosumers have solar-based distrbuted generation, and the other 1/5 have wind-based. In the simulations, we use System Advisor Model (SAM) [99] developed by National Renewable Energy Laboratory (NREL) to model residential generation by solar and wind. The weather resource data for Arizona State by NREL is used for the simulations in SAM.

For the solar generation, we consider all panels have nameplate capacity as 2 KWdc with DC to AC ratio of 1.2 and inverter efficiency of 96%. For each distributed solar resource owner, the module type and array type have equal chance to be one of

Table 4.1.: Wind turbine models

Model	KW Rating
Energy Ball HEA V100 1.1m 0.6KW	0.5
Bergey BWC XL.1	1
True North Power Arrow 2m 1KW	1.23
Future Energy FE1048U 1.8m 1KW	1.5
Hummer 3.1m 1KW	2
Energy Ball HEA V200 1.98m 2.5KW	2.23
Southwest Windpower Skystream 3.7m 1.9KW	2.63
Westwind 3.7m 3KW	3.1

{*Standard, Premium, Thin Film*} and {*Fixed Open Rack, Fixed Root Mount, 1 Axis Tracking, 1 Axis Backtracking, 2 Axis Tracking*}, respectively. All other inputs are set as default in the *Photovoltaic PVWatts* simulations for distributed residential in SAM. More details about photovoltaic simulations can be found in [99–101].

For the simulations of distributed residential wind generation, each wind-based prosumer samples its turbine model uniformly from the 8 wind turbine models listed in Table 4.1, and the number of turbines owned by the prosumer is uniformly sampled among 1 through 4. All other inputs are set as default in the *Wind Residential* simulations in SAM. The turbines' specifications, such as wind power curves and turbine layout, can be found in [99, 101].

4.3.2 Numerical Results

The three different auction designs are simulated with the input data. We use UP, VV, and MV to denote uniform price auction, Vickrey variant auction, and maximum volume matching auction, respectively.

In Figure 4.4, the clear quantity results of the auctions are presented, and we can see the results all have a trend of convergence. The counter-intuitive phenomenon is that in the later phase, UP is more likely to have a higher level of traded volume than MV which is designed to maximize traded volume. The reason is that with bandit learning, agents are updating their bids/asks dynamically, and thus the collective bids/asks schedules are not necessarily the same for different auctions. Besides the volume, we can see after a while of learning, UP's total clear quantity has smaller volatility than the other two auction designs. Therefore, in terms of auction clear quantity, UP outperforms VV and MV, and thus the auction design can let more renewable DERs be utilized.

Similar to the clear quantity, agents' total social welfare also display the convergence trend in the auctions, as shown in Figure 4.5. Associated with more clear quantity, buyers and sellers in UP have higher social welfare (in $) than in the other two auctions in the later auctions. The performance of VV and MV are close to each other. Accordingly, for the total normalized reward, the results display very similar patterns as shown in Figure 4.6. Though UP outperforms the other two auctions for benefiting market participants and incentivizing DERs, it is not necessary that the auctioneer prefers it as well. As discussed in Section 4.2, the auctioneer has no profit in UP due to the zero trading surplus, which is validated by our simulations as shown in Figure 4.7. According to the results, the auctioneer can achieve the most profit in MV, though the profit fluctuations of MV are much higher than VV's.

To further validate the results, we conduct four simulation epochs with the same input settings for each auction mechanism. We compute the average of 300 auction rounds for each simulation epoch. The summaries for energy clear quantity, total social welfare of all agents, total normalized rewards of all agents, and auctioneer's profit are presented in Table 4.2 to Table 4.5, respectively. We can clearly see that the UP auction has the best performance on average from agents' perspective and it can clears the most energy quantity for DERs. The MV auction can bring the auctioneer the most profit on average.

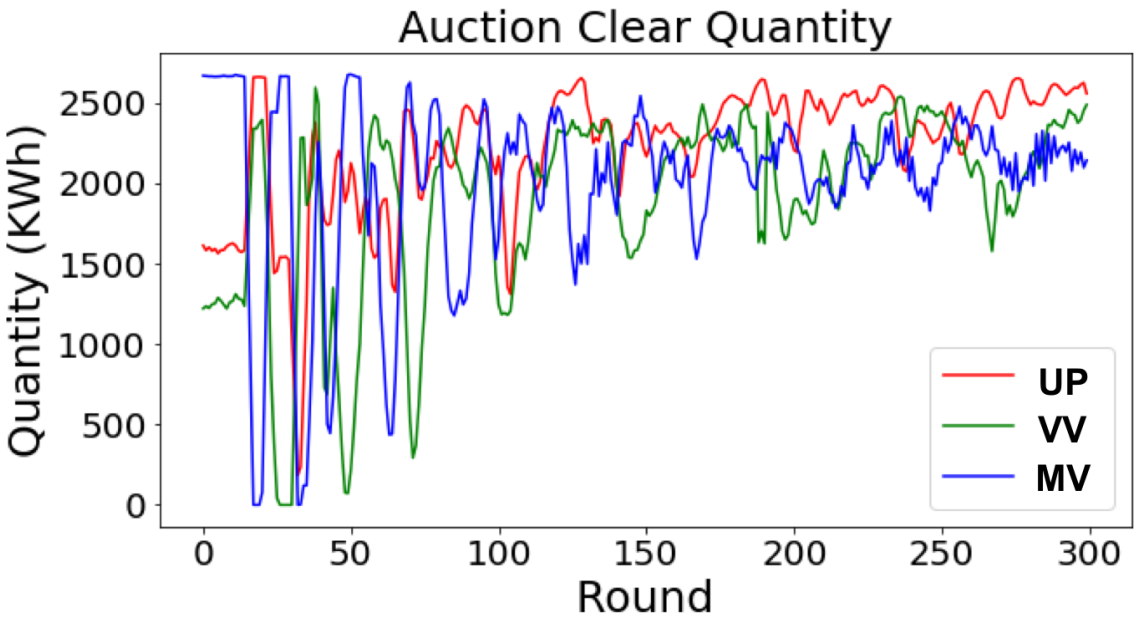

Figure 4.4.: Total clear energy quantities (KWh) in the auctions.

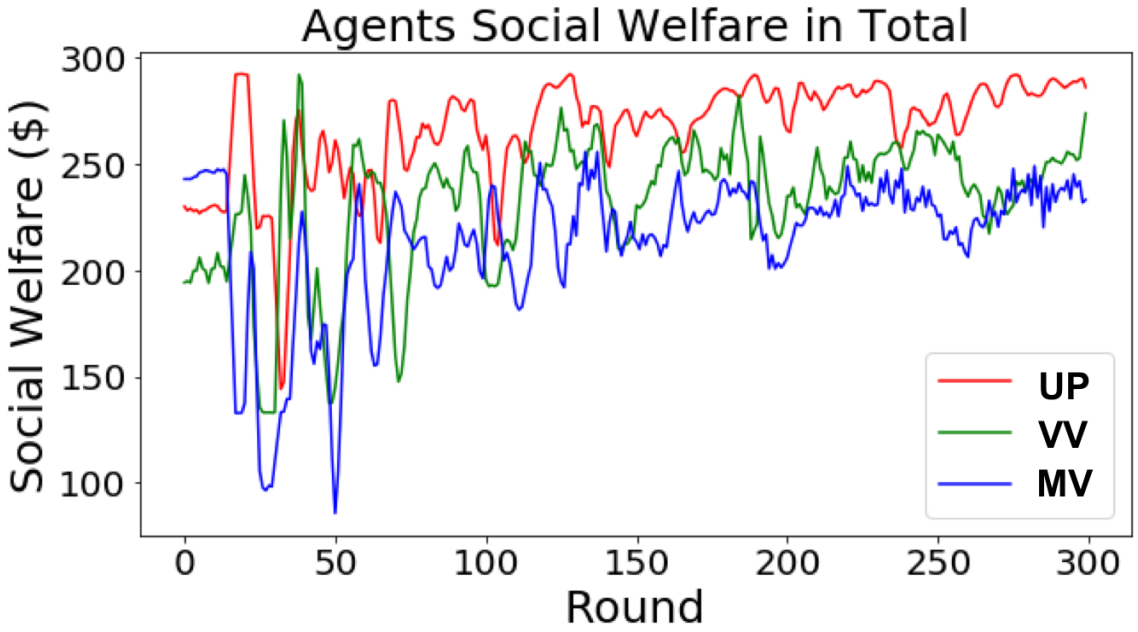

Figure 4.5.: Total social welfare ($) of all buyers and sellers in the auctions.

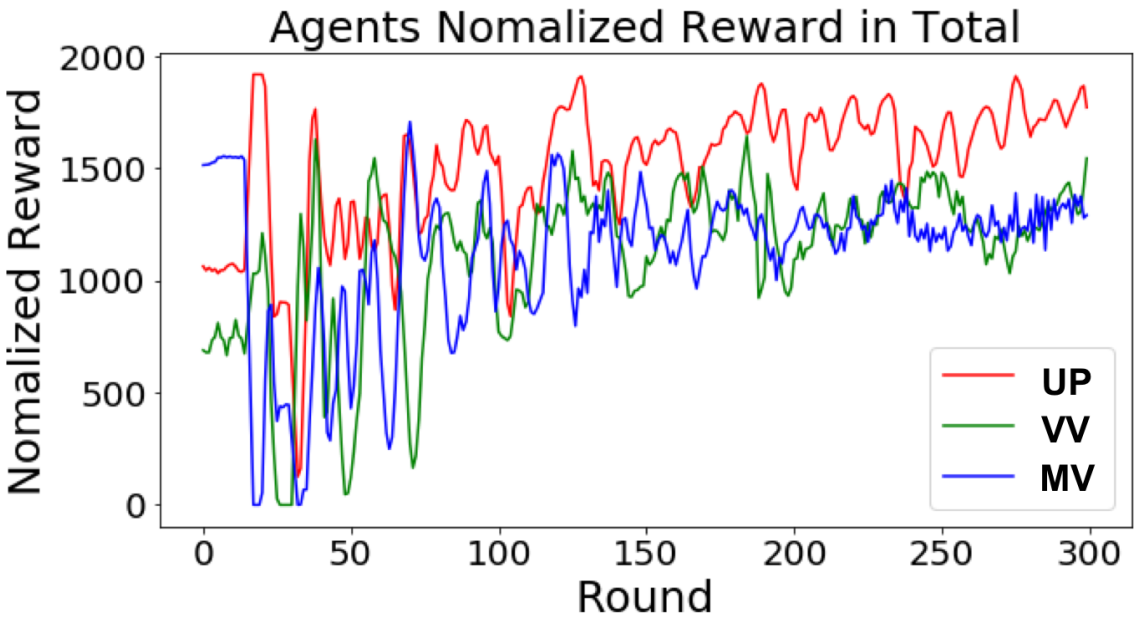

Figure 4.6.: Total normalized reward of all buyers and sellers in the auctions.

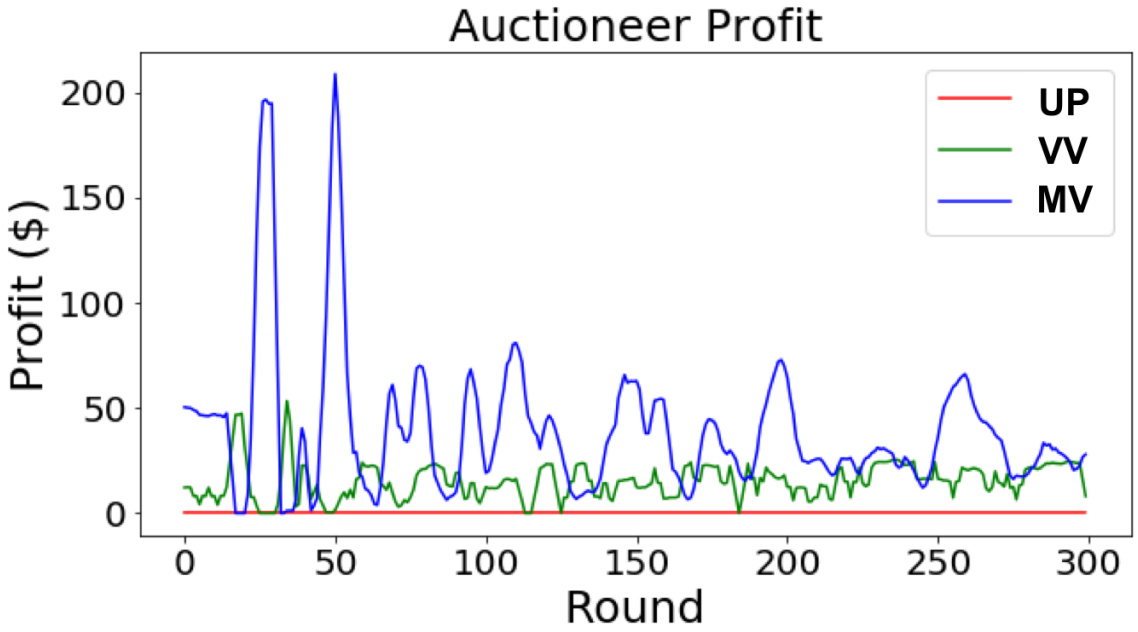

Figure 4.7.: Auctioneer's profit ($) in the auctions.

Table 4.2.: Average total clear quantity (KWh) of all agents in the auctions.

Auction	Sim 1	Sim 2	Sim 3	Sim 4	Average
UP	2238.63	2238.13	2234.67	2242.38	2238.45
VV	1876.11	1859.64	1887.67	1840.89	1866.08
MV	2031.05	2035.79	2018.3	2024.96	2027.52

Table 4.3.: Average total social welfare ($) of all agents in the auctions.

Auction	Sim 1	Sim 2	Sim 3	Sim 4	Average
UP	267.61	267.55	267.30	267.87	267.58
VV	230.19	229.94	231.77	229.13	230.26
MV	214.19	215.31	213.06	214.57	214.28

Table 4.4.: Average total normalized reward of all agents in the auctions.

Auction	Sim 1	Sim 2	Sim 3	Sim 4	Average
UP	1521.38	1520.91	1518.39	1523.36	1521.01
VV	1083.80	1083.24	1110.08	1075.95	1088.27
MV	1117.32	1127.74	1103.35	1113.70	1115.53

Table 4.5.: Average auctioneer profit ($) in the auctions.

Auction	Sim 1	Sim 2	Sim 3	Sim 4	Average
UP	0	0	0	0	0
VV	16.20	15.82	15.05	15.10	15.54
MV	40.96	40.10	41.26	40.26	40.64

5. CONCLUSIONS AND FUTURE WORK

In this dissertation, we study different decentralized approaches to implement real-time pricing and demand response in an energy market: the naive-response model, the adaptive-response model, the MAB-game model, and the LAR-game model. The resulting real-time prices from the naive-response model exhibit large variations on a daily basis, confirming the concerns raised in [38]. 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 game-theoretical approaches can achieve much greater benefits from a system perspective. Also, we establish the performance bounds for both MAB-game and LAR-game approaches, and we can see as agents in LAR-game utilize full information feedback, the overall system thus converges to more efficient outcomes.

The game-theoretical approaches introduced in this work, however, has several limitations. First and foremost, while its feature of not relying on any price forecasts (and only learns through the history under real-time pricing) 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 bandit learning algorithms. Second, the theoretical results 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 this dissertation to the case of exogenous uncertainty (faced by all agents) would be a significant endeavor.

We propose a MAB-game approach for market participants choosing price for their energy bid/ask in P2P double-side auctions. The bandit learning approach allows each individual agent to make a decision according to its own history other than its belief about other agents which is impractical and implausible to maintain under a large pop- ulation. We conduct simulations for the approach under three different double auction designs, and the results indicate the convergence of clear quantities, total social welfare and total normalized reward for agents. Moreover, the uniform-price double auction outperforms the other two in terms of market participants' benefits. For auctioneer, the maximum volume matching offers the highest profit. In future, potential research directions are studying about the robustness of the approach under large external uncertainties, and the interactions between auctions in different locations or distribution systems under transmission constraints.

REFERENCES

REFERENCES

[1] T. Mai, D. Sandor, R. Wiser, and T. Schneider, "Renewable electricity futures study: executive summary," National Renewable Energy Laboratory (NREL), Golden, CO., Tech. Rep., 2012.

[2] L. Lu, J. Tu, C.-K. Chau, M. Chen, and X. Lin, "Online energy generation scheduling for microgrids with intermittent energy sources and co-generation," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 1, pp. 53–66, Jun. 2013.

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

[4] J. M. Carrasco, L. García Franquelo, J. T. Bialasiewicz, E. Galván, R. C. Portillo Guisado, M. d. l. Á. Martín Prats, J. I. León, and N. Moreno-Alfonso, "Power-electronic systems for the grid integration of renewable energy sources: A survey," *IEEE Transactions on Industrial Electronics, 53 (4), 1002-1016.*, 2006.

[5] G. M. Masters, *Renewable and efficient electric power systems*. John Wiley & Sons, 2013.

[6] Y. Kabalci, "A survey on smart metering and smart grid communication," *Renewable and Sustainable Energy Reviews*, vol. 57, pp. 302–318, 2016.

[7] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 425–436, 2015.

[8] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

[9] L. Jiang and S. Low, "Multi-period optimal energy procurement and demand response in smart grid with uncertain supply," in *IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, 2011, pp. 4348–4353.

[10] "Assessment of demand response and advanced metering," Federal Energy Regulatory Commission, December, Tech. Rep., 2008.

[11] M. Milligan and B. Kirby, "Utilizing load response for wind and solar integration and power system reliability," National Renewable Energy Laboratory (NREL), Golden, CO., Tech. Rep., 2010.

[12] D. S. Callaway and I. A. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184–199, 2011.

[13] S. Lu, N. Samaan, R. Diao, M. Elizondo, C. Jin, E. Mayhorn, Y. Zhang, and H. Kirkham, "Centralized and decentralized control for demand response," in *ISGT 2011*. Ieee, 2011, pp. 1–8.

[14] P. Cappers, C. Goldman, and D. Kathan, "Demand response in us electricity markets: Empirical evidence," *Energy*, vol. 35, no. 4, pp. 1526–1535, 2010.

[15] G. Sharma, L. Xie, and P. Kumar, "Large population optimal demand response for thermostatically controlled inertial loads," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2013, pp. 259–264.

[16] S. Meyn, P. Barooah, A. Busic, and J. Ehren, "Ancillary service to the grid from deferrable loads: The case for intelligent pool pumps in florida," in *IEEE Conference on Decision and Control (CDC)*, 2013, pp. 6946–6953.

[17] H. Hao, T. Middelkoop, P. Barooah, and S. Meyn, "How demand response from commercial buildings will provide the regulation needs of the grid," in *IEEE Annual Allerton Conference on Communication, Control, and Computing*, 2012, pp. 1908–1913.

[18] F. Rahimi and A. Ipakchi, "Demand response as a market resource under the smart grid paradigm," *IEEE Transactions on smart grid*, vol. 1, no. 1, pp. 82–88, 2010.

[19] 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.

[20] L. Yang, J. Zhang, and D. Qian, "Risk-aware day-ahead scheduling and real-time dispatch for plug-in electric vehicles," in *IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 3026–3031.

[21] M. He, S. Murugesan, and J. Zhang, "A multi-timescale scheduling approach for stochastic reliability in smart grids with wind generation and opportunistic demand," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 521–529, 2013.

[22] Y. Xu and F. Pan, "Scheduling for charging plug-in hybrid electric vehicles," in *IEEE Conference on Decision and Control (CDC)*, 2012, pp. 2495–2501.

[23] I. C. Paschalidis, B. Li, and M. C. Caramanis, "A market-based mechanism for providing demand-side regulation service reserves," in *IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, 2011, pp. 21–26.

[24] 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.

[25] M. H. Albadi and E. F. El-Saadany, "Demand response in electricity markets: An overview," in *2007 IEEE power engineering society general meeting*. IEEE, 2007, pp. 1–5.

[26] ——, "A summary of demand response in electricity markets," *Electric power systems research*, vol. 78, no. 11, pp. 1989–1996, 2008.

[27] H. Zhong, L. Xie, and Q. Xia, "Coupon incentive-based demand response: Theory and case study," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1266–1276, 2012.

[28] G. Barbose, C. Goldman, R. Bharvirkar, M. Ting, and B. Neenan, "Real time pricing as a default or optional service for commercial and industrial customers: a comparative analysis of eight case studies," Lawrence Berkeley National Laboratory, CA, Tech. Rep., 2006.

[29] P. Samadi, H. Mohsenian-Rad, R. Schober, and V. W. Wong, "Advanced demand side management for the future smart grid using mechanism design," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1170–1180, 2012.

[30] A.-H. Mohsenian-Rad, V. W. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 320–331, 2010.

[31] W. Saad, Z. Han, H. V. Poor, and T. Başar, "A noncooperative game for double auction-based energy trading between phevs and distribution grids," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2011, pp. 267–272.

[32] J. Zarnikau and D. Thal, "The response of large industrial energy consumers to four coincident peak (4cp) transmission charges in the Texas (ERCOT) market," *Utilities Policy*, vol. 26, pp. 1–6, 2013.

[33] A. Faruqui, R. Hledik, and J. Palmer, *Time-varying and dynamic rate design*. Regulatory Assistance Project, 2012.

[34] 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*, vol. 15, no. 7, pp. 42–47, 1978.

[35] W. W. Hogan, "Electricity market design and efficient pricing: Applications for new england and beyond," *The Electricity Journal*, vol. 27, no. 7, pp. 23–49, 2014.

[36] S. Borenstein, "The long-run efficiency of real-time electricity pricing," *The Energy Journal*, pp. 93–116, 2005.

[37] M. Roozbehani, M. Dahleh, and S. Mitter, "On the stability of wholesale electricity markets under real-time pricing," in *IEEE Conference on Decision and Control (CDC)*, 2010, pp. 1911–1918.

[38] M. Roozbehani, M. A. Dahleh, and S. K. Mitter, "Volatility of power grids under real-time pricing," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 1926–1940, 2012.

[39] O. Dalkilic, A. Eryilmaz, and X. Lin, "Stable real-time pricing and scheduling for serving opportunistic users with deferrable loads," in *IEEE Annual Allerton Conference on Communication, Control, and Computing*, 2013, pp. 1200–1207.

[40] H. Mohsenian-Rad, "Optimal demand bidding for time-shiftable loads," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 939–951, 2014.

[41] J. H. Yoon, R. Baldick, and A. Novoselac, "Dynamic demand response controller based on real-time retail price for residential buildings," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 121–129, 2014.

[42] A. J. Conejo, J. M. Morales, and L. Baringo, "Real-time demand response model," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 236–242, 2010.

[43] K. M. Tsui and S.-C. Chan, "Demand response optimization for smart home scheduling under real-time pricing," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1812–1821, 2012.

[44] Z. Chen, L. Wu, and Y. Fu, "Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1822–1831, 2012.

[45] L. Huang, J. Walrand, and K. Ramchandran, "Optimal demand response with energy storage management," in *IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, 2012, pp. 61–66.

[46] Q. Wang, Y. Guan, and J. Wang, "A chance-constrained two-stage stochastic program for unit commitment with uncertain wind power output," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 206–215, 2012.

[47] J. Wang, M. Shahidehpour, and Z. Li, "Security-constrained unit commitment with volatile wind power generation," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1319–1327, 2008.

[48] L. Zhao and B. Zeng, "Robust unit commitment problem with demand response and wind energy," in *IEEE Power and Energy Society General Meeting*, 2012, pp. 1–8.

[49] R. Jiang, J. Wang, and Y. Guan, "Robust unit commitment with wind power and pumped storage hydro," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 800–810, 2012.

[50] Q. Wang, J.-P. Watson, and Y. Guan, "Two-stage robust optimization for nk contingency-constrained unit commitment," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2366–2375, 2013.

[51] R. Barth, H. Brand, P. Meibom, and C. Weber, "A stochastic unit-commitment model for the evaluation of the impacts of integration of large amounts of intermittent wind power," in *IEEE International Conference on Probabilistic Methods Applied to Power Systems*, 2006, pp. 1–8.

[52] 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*, vol. 26, no. 4, pp. 2197–2206, 2011.

[53] J. M. Morales, A. J. Conejo, and J. Pérez-Ruiz, "Economic valuation of reserves in power systems with high penetration of wind power," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 900–910, 2009.

[54] J. Xiao, "Grid integration and smart grid implementation of emerging technologies in electric power systems through approximate dynamic programming," Ph.D. dissertation, Purdue University, 2013.

[55] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[56] A. Veit, Y. Xu, R. Zheng, N. Chakraborty, and K. Sycara, "Demand side energy management via multiagent coordination in consumer cooperatives," *Journal of Artificial Intelligence Research*, vol. 50, pp. 885–922, 2014.

[57] D. Guo, W. Zhang, G. Yan, Z. Lin, and M. Fu, "Decentralized control of aggregated loads for demand response," in *American Control Conference (ACC), 2013*. IEEE, 2013, pp. 6601–6606.

[58] 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*. International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 5–12.

[59] 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.

[60] T. Roughgarden, "Intrinsic robustness of the price of anarchy," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 2009, pp. 513–522.

[61] D. J. Foster, Z. Li, T. Lykouris, K. Sridharan, and E. Tardos, "Learning in games: Robustness of fast convergence," in *Advances in Neural Information Processing Systems*, 2016, pp. 4734–4742.

[62] H. Jiayi, J. Chuanwen, and X. Rong, "A review on distributed energy resources and microgrid," *Renewable and Sustainable Energy Reviews*, vol. 12, no. 9, pp. 2472–2483, 2008.

[63] M. F. Akorede, H. Hizam, and E. Pouresmaeil, "Distributed energy resources and benefits to the environment," *Renewable and sustainable energy reviews*, vol. 14, no. 2, pp. 724–734, 2010.

[64] P. Huang, A. Scheller-Wolf, and K. Sycara, "Design of a multi–unit double auction e–market," *Computational Intelligence*, vol. 18, no. 4, pp. 596–617, 2002.

[65] J. Niu and S. Parsons, "Maximizing matching in double-sided auctions," 2013, available at arXiv.org: https://arxiv.org/abs/1304.3135, last revised: Feb 11, 2013.

[66] D. Krishnamurthy, W. Li, and L. Tesfatsion, "An 8-zone test system based on iso new england data: Development and application," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 234–246, 2015.

[67] "Regional transmission organizations (rto)/independent system operators (iso)," 2019, available at arXiv.org: https://www.ferc.gov/industries/electric/indus-act/rto.asp.

[68] Agera Energy, "Independent system operator (iso)," https://www.ageraenergy.com/energy-terms/independent-system-operator-iso/.

[69] M. Roozbehani, M. Dahleh, and S. Mitter, "Dynamic pricing and stabilization of supply and demand in modern electric power grids," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*. IEEE, 2010, pp. 543–548.

[70] T. Zheng and E. Litvinov, "Ex post pricing in the co-optimized energy and reserve market," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1528–1538, 2006.

[71] ComEd, "Comed's real data of day-ahead and real-time rtp prices," https://hourlypricing.comed.com/live-prices/.

[72] M. Ilic, J. W. Black, and J. L. Watz, "Potential benefits of implementing load control," in *Power Engineering Society Winter Meeting, 2002. IEEE*, vol. 1. IEEE, 2002, pp. 177–182.

[73] D. Fudenberg and J. Tirole, "Game theory," *MIT press, Cambridge, Massachusetts*, vol. 393, no. 12, p. 80, 1991.

[74] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, May 2002.

[75] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[76] A. Mahajan, D. Teneketzis, A. Hero, D. Castanon, D. Cochran, and K. Kastella, "Foundations and applications of sensor management," *Chapter Multi-armed Bandit Problems*, pp. 121–151, 2007.

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

[78] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *International Conference on Machine Learning*, 2014, pp. 100–108.

[79] T. Roughgarden, "Intrinsic robustness of the price of anarchy," *Journal of the ACM (JACM)*, vol. 62, no. 5, p. 32, 2015.

[80] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *Annual Symposium on Theoretical Aspects of Computer Science*. Springer, 1999, pp. 404–413.

[81] G. Christodoulou and E. Koutsoupias, "The price of anarchy of finite congestion games," in *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. ACM, 2005, pp. 67–73.

[82] E. Hazan *et al.*, "Introduction to online convex optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[83] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, "Fast convergence of regularized learning in games," in *Advances in Neural Information Processing Systems*, 2015, pp. 2989–2997.

[84] T. Lykouris, V. Syrgkanis, and É. Tardos, "Learning and efficiency in games with dynamic population," in *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2016, pp. 120–129.

[85] E. Hazan and C. Seshadhri, "Efficient learning algorithms for changing environments," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 393–400.

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

[87] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *Journal of climate and applied meteorology*, vol. 23, no. 8, pp. 1184–1195, 1984.

[88] M. Brower *et al.*, "Development of eastern regional wind resource and wind plant output datasets," *Rep. No. NREL/SR-550*, vol. 46764, 2009.

[89] C. L. Archer and M. Z. Jacobson, "Supplying baseload power and reducing transmission requirements by interconnecting wind farms," *Journal of Applied Meteorology and Climatology*, vol. 46, no. 11, pp. 1701–1717, 2007.

[90] B. Deler and B. L. Nelson, "Input modeling and its impact: modeling and generating multivariate time series with arbitrary marginals and autocorrelation structures," in *Proceedings of the 33nd conference on Winter simulation*. IEEE Computer Society, 2001, pp. 275–283.

[91] 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*. International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 5–12.

[92] J. A. Lesser and X. Su, "Design of an economically efficient feed-in tariff structure for renewable energy development," *Energy policy*, vol. 36, no. 3, pp. 981–990, 2008.

[93] D. Friedman, *The double auction market: institutions, theories, and evidence*. Routledge, 2018.

[94] J. Nicolaisen, V. Petrov, and L. Tesfatsion, "Market power and efficiency in a computational electricity market with discriminatory double-auction pricing," *IEEE transactions on Evolutionary Computation*, vol. 5, no. 5, pp. 504–523, 2001.

[95] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[96] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*. Springer, 2008, pp. 121–151.

[97] D. Fudenberg, F. Drew, D. K. Levine, and D. K. Levine, *The theory of learning in games*. MIT press, 1998, vol. 2.

[98] U.S. Energy Information Admisnistration (EIA), "Residential energy consumption survey (RECS)," available at https://www.eia.gov/consumption/residential/data.

[99] National Renewable Energy Lab (NREL), Golden, CO (United States), "System advisor model (SAM)," avaible at https://sam.nrel.gov.

[100] P. Gilman, "Sam photovoltaic model technical reference," National Renewable Energy Lab (NREL), Golden, CO (United States), Tech. Rep., 2015, avaible at https://www.osti.gov/biblio/1215213.

[101] N. Blair, A. P. Dobos, J. Freeman, T. Neises, M. Wagner, T. Ferguson, P. Gilman, and S. Janzou, "System advisor model, sam 2014.1.14: General description," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2014, avaible at https://www.osti.gov/biblio/1126294.