informs
https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *Operations Research*

ISSN 0030-364X, EISSN 1526-5463

Decentralized Integration of Grid Edge Resources into Wholesale Electricity Markets via Mean-field Games

Feng Chen
Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN, USA fc123good@gmail.com

Andrew L. Liu
Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN, USA andrewliu@purdue.edu.

Abstract. Grid edge resources refer to distributed energy resources (DERs) located on the consumer side of the electrical grid, controlled by consumers rather than utility companies. Integrating DERs with real-time electricity pricing can better align distributed supply with system demand, improving grid efficiency and reliability. However, DER owners, known as prosumers, often lack the expertise and resources to directly participate in wholesale energy markets, limiting their ability to fully realize the economic potential of their assets. Meanwhile, as DER adoption grows, the number of prosumers participating in the energy system is expected to increase significantly, creating additional challenges in coordination and market participation.

To address these challenges, we propose a mean-field game framework that enables prosumers to autonomously learn optimal decision policies based on dynamic market prices and their variable solar generation. Our framework is designed to accommodate heterogeneous agents and demonstrates the existence of a mean-field equilibrium (MFE) in a wholesale energy market with many prosumers. Additionally, we introduce an algorithm that automates prosumers' resource control, facilitating real-time decision-making for energy storage management. Numerical experiments suggest that our approach converges towards an MFE and effectively reduces peak loads and price volatility, especially during periods of external demand or supply shocks. This study highlights the potential of a fully decentralized approach to integrating DERs into wholesale markets while improving market efficiency.

Funding: This research was supported by DOE DE-OE0000921 and NSF ECCS-2129631.

Key words: solar, energy storage, DER integration, mean field games, mulitagent systems, transactive energy, demand response

1. Introduction

The growing adoption of distributed energy resources (DERs), such as rooftop solar panels and energy storage, presents significant opportunities to enhance grid efficiency and resilience. To fully leverage these resources, integrating them into wholesale energy markets is essential for enabling more flexible and reliable grid operations. However, traditional market structures impose high entry barriers for small-scale DER participation due to minimum size requirements and complex market rules. To address these challenges, the Federal Energy Regulatory Commission (FERC) issued Order 2222 (FERC 2020), mandating that DERs be granted access to wholesale markets. While this regulatory change facilitates DER integration, effective mechanisms for small prosumers' (aka DER owners') participation remain unclear. A key concern is that prosumers often hesitate to relinquish direct control of their assets to aggregators – entities that pool multiple small-scale DERs to meet market size thresholds. Existing literature primarily focuses on how aggregators bid into wholesale markets on behalf of their customers and how contracts for direct load control can be structured, largely overlooking decentralized alternatives that preserve prosumer autonomy.

Our work addresses this gap with three key contributions. First, we develop a fully decentralized framework that enables prosumers to optimize DER operations independently, guided by real-time locational marginal prices (LMPs). Second, we formulate this problem as a mean-field game, where prosumers operate under a mean-field assumption – treating their individual bids as having negligible impact on LMPs. The collective bids of all prosumers, however, can influence market outcomes. Within this framework, we prove the existence of a mean-field equilibrium (MFE) for an infinite population of agents and an ϵ-Markov-Nash equilibrium for a large but finite population. Third, we propose a scalable, low-overhead learning algorithm that allows prosumers to adapt their strategies based on LMP fluctuations, supporting real-time storage management and bid optimization without requiring centralized coordination. Numerical results demonstrate that our approach reduces price volatility and peak loads, even in the presence of supply or demand shocks, thereby improving market stability.

We want to emphasize that our framework is prescriptive rather than descriptive – it is designed to prescribe optimal prosumer actions rather than merely describe observed behaviors. While our primary focus is electricity markets, the proposed mean-field approach is broadly applicable to any multiagent system where decisions are influenced by a shared external variable, such as market prices or other aggregate signals. Furthermore, the algorithm developed in this work is generalizable

and extends beyond energy markets to settings where large-scale agent interactions shape market dynamics in different sectors.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on models and algorithms for multiagent bidding in wholesale markets. Section 3 outlines the wholesale electricity market model. Section 4 formulates the prosumer optimization problem. Section 5 integrates wholesale market dynamics and prosumer decision-making into a mean-field game framework and establishes the existence of a mean-field equilibrium. Section 6 describes the proposed learning algorithm. Section 7 presents the numerical experiments and results. Finally, Section 8 concludes with a summary and potential future research directions.

2. Literature Review

There is extensive literature on individual agents' strategies for bidding into wholesale markets, using either optimization-based or learning-based approaches. In contrast, we focus on systems involving multiple agents. Existing research in this area can be broadly divided into two categories: agent-based simulations and game-theoretic approaches. Agent-based simulation (ABS) is widely used to model bidding behaviors in wholesale energy markets, offering a natural approach for studying multi-agent systems. Reviews such as Ringler et al. (2016), Sensfuß et al. (2007), Guerci et al. (2010) highlight its role in this field, with early works Price (1997) and later studies North et al. (2002), Macal et al. (2014), Shafie-khah and Catalão (2014) advancing the method. A key aspect of ABS is defining appropriate behavioral models for each agent type, creating a heterogeneous artificial economy (Guerci et al. 2010). While agents could be modeled as utility-maximizers considering other agents' actions – akin to game theory – ABS often avoids this complexity due to computational challenges.

An alternative, introduced by Roth and Erev (1995), uses a simpler adaptive strategy based on action propensities, termed reinforcement learning (RL). Unlike modern RL (as presented in Sutton and Barto (2018)), this model updates action probabilities based on past rewards without state-based feedback or value functions. Despite its simplicity, it effectively predicts human behavior in certain games (Erev and Roth 1998), inspiring adaptive multi-agent learning studies in energy markets, such as Bunn and Oliveira (2001), Sun and Tesfatsion (2007). Similar adaptive methods Visudhiphan and Ilic (1999), Ramchurn et al. (2011) explore bidding behavior and dynamic pricing responses but lack the ability to handle intertemporal decisions, such as energy storage management, which is central to this work.

With advancements in modern RL theories and algorithms, multi-agent reinforcement learning (MARL) offers sophisticated methods that avoid preset behavioral assumptions, relying only on utility maximization over time. Naturally, MARL has been applied to model agent participation in energy markets Du et al. (2021), Ye et al. (2022). However, MARL still faces two significant challenges: a lack of theoretical guarantees – specifically, whether multi-agent interactions will converge to an equilibrium, a steady state, or result in chaotic behavior – in complex environments, and scalability issues, particularly in large systems involving thousands or more agents.

On the game theory side, there is a rich body of literature analyzing bidding strategies and market interactions. Nash-Cournot models, where agents act as quantity setters to maximize their own profits while accounting for market-clearing prices based on total quantities, are widely used in electricity market studies to analyze market power and strategic interactions among generators (for example, Hobbs (1986), Willems (2002), Neuhoff et al. (2005), Metzler et al. (2003)). Another widely used framework is the supply function equilibrium (SFE), where agents compete by submitting supply functions instead of fixed quantities. SFE models are particularly suitable for wholesale electricity markets, as they capture the price-quantity relationship under market clearing (see, for example, Baldick et al. (2004), Rudkevich (2005), Anderson and Philpott (2002), Anderson and Xu (2005), Holmberg and Newbery (2010)). While these models provide valuable insights, they are inherently static and fail to capture the intertemporal dynamics that are critical in energy systems with energy storage.

Dynamic game-theoretic models address some limitations of static frameworks by incorporating intertemporal decision-making, allowing for the analysis of strategic behaviors over time. These models have been extensively studied in the economics and game-theory literature and applied to examine the strategic behaviors of electricity market participants. For example, works such as Liu and Hobbs (2013), Liu (2010), Fabra and Toro (2005), Anderson and Cau (2011) investigate repeated interactions among power producers, focusing on equilibrium concepts of subgame perfect equilibrium. However, these models typically assume complete and perfect information, meaning that all participants have full knowledge of each other's payoff functions, strategies, and the entire history of the game. In this work, however, the games involve incomplete information, as consumers and prosumers may lack precise knowledge of others' payoff functions, be unable to observe their actions, or not have access to the full history of the game. The standard equilibrium concept for such dynamic games is the Perfect Bayesian Nash Equilibrium (PBNE) (see Fudenberg and Tirole (1991)). PBNE requires players to update their beliefs using Bayes' rule and to select strategies that

maximize their expected payoffs across all possible game histories. While theoretically appealing, PBNE is impractical for real-world applications, as it assumes that agents possess an unrealistic level of strategic sophistication. Moreover, the computational complexity of these models grows significantly as the number of agents increases, making them difficult to scale to large systems.

Mean-field game (MFG) theory[1] offers a promising solution to the challenges posed by dynamic games with incomplete information by approximating interactions among a large number of agents through an aggregate mean-field effect. MFGs provide several advantages over PBNE, particularly in terms of computational tractability and scalability. Extensive research has explored the existence and uniqueness of mean-field equilibria (MFE) (Adlakha and Johari (2013), Light and Weintraub (2022), Saldi et al. (2018)). In addition to theoretical advancements, provably convergent algorithms for computing MFE have been developed, including Guo et al. (2019), Gu et al. (2024), Xie et al. (2021).

Building on the theoretical foundations, MFGs have been applied across various domains, including energy markets, where decentralized decision-making and large populations of interacting agents play a critical role. Notable applications include electricity demand management (Bagagiolo and Bauso 2014) and electric vehicle (EV) charging coordination (Tajeddini and Kebriaei 2018, Zhu et al. 2016). MFGs have also been used to study electricity price dynamics, a topic closely aligned with the focus of this paper. However, all these energy-related MFG applications adopt continuous-time models, which can be impractical for electricity markets where pricing and market clearing occur at discrete intervals (unlike financial markets). Additionally, current models often overlook interactions between energy system operators and the influence of transmission constraints, which are essential factors in determining electricity prices. we applied discrete-time MFGs to analyze how DERs' decentralized actions influence wholesale electricity markets in He and Liu (2024), though without theoretical foundations. In this paper, we expand on this by rigorously studying how the collective and decentralized actions of prosumers with solar PVs and energy storage affect wholesale electricity prices. We establish formal conditions for the existence of MFE and propose a scalable heuristic algorithm, making it well-suited for large-scale energy systems integrating increasing levels of decentralized resources. A major advantage of our algorithm is its minimal computational and memory requirements for each agent, unlike distributed optimization methods such as the alternating direction method of multipliers (ADMM), which require solving (proximal)

[1] We focus here on discrete-time MFGs.

optimization problems at each step. This makes our approach highly scalable and well-suited for large systems with thousands of DERs. By facilitating control automation with low overhead, our algorithm helps unlock tangible benefits for DER owners, supports DER integration into wholesale markets, and enhances scalability.

Our work addresses critical gaps in the study of decentralized decision-making in energy. A defining feature of our model is the use of continuous state and action spaces (such as energy storage charging/discharging), providing a more realistic representation of prosumer behavior compared to discretized models. We further extend the framework to include multiple heterogeneous agent types, capturing the diversity in prosumer characteristics and decision-making. This aspect draws inspiration from Mondal et al. (2022), which incorporates heterogeneity in a mean-field control (MFC) setting. However, while their work focuses on cooperative agents, our model explores non-cooperative interactions. This market-driven approach extends beyond energy markets to any domain where aggregate behavior shapes price signals. Leveraging this structure, the highly scalable heuristic algorithm developed in this work can be applied across a wide range of settings.

3. Wholesale Energy Market Operations

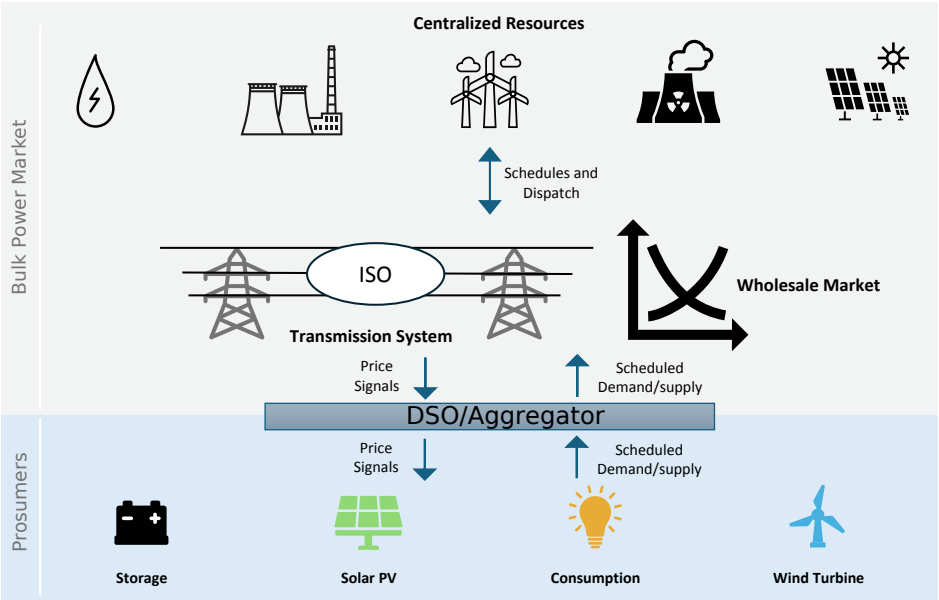


Figure 1 Conceptual framework of a wholesale energy market with aggregators participation

In this section, we first present the optimization problem solved by an Independent System Operator (ISO) for a wholesale energy market. As illustrated in Fig. 1, in each time period (such

as an hour), the ISO collects supply and demand bids and runs optimization problems to match the supply and demand with the lowest cost, subject to various engineering and transmission network constraints. The marginal costs of supplying one additional unit of demand at each node, known as the locational marginal prices or LMPs, can be calculated based on the shadow prices of constraints on supply-demand balancing and transmission capacity constraints. We show a key result in this section regarding the Liptschitz continuity of the LMPs with respect to energy demand.

Consider a (bidirectional) power system network with N nodes and L transmission lines. For simplicity, we assume that each node $n \in \{1, ..., N\}$ has only one energy supplier[2] with an operation cost function $C_n(\cdot)$. The node has I_n agents, including both consumers and prosumers. One hour before the operating period t, the i-th agent at node n submits its energy demand/supply bid, $b^n_{i,t}$, for period t to the ISO. Note that $b^n_{i,t}$ represents the net demand. For consumers, this value is simply their actual energy demand with $b^n_{i,t} > 0$. For prosumers, $b^n_{i,t}$ represents net energy demand, calculated as actual demand plus the energy used for charging batteries, minus PV generation and any energy withdrawn from batteries. This net demand can be either positive or negative, where a negative value indicates that the prosumer is supplying energy back to the grid. The decision-making problem for prosumers' bidding is presented in the next section.

Throughout the paper, we make the blanket assumption that the total net demand of all agents in the entire system is positive, that is, $\sum_{n=1}^{N} \sum_{i=1}^{I_n} b^n_{i,t} > 0, \quad \forall\ t \in \{1, ..., \}$. This assumption is reasonable, as it will likely take considerable time in the future before prosumers can meet all consumer energy demands and still produce surplus energy.

In the hour-ahead market, the ISO solves an optimization problem, known as economic dispatch (ED), for the upcoming operating period t. This optimization determines the amount of real power

[2] If there are multiple suppliers, we can assume that each supplier is on a separate node with such nodes connected by transmission lines of unlimited capacities.

to be dispatched from each electric power-generating resource to match the total demand as follows:

$$\underset{\mathbf{g}_t}{\text{minimize}} \ \sum_{n=1}^{N} C_n(g_t^n) \tag{1}$$

$$\text{subject to} \ \sum_{n=1}^{N} g_t^n \geq \sum_{n=1}^{N}\sum_{i=1}^{I_n} b_{i,t}^n \tag{2}$$

$$-\widehat{\mathrm{F}}_l \leq \sum_{n=1}^{N} \mathrm{PTDF}_{l,n}(g_t^n - \sum_{i=1}^{I_n} b_{i,t}^n) \leq \widehat{\mathrm{F}}_l,$$

$$\forall l \in \{1, ..., L\} \tag{3}$$

$$0 \leq g_t^n \leq \widehat{\mathrm{G}}_n, \quad \forall n \in \{1, ..., N\}, \tag{4}$$

where $\mathbf{g}_t := \{g_t^n\}_{n=1}^N$ is the collection of decision variables, representing the energy generation at node n in time period t. Constraint (2) specifies that the total supply must not be less than the total demand, often referred to as the supply and demand balancing constraint. Constraint (3) represents the transmission network capacity constraints, with the capacity limit denoted by $\widehat{\mathrm{F}}_l$. The network, which is assumed to be a connected graph, is modeled as a hub-spoke network, in which energy sent from node n to n' is assumed to be routed from n to a hub (an arbitrary node in the system) first and from the hub to n'. The parameter $\mathrm{PTDF}_{l,n}$ in (3) represents the power transfer distribution factor, which indicates the fraction of power injected at node n that flows through line l.[3] Last, $\widehat{\mathrm{G}}_n$ in (4) represents the generation capacity for the power plant at node n.

To write out the exact formula of nodal electricity prices, aka the LMPs, we first use $\mathcal{L}$ to denote the Lagrangian function of the ED problem. For the ease of argument, we use B_h^t to denote the aggregate demand at node n in time period h; that is, $B_t^n = \sum_{i=1}^{I_n} b_{i,t}^n$. Let λ denote the dual variable associated with constraint (2), and $\overline{\mu}_l$ and $\underline{\mu}_l$ be the dual variables corresponding to (3). Then the LMPs, denoted by $P_t^n(B_t^1, \ldots, B_t^N)$ for $n = 1, \ldots, N$ at time t, are the derivatives of the Lagrangian function with respect to the demand:

$$\begin{aligned}\mathrm{LMP}_t^n &:= P^n(B_t^1, \ldots, B_t^N) \\ &= \frac{\partial \mathcal{L}}{\partial B_t^n} = \lambda - \sum_{l=1}^{L} \mathrm{PTDF}_{l,n}(\overline{\mu}_l - \underline{\mu}_l).\end{aligned} \tag{5}$$

[3] For simplicity, we ignore transmission losses in this formulation. However, they can be incorporated as long as the resulting formulation remains a convex optimization problem. In that case, all results presented in this work still hold.

To establish the main result of this paper, which is the existence of an MFE of the multiagent system, it is crucial to prove that the LMPs are Lipschitz continuous with respect to the demand vector $\mathbf{B}_t := (B_t^1, \ldots, B_t^N)$. Achieving this requires an assumption regarding the constraint qualification for the ED problem. We state this assumption below and then present the Lipschitz continuity result.

ASSUMPTION 1. *(LICQ) Let $X(\mathbf{B}_t)$ denote the feasible region of the ED problem* (1) – (4)*. Define the set $\mathcal{F}_B$ such that for all $\mathbf{B}_t \in \mathcal{F}_B$, $X(\mathbf{B}_t) \neq \emptyset$. We assume that for all t and for all $\mathbf{B}_t \in \mathcal{F}_B$, the linear independence constraint qualification (LICQ) holds at all points in $X(\mathbf{B}_t)$.*

PROPOSITION 1. *Assume that the generation cost function $C_n(\cdot)$ in* (1) *is a strongly convex quadratic function in the form of $C_n(g) = \frac{1}{2}\alpha_n g^2 + \beta_n g + \gamma_n$, with $\alpha_n > 0$ for all $n = 1, \ldots, N$. Under Assumption 1, with $\mathbf{B}_t \in \mathcal{F}_B$, the LMP at each node $n = 1, \ldots, N$, $P^n(\mathbf{B}_t)$, is a single-valued function and Lipschitz continuous with respect to $\mathbf{B}_t$.*

The proof is in Online Appendix A.1.

4. A Prosumer's Markov Decision Process

The previous section focuses on the system operator's optimization problem. In this section, we shift the focus to how individual agents participate in a wholesale market. We first introduce a model for a single agent who makes *repeated* decisions regarding the charging and discharging of their energy storage over time, in response to real-time pricing tied to the LMPs. The agents make their decisions under the assumption that the system is in an MFE due to the large number of agents. We then show that an MFE can indeed emerge with heterogeneous agents holding this belief. This is a direct extension of our earlier work in Zhao et al. (2018) where each agent solves a multiarmed bandit problem, which cannot accommodate intertemporal decisions.

4.1. Assumptions on the Agents

To accommodate agents' heterogeneity, we assume that each consumer or prosumer has a type $\theta \in \Theta$, with Θ being a finite set. These types can include characteristics such as location (e.g., agents at different nodes in the transmission network belong to different types), varying solar PV capacities, battery capacities or types, and distinct load profiles. We assume that agents of the same type are homogeneous in their payoff function, state transition function, battery capacity, PV generation profile, and load distribution. Specifically, each agent of type θ has a battery capacity $\overline{e}^\theta = \frac{\overline{C}^\theta}{I^\theta}$, where I^θ is the number of agents of type θ, and $\overline{C}^\theta$ is the aggregated battery capacity of all type-θ agents. This definition ensures that as I^θ approaches infinity, each individual's capacity becomes infinitesimally small, yet the aggregate capacity remains well-defined and finite.

4.2. Single-agent's Dynamic Optimization

In the following, we provide the key elements in building an individual agent's decision-making model, with a given agent type $\theta \in \Theta$.

Action. At each time period t, agent i determines the fraction of energy to charge or discharge from their battery, expressed as a percentage of battery capacity and denoted by $a_{i,t} \in \mathcal{A} := [-1, 1]$. A positive value of $a_{i,t}$ signifies a charging action, whereas a negative value indicates discharging.

State. The state of an agent consists of three elements: the net load, the state of charge (SoC) of the energy storage, and time of day. The net load, which is a random variable, is defined as the firm (or inflexible) demand minus the energy output from solar PVs. We assume that agents of the same type share an identical daily net load shape, representing the expected value of the net load at the corresponding time of the day. Let $Q^{\theta}_{i,t}$ denote the net load for agent i at time period t, where Q is used to represent 'quantity.' Since actions (and later, the SoC) are defined as percentages, it is convenient to consider net load as a percentage as well. We introduce the ratio $q^{\theta}_{i,t} := Q^{\theta}_{i,t}/\bar{e}^{\theta}$, where $\bar{e}^{\theta}$ is the storage capacity as defined in Section 4.1. The transition from $q^{\theta}_{i,t}$ to $t+1$ is assumed to be purely driven by weather conditions and by random noise, which accounts for variations in real-time electricity usage among agents. Mathematically speaking, we have that

$$q^{\theta}_{i,t} = \omega^{\theta}_{t} + \zeta^{\theta}_{i,t}, \tag{6}$$

where both ω^{θ}_{t} and $\zeta^{\theta}_{i,t}$ are random variables. The first term, ω^{θ}_{t}, represents weather-related randomness and is location-specific (depending on the type θ) but not agent-specific (hence, no agent index i). The second term, $\zeta^{\theta}_{i,t}$, represents agent-specific random noise in electricity demand. Both variables are assumed to have compact supports, as each agent's electricity demand and PV/storage capacity are finite.

For the SoC of energy storage, we use $e_{i,t} \in \mathcal{E} \equiv [0, 1]$ to denote the fraction of remaining energy in the battery at the beginning of period t for agent i. The state transition of the SoC, denoted by $E(\cdot, \cdot)$, can be expressed as:

$$\begin{aligned} e_{i,t} &:= E(e_{i,t-1}, a_{i,t-1}) \\ &= \max\{\min\{e_{i,t-1} + a_{i,t-1}, 1\}, 0\}, \ t = 1, 2, \ldots, \end{aligned} \tag{7}$$

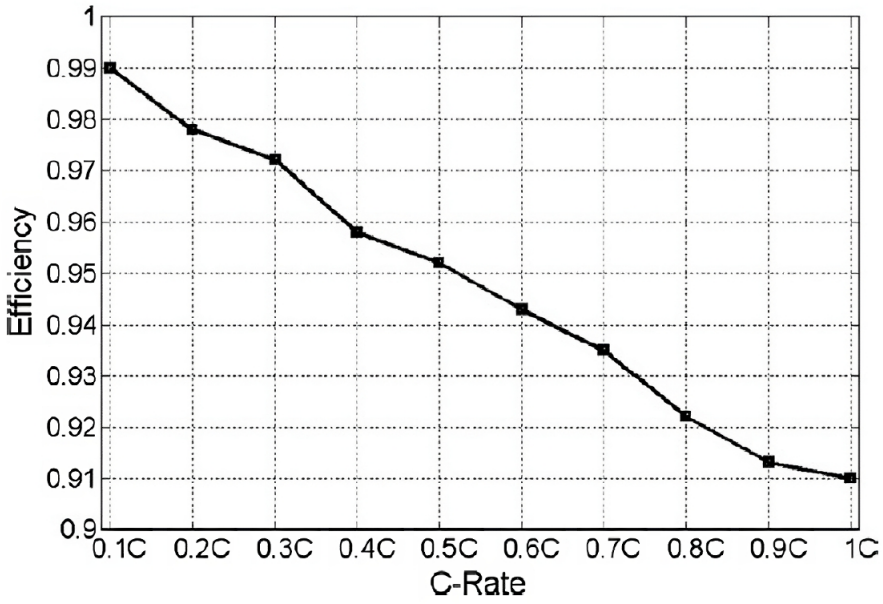
with the 'max' and 'min' operations to ensure that the actions will not lead to over-charging or over-discharging of the battery. In the following, we use $\mathcal{E} = [-1, 1]$ to denote the general space of SoC.

For the time of day, $h = 1, \ldots, H$, we account for the fact that an agent's policy should vary throughout the day. For example, even if the state of charge of the energy storage is the same at 10 AM and 6 PM in a day, the corresponding optimal strategy should be different. At 10 AM, solar energy is generally abundant, and household electricity usage is typically lower, as many people are at work. In this case, a good strategy might be to charge the battery to full. On the other hand, at 6 PM, solar energy is diminishing, and people are returning home, causing residential energy demand to increase. Even with the decrease in commercial and industrial loads at that time, the overall energy demand is expected to rise in the early evening. Hence, a good strategy at this time might be to discharge the energy storage. For a general time period index t, we use $T_{day}(t)$ to denote the time of day of t, and denote the set of times of day as $\mathcal{H}$, where $\mathcal{H} = \{1, \ldots, H\}$.

In the following, we treat $e_{i,t}$ and $T_{day}(t)$ as the state variable, denoted by the generic notation $s_{i,t}$, while considering the random variable $q^{\theta}_{i,t}$ as exogenous. Note that the state of charge transition equation (7) does not involve any uncertainties; that is, the net load $q^{\theta}_{i,t}$ does not directly affect the state transition. This is a specific modeling choice, which we will explain further after introducing the agents' bid functions in (9). This deterministic transition simplifies both the analysis and algorithm design in the subsequent sections. Additionally, the transition of the time of day is trivially deterministic. We use $Tr(s, a)$ to denote the general state transition, which includes both the state of charge transition and the time of day transition, which simply increments by one (that is, moves to the next time of day). It is understood that when $T_{day}(t) = H$, the time of day cycles back to 1 in $t + 1$, representing the start of the next day.

Charging/Discharging Efficiency. For most energy storage batteries, both charging and discharging efficiencies decrease as the respective rates increase. As demonstrated in Figure 2 (taken from (Amoroso and Cappuccino 2012)), the efficiency of a lithium-ion battery approximately follows a linear relationship with the charging rate. We model the energy charged to or discharged from the battery at a constant rate of $\frac{a_{i,t}}{\Delta t}$ for each period, where $a_{i,t}$ is agent i's charging/discharging action as defined earlier, and Δt represents the duration of the period. Consequently, we assume that charging

Figure 2 Experimental results showing the dependency of charging efficiency on the charging rate for a Li-ion cell (C represents battery capacity) Amoroso and Cappuccino (2012)



efficiency decreases linearly as $a_{i,t}$ increases for $a_{i,t} \in [0,1]$, while discharging efficiency increases linearly as $a_{i,t}$ decreases for $a_{i,t} \in [-1,0]$. The efficiency function is defined as:

$$\eta(a_{i,t}) = \begin{cases} \alpha_0 + \alpha_d \cdot a_{i,t}, & \text{if } a_{i,t} < 0, \\ \alpha_0 - \alpha_c \cdot a_{i,t}, & \text{if } a_{i,t} \geq 0, \end{cases} \tag{8}$$

where $\alpha_0 \in (0,1)$ represents the baseline charging/discharging efficiency. The coefficients $\alpha_c > 0$ and $\alpha_d > 0$ represent the rates at which efficiency decreases with increasing charging and discharging percentages, respectively. To ensure that efficiencies across all $a_{i,t} \in [-1,1]$ are non-negative, we impose the conditions that $\alpha_0 - \alpha_c > 0$ and $\alpha_0 - \alpha_d > 0$.

Energy bid. With solar panels and energy storage, a prosumer's bid, $b^{\theta}_{i,t}$, can be represented as a function of the state variable $s_{i,t}$, action $a_{i,t}$, and the exogenous uncertainty $q^{\theta}_{i,t}$:

$$b^{\theta}_{i,t}(s_{i,t}, a_{i,t}, q^{\theta}_{i,t}) = \begin{cases} q^{\theta}_{i,t} \cdot \overline{e}^{\theta} + \eta(a_{i,t})\overline{e}^{\theta} \cdot \max\left\{-e_{i,t},\ a_{i,t}\right\}, \\ \qquad\qquad \text{if } a_{i,t} < 0 \text{ (discharging)}, \\ q^{\theta}_{i,t} \cdot \overline{e}^{\theta} + \dfrac{\overline{e}^{\theta} \cdot \min\left\{1 - e_{i,t},\ a_{i,t}\right\}}{\eta(a_{i,t})}, \\ \qquad\qquad \text{if } a_{i,t} \geq 0 \text{ (charging)}. \end{cases} \tag{9}$$

The bids represent the sum of the net load and battery charging or discharging quantity (after accounting for efficiency losses). Since the state variables, action variables, and exogenous uncertainties – $e_{i,t}$, $a_{i,t}$, and $q^{\theta}_{i,t}$ – are all bounded, the bid $b_{i,t}$ is also bounded for all i and t.

The formulation in (9) defines the bidding strategy. The first case ($a_{i,t} < 0$, discharging) indicates that the agent first uses its energy storage to meet its net energy demand, $q^{\theta}_{i,t} \cdot \overline{e}^{\theta}$, measured in absolute terms rather than as a percentage. If there is excess energy after discharging, it is sold directly into the wholesale market. Conversely, if there is a shortfall, the agent purchases the required energy from the wholesale market. The second case ($a_{i,t} \geq 0$, charging) states that the bid represents the total energy purchased from the grid to meet the agent's energy demand plus the energy charged to storage. This bid formulation makes the state transition in (7) deterministic and simplifies the analysis considerably. While this is not the only way to design a bidding strategy, it has the advantage of giving prosumers precise control over the energy levels they wish to maintain in their storage.

The downside of this approach is that it assumes any excess supply or demand from prosumers can always be absorbed by or met in the wholesale market. This assumption holds when the collective size of prosumers is small relative to the overall grid's supply and demand, or when considering the geographical averaging effect – where excess energy from prosumers in one area can be used to meet the needs of another. However, as the number of prosumers increases, this assumption may become problematic, especially since most prosumers have solar generation, not wind. Unlike wind energy, which benefits from geographical diversity due to varying wind conditions, solar

power generally does not. During the day, solar energy is generated across all locations (with some variation due to irradiance), but in the evening, production drops to zero. A more sophisticated bidding strategy using reinforcement learning can be explored as a future research direction.

Population Profile. Before detailing the payoff functions for each agent, it is essential to establish the concept of a population profile, which aggregates the states and actions of all agents. In a large population game, although the actions of an individual agent do not directly affect the payoffs of others, the aggregated actions of the entire group do. Population profiles vary by both time of day and agent type. We begin by defining the empirical distribution of population profiles for a finite number of agents at time t. To account for general state and action spaces, we use $\mathcal{B}(X)$ to denote the Borel σ-algebra of a generic set X. Then, for a state space $S \in \mathcal{B}(\mathcal{S})$, defined as the Cartesian product of $\mathcal{E}$ and $\mathcal{H}$ – the SoC space and the set of all times of the day – and an action space $A \in \mathcal{B}(\mathcal{A})$, we define the following:

$$\begin{aligned} & p_t^{I^\theta}(S, A) \\ &= \frac{1}{I^\theta} \sum_{i=1}^{I^\theta} \sum_{h \in \mathcal{H}} \mathbb{I}_{\{e_{i,t} \in \mathcal{E}\}} \times \mathbb{I}_{\{T_{day}(t)=h\}} \times \mathbb{I}_{\{a_{i,t} \in A\}}, \end{aligned} \tag{10}$$

where $\mathbb{I}_{\{e_{i,t} \in S\}}$, $\mathbb{I}_{\{T_{day}(t)=h\}}$, and $\mathbb{I}_{\{a_{i,t} \in A\}}$ are indicator functions that respectively track the state of charge, time of day, and action of agent i. This formulation represents the empirical joint distribution of states and actions across the population.

Let $p_h^{\infty,\theta}$ be the limit as $I^\theta, t \to \infty$ for all θ and $h \in \mathcal{H}$. This limit represents a probability distribution over the joint state and action space, denoted by $\Xi := \mathcal{S} \times \mathcal{A}$. We use $\mathcal{P}(\Xi)$ to denote the set of all probability measures on Ξ, and let $p_h^\infty := [p_h^{\infty,\theta}]_{\theta \in \Theta} \in \mathcal{P}(\Xi)^{|\Theta|}$ denote the population profile of all types at time h of a day, with $|\Theta|$ being the cardinality of the type space Θ. Furthermore, we use p^∞ to denote the collection of p_h^∞ for all $h \in \{1, \ldots, H\}$; that is, $p^\infty := [p_h^\infty]_{h=1}^H \in \mathcal{P}(\Xi)^{|\Theta| \times H}$.

Payoff. The single-stage payoff function for a type-θ agent at time t, with a long-run equilibrium of the population profile $p^\infty_{T_{day}(t)}$, is denoted as $R^\theta_{i,t}(s_{i,t}, a_{i,t}, q^\theta_{i,t} \mid p^\infty_{T_{day}(t)}) : \mathcal{S} \times \mathcal{A} \times \mathcal{Q} \to \mathbb{R}$. To provide the explicit mathematical formulation of the payoff function, we first define $P_t^{n(\theta)}(\cdot) : \mathcal{P}(\Xi)^{|\Theta|} \to \mathcal{R}$ as a function that maps the population profile at time t to the LMP at node n in the transmission

network. With a slight abuse of notation, we use $n(\theta)$ to denote the location within the transmission network where agents of type θ are situated. The stage payoff function is:

$$\begin{aligned} &R^{\theta}_{i,t}(s_{i,t}, a_{i,t}, q^{\theta}_{i,t} | p^{\infty}_{T_{day}(t)}) \\ &= -P^{n(\theta)}_{t}(p^{\infty}_{T_{day}(t)}) \times b^{\theta}_{i,t}(e_{i,t}, a_{i,t}, q^{\theta}_{i,t}), \end{aligned} \tag{11}$$

where $b^{\theta}_{i,t}$ is agent i's energy bid at time t, as specified in (9). Since $b_{i,t} < 0$ indicates energy sales to the grid, this formula yields a positive payoff for the agent, while an energy purchase bid ($b_{i,t} > 0$) results in a cost, or a negative payoff, to the agent.

Note that the stage payoff is a random variable due to the stochastic nature of the LMPs, demand, and variable renewable outputs. When determining optimal policies, agents must rely on the expected value of the payoff. Therefore, to simplify the notation and analysis, we directly define the expected payoff and denote it by $\bar{R}^{\theta}_{t}(s, a \mid p^{\infty})$. Since each individual agent's bid is small (infinitesimal in the case of an infinite number of agents), we assume that the individual bid does not impact the LMPs and is thus independent of them. Consequently, we can write out the expected value of the payoff as follows:

$$\begin{aligned} &\bar{R}^{\theta}_{t}(s, a \mid p^{\infty}) \\ &:= \mathbb{E}\left[R^{\theta}(s, a, q^{\theta} \mid p^{\infty})\right] \\ &= -\bar{P}^{n(\theta)}_{t}(p^{\infty}_{T_{day}(t)}) \times \mathbb{E}_{q^{\theta}}\left[b^{\theta}_{i,t}(e_{i,t}, a_{i,t}, q^{\theta}_{i,t})\right], \end{aligned} \tag{12}$$

where $\bar{P}^{n(\theta)}_{t}(p^{\infty}_{T_{day}(t)})$ represents the expected LMP at node $n(\theta)$ and time t.

Remark 1. (Boundedness of $\bar{R}^{\theta}_{t}$.) Note that we assumed both the net load and energy storage capacity of each agent are bounded. Therefore, each agent's bid is bounded, regardless of external uncertainties. For net load, following a similar approach to how we define individual energy storage capacity, we assume that the total net load for each agent type θ is bounded by an upper limit $\bar{Q}^{\theta}$. As a result, the aggregate demand at each time t, $\mathbf{B}_t = (B^1_t, \ldots, B^N_t)$, lies within a compact region. By the Lipschitz continuity of the LMPs with respect to $\mathbf{B}_t$ (under the assumption that the LICQ holds at all the feasible points), the LMPs are uniformly bounded (over the feasible region

of $\mathbf{B}_t$). Hence, the payoff function $R^{\theta}_{i,t}(s_{i,t}, a_{i,t}, q^{\theta}_{i,t} \mid p^{\infty}_{T_{day}(t)})$, along with its expected value, is also uniformly bounded.

Remark 2. (Continuity of $\bar{R}^{\theta}_t$.) Based on the formulation of an agent's bid in (9), for a given $e \in S$, the function is continuous with respect to the action a. This is evident because the bid function consists of two parts: one for $a > 0$ and the other for $a < 0$. In both cases, the max and min functions are continuous, and so is the charging/discharging efficiency function (8). Hence, their product is continuous as well. At $a = 0$, whether approaching from $a \to 0^+$ or $a \to 0^-$, the bid function $b^{\theta}_{i,t}(e_{i,t}, a_{i,t}, q^{\theta}_{i,t})$ always converges to $q^{\theta}_{i,t} \cdot \bar{e}^{\theta}$. This reflects the fact that as the action approaches zero (i.e., no charging or discharging), the bid approaches the net load $q^{\theta}_{i,t} \cdot \bar{e}^{\theta}$. Therefore, for each $e \in S$, the expected payoff function $\bar{R}^{\theta}_t(s, a \mid p^{\infty})$ is continuous with respect to a.

4.3. Dynamic Optimization and Optimal Policy

The repeated decision-making problem of how to submit quantity bids and manage energy storage to maximize a prosumer's long-term payoff can be modeled as a stochastic dynamic programming (SDP) problem. Specifically, each agent i aims to maximize the following expected discounted payoff over an infinite time horizon:

$$\sup_{\pi_{i,t}} \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R^{\theta}_{i,t}(s_{i,t}, a_{i,t}, q^{\theta}_{i,t} \mid p^{\infty}_t) \,\middle|\, a_{i,t} \sim \pi_{i,t},\ s_{i,0},\ p^{\infty}_0\right], \tag{13}$$

where $\beta \in (0, 1)$ is the discount factor. The stochastic process begins with an initial individual state $s_{i,0}$ and population profile p^{∞}_0. At time $t = 0, 1, \ldots$, the agent selects an action $a_{i,t}$ according to a policy $\pi_{i,t}$.

Assume that the population profile is already at an equilibrium (its existence is the main subject of Section 5). Since the only actions are energy storage charging and discharging – which we model as percentages – the action space is compact, irrespective of the state. As discussed in Remarks 1 and 2, the expected stage reward function is bounded and continuous with respect to the actions. Additionally, the state transition function (7) is continuous with respect to the action a. Therefore, by the well-established result in Puterman (2014) (Theorem 6.2.12), an optimal stationary policy of (13) exists. Furthermore, according to a well-known result in stochastic dynamic programming Bertsekas et al. (2007) (Proposition 1.2.3), a stationary optimal policy must satisfy the Bellman equation.

To formulate the Bellman equation, we first define the value function for type-θ agents, which depends on both an agent's individual state and the population profile. For simplicity, we remove

the agent index i here, but still keep the type index θ. With a population profile $p^\infty = [p_h^\infty]_{h=1}^H$ and a stationary policy π^θ, the expected discounted present value for each state variable $s \in S$ can be expressed as follows:

$$V^{\pi^\theta}(s, p^\infty) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R_t^\theta(s_t, a_t, q_t^\theta | p^\infty) \middle| a_t \sim \pi^\theta,\ s_0\right]. \tag{14}$$

Let $V^{\pi^{\theta^*}}(s, p^\infty) = \max_{\pi^\theta \in \Pi^\theta} V^{\pi^\theta}(s, p^\infty)$, where Π^θ is the set of all admissible policies of the type-θ agent. The well-known Bellman equation can then be written as,

$$V^{\pi^{\theta^*}}(s, p^\infty) = \max_{a \in \mathcal{A}} \left\{\bar{R}^\theta(s, a | p^\infty) + \beta V^{\pi^{\theta^*}}\left[Tr(s, a),\ p^\infty\right]\right\}, \tag{15}$$

where the function $\bar{R}^\theta(s, a | p^\infty)$ represents the expected value of a one-period payoff, and $Tr(s, a)$ is the general state transition for both the storage's SoC and the time of day. Note that, as emphasized earlier, the state transition is deterministic. Therefore, the Bellman equation does not require a transition probability density function to describe how the state evolves. The corresponding optimal policy mapping is

$$\Pi^{\theta^*}(s, p^\infty) \equiv \arg\max_{a \in \mathcal{A}} \left\{\bar{R}^\theta(s, a | p^\infty) + \beta V^{\pi^{\theta^*}}\left[Tr(s, a),\ p^\infty\right]\right\}. \tag{16}$$

In the following, we show a key property regarding the agents' optimal stationary policy, which is crucial for proving the existence of an MFE in the next section.

PROPOSITION 2. *Under the assumptions of Proposition 1, for an agent of type θ, the optimal stationary policy mapping $\Pi^{\theta^*}(s,\ p^\infty)$ is single-valued and continuous with respect to $(s,\ p^\infty)$.*

The proof is in Online Appendix A.2.

5. Multiagent Mean-field Games

In this section, we first provide the precise definition of an MFE and show its existence in the context of DER integration into a wholesale electricity market. We then provide an algorithmic approach that enables agents to adaptively learn how to play in the mean-field game, which facilitates fully decentralized control automation.

5.1. Mean-Field Equilibrium: Definition and Existence

The essence of an MFE in this context is that each agent assumes the LMPs are at a long-run equilibrium and believes their individual actions do not influence this equilibrium. Based on this assumption, each agent selects an optimal strategy, which collectively leads to an equilibrium consistent with the assumed LMPs. This state is known as an MFE. A more precise definition of MFE is provided below.

DEFINITION 1. A collection of stationary strategy $\pi^* := [\pi^{\theta^*}]_{\theta\in\Theta}$ and a population profile $p^\infty := \left[[p_1^{\infty,\theta}]_{\theta\in\Theta}, \cdots, [p_H^{\infty,\theta}]_{\theta\in\Theta}\right] \in \mathcal{P}(\mathcal{S})^{|\Theta|\times H}$ constitute an MFE if for each $\theta \in \Theta$ and $h = 1, \ldots, H$, the following two conditions hold:

- Optimality: for a given state $s \in \mathcal{S}$, $\pi^{\theta^*} \in \Pi^{\theta^*}(s, p^\infty)$ as defined in (16).
- Consistency: for all $S \times A \in \mathcal{B}(\mathcal{S}) \times \mathcal{B}(\mathcal{A})$, where $\mathcal{B}(\cdot)$ is the Borel algebra of the corresponding set, and $s \in S$,

$$
\begin{aligned}
&p_h^{\infty,\theta}(S\times A) = \\
&\int_{S\times A} \mathbb{I}_{S\times A}\left\{E\left(e, \pi_{h-1}^{\theta^*}(e, p^\infty)\right),\right. \\
&\qquad \left.\pi_h^{\theta^*}\left(E\left(e, \pi_{h-1}^{\theta^*}(e, p^\infty)\right), p^\infty\right)\right\} dp_{h-1}^{\infty,\theta}(s,a),
\end{aligned} \tag{17}
$$

where $E(e,a)$ represents the state transition function for the energy storage's state of charge, as in (7). In (17), when $h = 1$, it is understood that the model interprets $h-1$ as H, which represents the final time period of the previous day. Additionally, with a slight abuse of notation, we use $\pi_h^{\theta^*}(e, p^\infty)$ to denote the policy at the state where the time of day is h; that is $\pi_h^{\theta^*}(e, p^\infty) := \pi^{\theta^*}\left(s = (e, T_{day} = h),\ p^\infty\right)$.

Definition 1 implies that under an MFE, the population profile at the same time of day on different days remains invariant when each agent adopts an optimal strategy according to (16). Equivalently,

(π^*, p^∞) is an MFE if and only if p^∞ is a fixed point of the MFE operator $\Phi : \mathcal{P}(\mathcal{S})^{H\times|\Theta|} \to \mathcal{P}(\mathcal{S})^{H\times|\Theta|}$ defined by:

$$\Phi(p^\infty)(S\times A)_{\theta\in\Theta} = \begin{bmatrix} [\Phi_1^\theta(p^\infty)(S\times A)]_{\theta\in\Theta} \\ \vdots \\ [\Phi_H^\theta(p^\infty)(S\times A)]_{\theta\in\Theta} \end{bmatrix}, \tag{18}$$

where

$$\begin{aligned} &\Phi_h^\theta(p^\infty)(S\times A) = \\ &\int_{S\times A} \mathbb{I}_{S\times A}\left\{E\left(e, \pi_{h-1}^{\theta^*}(e, p^\infty)\right), \right. \\ &\qquad \left. \pi_h^{\theta^*}\left(E\left(e, \pi_{h-1}^{\theta^*}(e, p^\infty)\right),\ p^\infty\right)\right\} dp_{h-1}^\infty(s,a), \\ &\text{for } h = 1\cdots H, \text{ and } \forall\theta\in\Theta. \end{aligned} \tag{19}$$

Therefore, to show the existence of an MFE, we will prove that there is a fixed point of Φ, to be presented in Proposition 3 below.

PROPOSITION 3 **(Existence of an MFE)**. *Under the assumptions in Proposition 1, an MFE, as defined in Definition 1, exists for the prosumer bidding game of direct participation in a wholesale electricity market.*

The proof uses the Schauder-Tychonoff Fixed Point Theorem; the details are provided in Online Appendix A.3.

5.2. Finite Agents and Approximate Markov-Nash Equilibrium

The existence of an MFE established in the previous subsection assumes an infinite number of agents. A natural question arises: What happens when the number of agents is large but finite? More specifically, if each agent in a finite system adopts the mean-field equilibrium policy, which was derived under the infinite-agent assumption, how does this affect the system's equilibrium properties?

To address this question, we first formally define the Markov-Nash equilibrium and its approximate counterpart, the ϵ-Markov-Nash equilibrium. For notational convenience, we omit the type index θ corresponding to an agent $i = 1, \ldots, I$. For a finite number of agents I, let M^i denote the set of all Markov policies for agent i, and define the Cartesian product $M^{\mathcal{I}} := \Pi_{i=1}^{I} M^i$. Let $\boldsymbol{\pi}^{\mathcal{I}} \in M^{\mathcal{I}}$ denote the collection of policies of the I agents, i.e., $\boldsymbol{\pi}^{\mathcal{I}} = (\pi_1, \ldots, \pi_I)$. The empirical population profile at time t, denoted by $p_t^{\mathcal{I}}$, is defined as in (10); that is, $p_t^{\mathcal{I}}(\cdot, \cdot) = [p_t^{\theta}]_{\theta \in \Theta}$. The initial state of each agent is given by $s_{i,0} = (e_{i,0}, T_{day}(0))$, where the initial state of charge $e_{i,0}$ follows a distribution in $\mathcal{P}[0, 1]$, and the initial time of the day $T_{day}(0)$ is arbitrary. The initial states of different agents are assumed to be independent. Let $\boldsymbol{\pi}_{-i}^{\mathcal{I}}$ denote the collection of policies for all agents except agent i. The discounted total reward for agent i in a finite-agent game is defined as:

$$J_i^{\mathcal{I}}(\pi_i^{\mathcal{I}}, \boldsymbol{\pi}_{-i}^{\mathcal{I}}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R_{i,t}^{\theta}(s_{i,t}, a_{i,t}, q_{i,t}^{\theta} \mid p_t^{\mathcal{I}}) \,\middle|\, a_{i,t} \sim \pi_i^{\mathcal{I}}, s_{i,0}, p_0^{\mathcal{I}}\right].$$

Definition 2. (Definition 2.2, Saldi et al. (2018)) A policy $\boldsymbol{\pi}^{\mathcal{I}^*} \in M^{\mathcal{I}}$ is a Markov-Nash equilibrium if

$$J_i^{\mathcal{I}}(\pi_i^{\mathcal{I}^*}, \boldsymbol{\pi}_{-i}^{\mathcal{I}^*}) = \sup_{\pi_i \in M_i} J_i^{\mathcal{I}}(\pi_i, \boldsymbol{\pi}_{-i}^{\mathcal{I}^*}),\ i = 1, \ldots, I. \tag{20}$$

It is an ϵ-Markov-Nash equilibrium if

$$J_i^{\mathcal{I}}(\pi_i^{\mathcal{I}^*}, \boldsymbol{\pi}_{-i}^{\mathcal{I}^*}) \geq \sup_{\pi_i \in M_i} J_i^{\mathcal{I}}(\pi_i, \boldsymbol{\pi}_{-i}^{\mathcal{I}^*}) - \epsilon,\ i = 1, \ldots, I. \tag{21}$$

It has been established in Saldi et al. (2018) that for any given $\epsilon > 0$, an ϵ-Markov-Nash equilibrium exists when the number of agents I is sufficiently large. However, this result relies on several technical conditions that may be challenging to verify in general settings. In our case, the specific modeling of the deterministic state transition for the SoC of energy storage significantly simplifies the verification of these conditions. In the following, we demonstrate that these assumptions hold in our framework, justifying the use of the MFE policy even in a finite-agent game.

Proposition 4 (**ϵ-Markov-Nash Equilibrium**). *Under Assumption 1, for any $\epsilon > 0$, there exists a positive integer $I(\epsilon)$ such that for all $I \geq I(\epsilon)$, the policy $\pi^{(I)} = (\pi_1, \pi_2, \ldots, \pi_I)$, where each π_i is defined as in* (16) *for $i = 1, \ldots, I$, constitutes an ϵ-Markov-Nash equilibrium for the game involving I prosumers participating in a wholesale energy market.*

Proof To prove the result, we need to verify that the required nine conditions, as outlined in Saldi et al. (2018), are satisfied under our model. The continuity of the one-stage reward function with respect to the state, action, and population profile, as established in the previous section, along with the compactness of the action space and the boundedness of the reward function (as stated in Remark 1), ensures that several key conditions are satisfied. Furthermore, the specific modeling of the state transition in (7), which is deterministic and independent of the population profile, automatically satisfies the remaining related to the transition dynamics. □

6. Learning in a Mean-Field Game: An Algorithmic Approach

While previous results establish the existence of an MFE, they do not provide a direct strategy for agents to follow in games with a large number of participants. Various RL-based methods have been proposed to approximate the fixed-point iteration needed to converge to an MFE (Guo et al. 2019, 2023). These approaches typically employ a double-loop structure: first, the population profile is fixed while each agent solves an MDP using an RL algorithm, such as Q-learning; then, the population profile is updated.

As highlighted by Xie et al. (2021), the double-loop approach presents two major challenges: (1) the population profile usually evolves simultaneously with agents' policy updates, and (2) for large state spaces, function approximations (such as neural networks) used to represent value and policy functions make solving each subproblem computationally demanding. To address these issues, Xie et al. (2021) proposed a single-loop, online algorithm. In their method, the mean-field state is updated via a single step of gradient descent, while agents' policies are updated by one step of policy optimization, informed by real-time feedback from the game. To ensure convergence, the algorithm employs a fictitious play mechanism, where agents probabilistically update their policies or do nothing. This mitigates instability by smoothing the learning process, allowing the mean-field state to evolve more gradually alongside policy updates.

While single-loop methods improve computational efficiency and stability compared to the double-loop approach, both methods still rely on a fundamental assumption: agents must have access to the global population profile, which is a probability distribution. However, this assumption is often impractical in real-world settings. In contrast, we propose an approach that leverages a specific feature of our setting – namely, fluctuations in electricity prices (aka the LMPs) reflect underlying changes in the population profile as well as external uncertainties. Instead of requiring direct access to the population profile, we allow agents to form beliefs about future LMPs at different

times of the day. These beliefs guide agents' actions by solving their SDP problems, eliminating the need for explicit knowledge of the population profile. Agents then update their beliefs adaptively based on realized prices, facilitating decentralized and scalable decision-making.

Specifically, let $\widetilde{P}_h$ denote the agent's belief about the LMP at time of day h, and $P_{T_{day}(t)=h}$ is the actual price observed at period t. The belief update rule for the h-th time of day is given by:

$$\widetilde{P}_h \leftarrow \widetilde{P}_h - \delta \cdot (\lfloor t/H \rfloor + 1)^{-0.5} (\widetilde{P}_h - P_{T_{day}(t)=h}). \tag{22}$$

The parameter δ is a learning rate in $(0, 1)$, and $\lfloor t/H \rfloor$ accounts for the total number of days elapsed. This rule adjusts an agent's belief using a diminishing step size, ensuring that recent observations have a greater impact while older data becomes less influential over time.

With an agent's belief and a given state $e_{t,h}$, an optimal decision a^* is determined by solving:

$$\begin{aligned} a^* = \arg \max_{a \in [-e_{t,h}, 1-e_{t,h}]} & -\eta(a) \cdot \min\{a, 0\} \cdot \widetilde{P}_h \\ & - \max\{a, 0\} \cdot \frac{\widetilde{P}_h}{\eta(a)} + \beta V_{h+1}(e_{t,h} + a, \widetilde{\mathbf{P}}), \end{aligned} \tag{23}$$

where $\eta(a)$ is the charging/discharging efficiency and β is the discount factor, as defined earlier, and $\widetilde{\mathbf{P}}$ denotes the vector of LMP beliefs for all times of the day, given by $(\widetilde{P}_h)_{h=1}^H$. The value function V_{h+1} – despite being defined for a single period $h + 1$ – depends on LMP beliefs across all time periods.

At this stage, the proposed method remains heuristic; yet numerical experiments consistently demonstrate convergence to a steady state. This suggests that underlying theoretical convergence properties may exist, which we leave as an avenue for future research.

To enhance our algorithm's realism and demonstrate the demand response capability of DERs without direct load control, we incorporate demand and supply shocks to reflect real-world conditions, such as unexpected fluctuations in energy demand or renewable generation. For scenarios involving these shocks, we assume the ISO issues an emergency signal one hour before the event. During these periods, agents adapt their actions based on alternative sets of LMP beliefs corresponding to the type of shock. The value function and optimal decision rules are computed separately for regular and shock periods. Specifically, during a demand shock – when demand surges and supply is likely insufficient, risking blackouts if no action is taken – agents replace their regular LMP

belief $\widetilde{P}$ with $\widetilde{P}^{\mathrm{DS}}$. In contrast, during a supply shock – where electric power supply likely exceeds demand, such as at night when wind energy surges but electricity demand is low – agents switch to the LMP beliefs $\widetilde{P}^{\mathrm{SS}}$. Agents then use these alternative beliefs to determine optimal actions under supply or demand shocks, following the same approach as in (23). To track the frequency of such events, agents maintain counters for demand and supply shocks, denoted as τ_d and τ_s, respectively. After observing the actual LMP P_t, agents update the LMP beliefs for supply or demand shocks using the same adaptive rule as in (22), with the counters τ_d and τ_s replacing $\lfloor t/H \rfloor$ for demand and supply shocks, respectively.

The pseudocode summarizing the single agent's value-iteration algorithm, including responses to supply and demand shocks, is presented in Algorithm 1. Using the LMP beliefs in (23), the problem becomes a typical SDP problem. We do not specify a particular algorithm for solving (23), nor is exact computation required. Therefore, approximate dynamic programming methods, such as those in Bertsekas et al. (2007), Powell (2011), are all applicable. In an extreme case, the algorithm can involve just one step of a gradient-descent-like method to enable a single-loop, online approach. This flexibility makes the framework scalable and adaptable, allowing agents to learn and act effectively in a mean-field game setting, where the mean-field is reflected through market prices. In our implementation, we introduce a small probability for each agent to restart in a random state with random LMP beliefs, a process referred to as regeneration. This serves two purposes. First, it ensures the multi-agent system remains active and adaptive, allowing agents to continue learning. When some agents regenerate, they must relearn, preventing the system from becoming static once it reaches a mean-field equilibrium. Second, it simulates real-world scenarios, where some agents may leave the market while new agents enter, reflecting the natural turnover in such systems.

Algorithm 1 Single Agent's Value-Iteration Algorithm

Initialization:

Randomly initialize $e_{0,1} \in [0,1]$ (initial battery state).

Randomly initialize $\widetilde{P}$ and shock-specific beliefs $\widetilde{P}^{\mathrm{DS}}, \widetilde{P}^{\mathrm{SS}}$.

Set learning rate $\delta \in (0,1)$; initialize shock counters $\tau_d = 0$ and $\tau_s = 0$.

for $t = 0, 1, \ldots$ **do**

  for each hour $h = 1, \ldots, H$ **do**

    if no emergency signal received **then**

      Submit bid a^* based on (23) using the LMP belief $\widetilde{P}$.

      Update beliefs using (22) after the market price P_t is observed.

    else if demand shock signal received **then**

      Submit bid a^* based on (23) using the LMP belief $\widetilde{P}^{\mathrm{DS}}$.

      Update beliefs after the market price P_t is observed as follows

$$\widetilde{P}_h^{DS} \leftarrow \widetilde{P}_h^{DS} - \delta \cdot (\tau_d + 1)^{-0.5} (\widetilde{P}_h^{DS} - P_t).$$

      Set $\tau_d \leftarrow \tau_d + 1$.

    else if supply shock signal received **then**

      Submit bid a^* based on (23) using the LMP belief $\widetilde{P}^{\mathrm{SS}}$.

      Update beliefs after the market price P_t is observed as follows

$$\widetilde{P}_h^{SS} \leftarrow \widetilde{P}_h^{SS} - \delta \cdot (\tau_s + 1)^{-0.5} (\widetilde{P}_h^{SS} - P_t).$$

      Set $\tau_s \leftarrow \tau_s + 1$.

    end if

  end for

end for

7. Numerical Results

In this section, we apply Algorithm 1 to a test power system comprising both bulk generators and thousands of prosumers and consumers. Our objectives are as follows: (1) to assess if the algorithm can achieve convergence numerically; (2) to observe whether, upon convergence, the algorithm encourages the desired behavior of charging during peak sunshine hours and discharging in the

evening when demand increases, ultimately smoothing LMPs and reducing volatility; and (3) to evaluate the algorithm's performance under random supply and demand shocks.

7.1. Test Case

In our experiment, we use the IEEE 14-bus system[4] as the test case. We assume there is one generator at each bus, with each generator's total generation cost represented by a quadratic function: $C_n(g) = \frac{1}{2}\alpha_n g^2 + \beta_n g$. The parameters α_n and β_n are chosen uniformly from the ranges $[0.0118, 0.0684]\$/\text{MW}^2\text{h}$ and $[150, 233]\$/\text{MWh}$, respectively, based on data from Krishnamurthy et al. (2015), for all n. Each power plant is assumed to have a 600 MW capacity. All transmission lines' capacity is set to be 1,000 MW. Each node (bus) contains two types of agents: prosumers with DERs (solar, small wind and energy storage) and pure consumers.

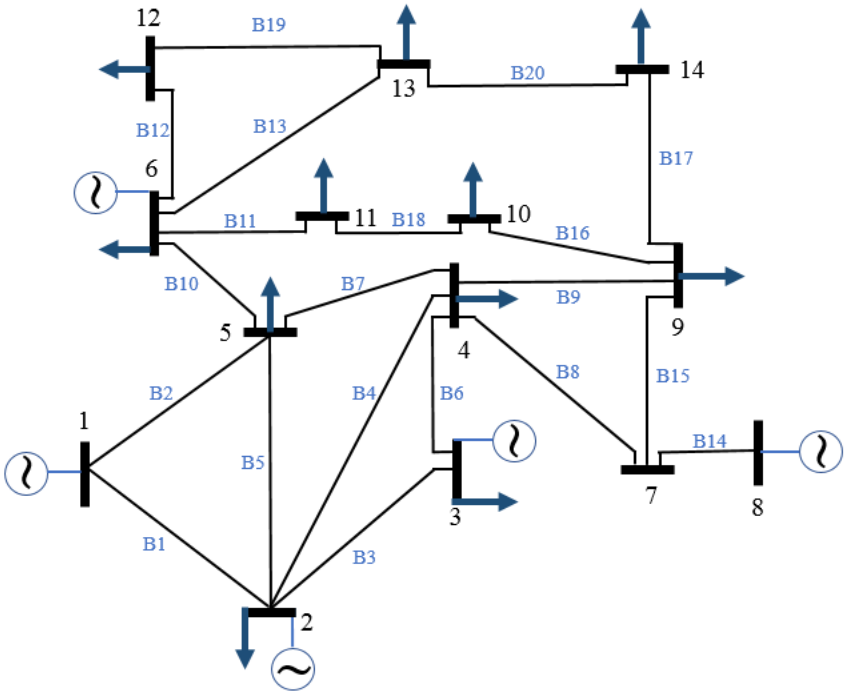


Figure 3 IEEE-14 test bus system

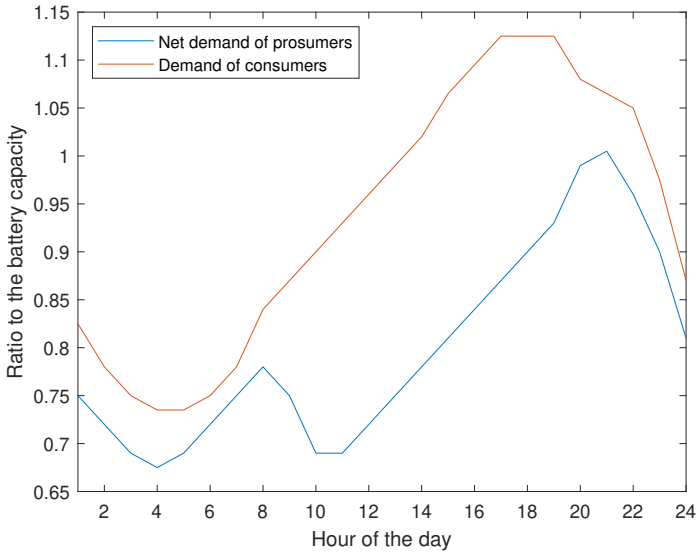
For demand, we use CAISO data,[5] which includes both total aggregated load and net load, where net load is defined as gross load minus distributed wind and solar generation. Using CAISO's

[4] Power Systems Test Case Archive – 14 Bus Power Flow Test Case (`https://labs.ece.uw.edu/pstca/pf14/pg_tca14bus.htm`).

[5] Both aggregate load and net load data are available at `https://www.caiso.com/todays-outlook#section-net-demand-trend`.

2022 data, we compute two aggregated load shapes – one for gross load and one for net load – each representing the average 24-hour profile, as shown in Fig. 4. In our model, prosumers and

Figure 4 Daily shapes for agents



consumers at the same bus are classified under the same location type, with all prosumers sharing a common net load shape and all consumers sharing a common gross load shape. To introduce variability across the 14 buses in our test network, we scale the CAISO load data by assigning each bus a unique load shape. Specifically, the base load profile is multiplied by a scaling factor, uniformly drawn from (0.9, 1.1), ensuring differentiation in demand patterns across buses.

To further introduce variability at the agent level, individual demand values in our simulation are sampled from a triangular distribution, with a lower bound of 0.8 times, an upper bound of 1.2 times the bus-specific average demand, and a mode equal to the average demand. This ensures that while all agents at a given bus follow the same general load pattern, their individual demand levels still vary, better capturing the diversity in consumption behaviors.

Each trading period in our simulation is set to one hour. To model demand surges, we introduce significant increases in energy demand between 6 PM and 9 PM on random days. This period aligns with the hours when solar output diminishes, providing an opportunity to evaluate how energy storage can be leveraged to mitigate early evening demand spikes within a completely decentralized decision-making framework. For supply shocks, we simulate increased generation, primarily driven by surges in distributed wind output, between 1 AM and 4 AM. These shocks occur on random

days and are independent of demand shocks. The arrival of both demand and supply shocks is assumed to follow independent Poisson distributions, each with an arrival rate of 0.1 events per day. During a demand shock, the surge is represented as a percentage increase relative to typical demand, modeled using a triangular distribution with bounds [30%, 50%] and a mode of 40%. Similarly, supply shocks involve increases in wind generation, modeled with a triangular distribution [20%, 30%] and a mode of 25%. Agents are notified one hour in advance if the system operator anticipates a shock in the upcoming period.

The simulation includes 3,000 agents at each node, with each agent having a probability of 0.0001 of regenerating in each hour. To represent battery levels, the state space is discretized into 100 evenly spaced points between 0 and 100%.

7.2. Result Analysis

We run simulations using Algorithm 1 to model a 100-day period, repeated 10 times with different random seeds. Additionally, we introduce two comparative scenarios: one where each agent maintains a single set of mean-field beliefs and does not adjust strategies in response to demand or supply shocks, and another without mean-field learning, where agents lack battery storage and bid solely based on their net load. This latter scenario represents a 'grid-tied' setup in which solar or small wind generation is directly connected to the grid; any excess generation is immediately fed back into the grid without storage. Figure 5 shows the average of relative difference between the belief of LMPs from an agent on Bus 3 and actual LMPs for Bus 3 across 10 runs. We select the LMPs from three typical hours – 4 AM, 9 AM, and 9 PM – when no supply or demand shocks occurred, as representative examples. The shaded region represents one standard deviation. It can be seen that the relative difference converges to almost zero quickly after about 10 days for all three hours, which indicates that agents' beliefs and their policies converge to a (mean-field) steady state quickly under our framework.

Figure 6 shows the realized LMPs for Bus 3, averaged over 10 runs, in chronological order for the first 10 days. The results indicate that LMPs stabilize quickly, reaching a steady state within less than 10 days, similar to the convergence pattern of LMP belief errors as in Figure 5. Compared to the LMPs in the scenario without energy storage (and therefore no agent learning), the LMPs with learning are lower during peak hours and higher during off-peak hours, resulting in reduced daily fluctuations.

Figure 5 Relative difference between the belief and true LMP: mean-field learning without prior shock information

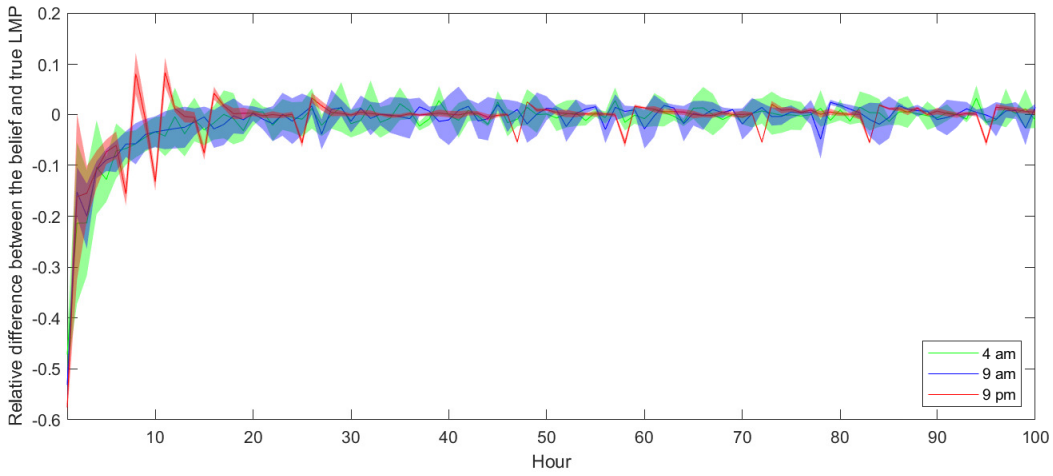
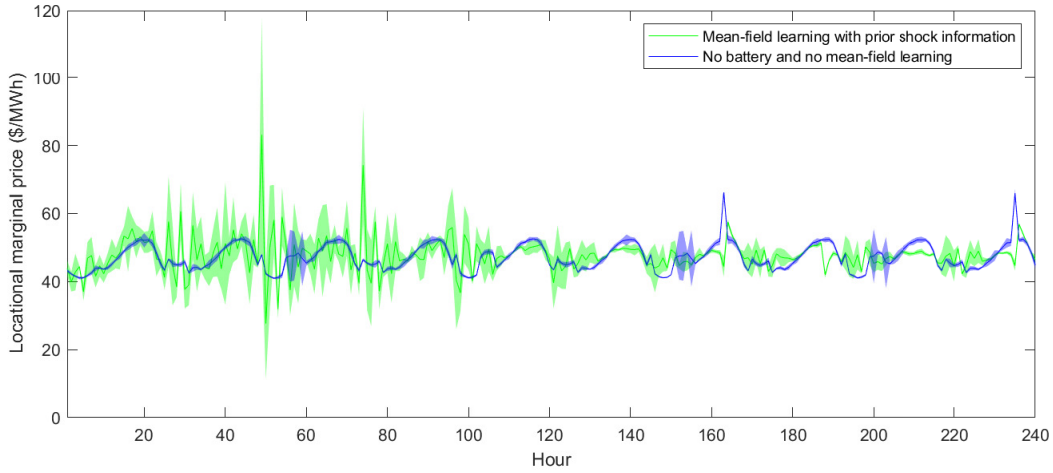
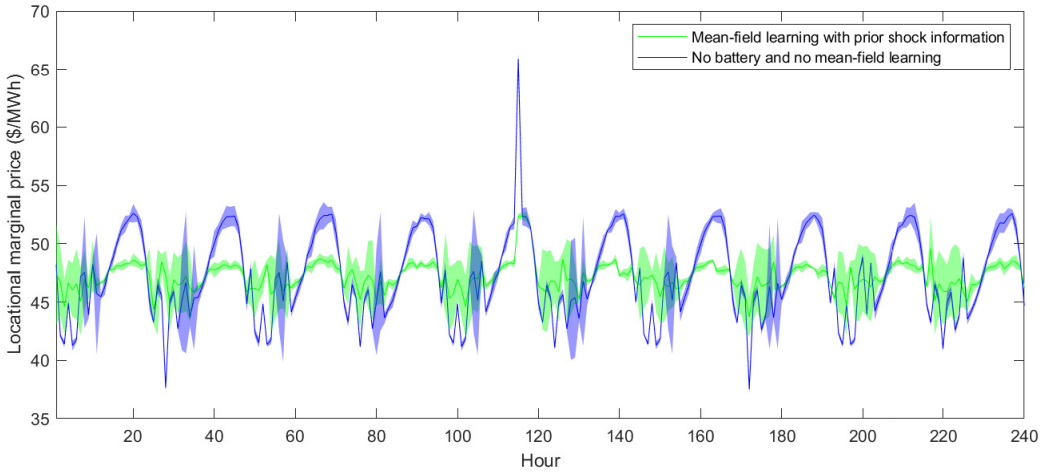


Figure 6 Hourly marginal prices of Bus 3 over the first 10 days: mean-field learning with prior shock information vs. no battery and no mean-field learning



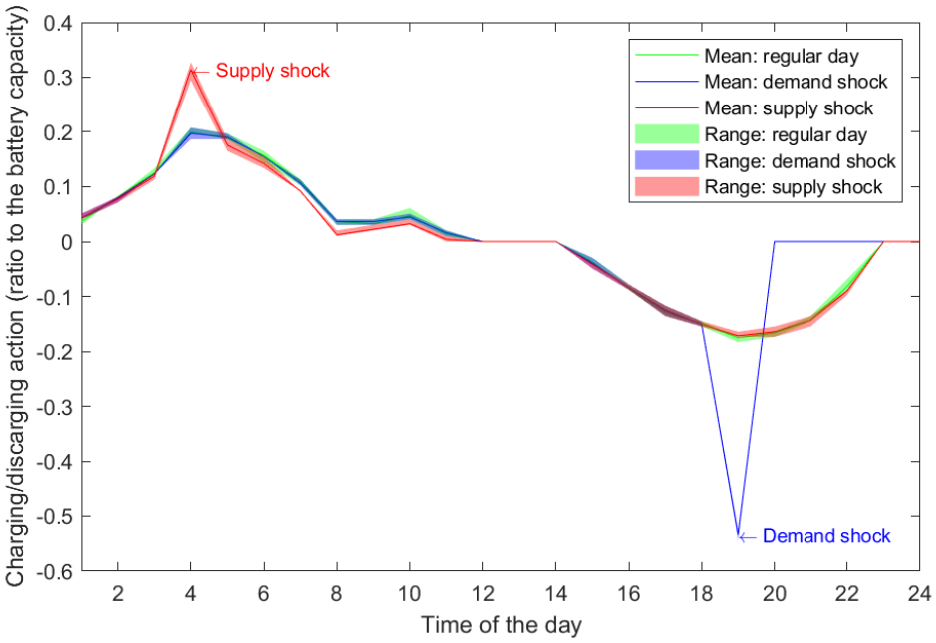
When there are supply or demand shocks, based on the results in Figure 7, where the LMP for the last 10 days is presented in chronological order, the LMPs without mean-field learning can fluctuate dramatically. In comparison, with mean-field learning, the LMPs during demand shocks show only a slight increase from regular levels, while during supply shocks, they remain nearly unchanged. As outlined in Section 6, agents with prior shock arrival information adjust their LMP beliefs for demand or supply shocks upon receiving the corresponding signals. Since the LMPs anticipated by agents during demand shock scenarios are significantly higher than those under regular conditions, prosumers discharge more energy from their batteries according to their

Figure 7 Hourly marginal prices of Bus 3 over the last 10 days: mean-field learning with prior shock information vs. no battery and no mean-field learning



optimal strategies. Similarly, during supply shocks, when anticipated LMPs are lower, prosumers may choose to charge more energy into storage to take advantage of the lower prices.

Figure 8 Charging/discharging actions over one day



To evaluate how strategy adaptation during supply or demand shocks helps mitigate these events, we compare the performance of mean-field learning frameworks with and without prior knowledge of shock arrivals. This comparison is presented in Figures 9 and 10, which display the average

LMPs for Bus 3 over ten independent runs during the first and last ten days, respectively. While both frameworks perform similarly during regular hours, substantial differences arise during demand and supply shocks. Notably, the framework not using prior shock information and without pre-shock strategy adjustments struggles to manage significant price fluctuations during these critical periods, underscoring the importance of incorporating built-in mechanisms to address diverse emergency scenarios within the algorithm.

Figure 9 Hourly marginal prices of Bus 3 over the first 10 days: mean-field learning with and without prior shock arrival information

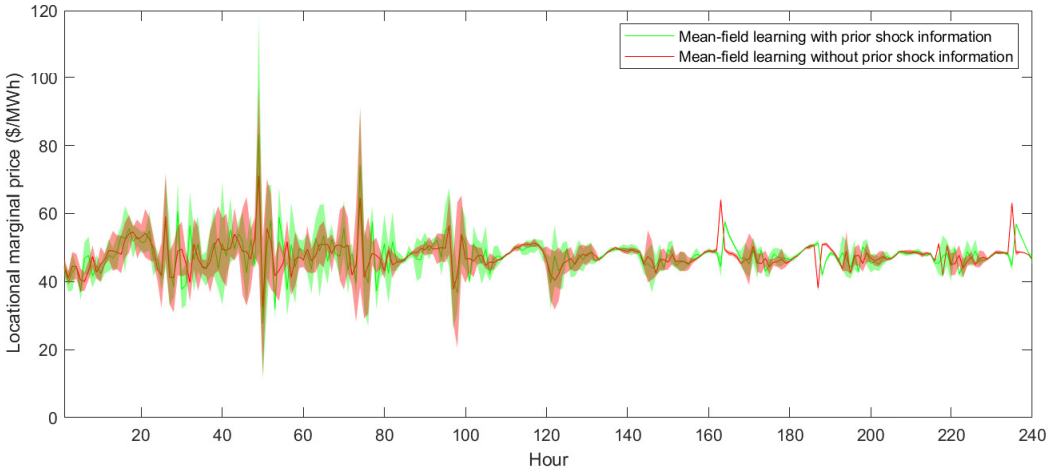
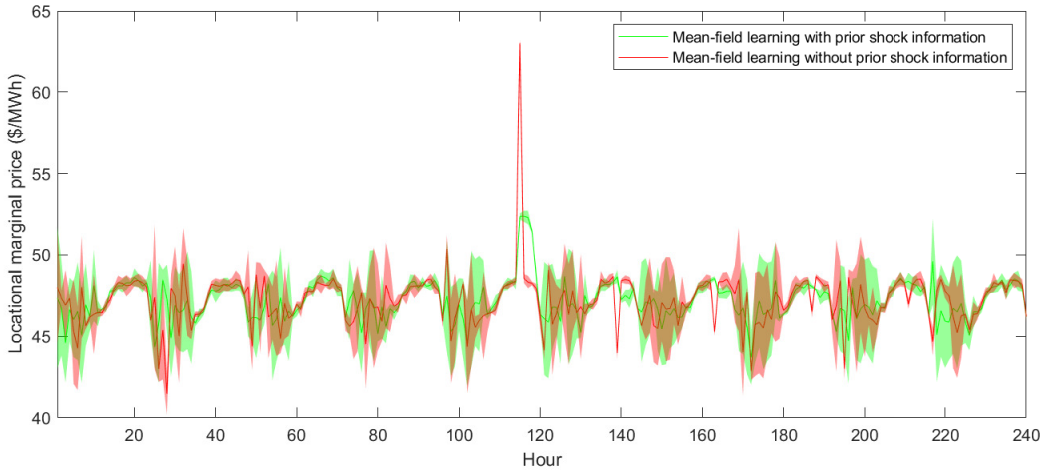


Figure 10 Hourly marginal prices of Bus 3 over the last 10 days: mean-field learning with and without prior shock arrival information



To further compare the volatility across different cases, we adopt the volatility measure presented in Roozbehani et al. (2012), which is the log-scaled incremental mean volatility (IMV). The IMV of a sequence $\{p_t\}_{t=1}^{\infty}$ is defined as

$$\text{IMV} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} |p_{t+1} - p_t|. \tag{24}$$

We approximate the IMV of a sequence of LMPs in our simulations using the prices from the last ten days, once the LMPs have reached a steady state. The average IMV over these ten days is computed as: $\overline{IMV} = \frac{1}{10}\sum_{i=1}^{10} IMV^i$, where IMV^i represents the IMV of the i-th run. Table 1 presents the average IMVs at Bus 3, along with its standard deviation, over ten runs across three different learning approaches. The results indicate that the scenario without mean-field learning

Table 1 Averaged IMV of the LMPs at Bus 3 over 10 runs under three different scenarios

Scenario	averaged IMV	Standard deviation
Mean-field learning with prior shock information	0.348	0.0038
Mean-field learning without prior shock information	0.370	0.0035
No mean-field learning	0.484	0.0018

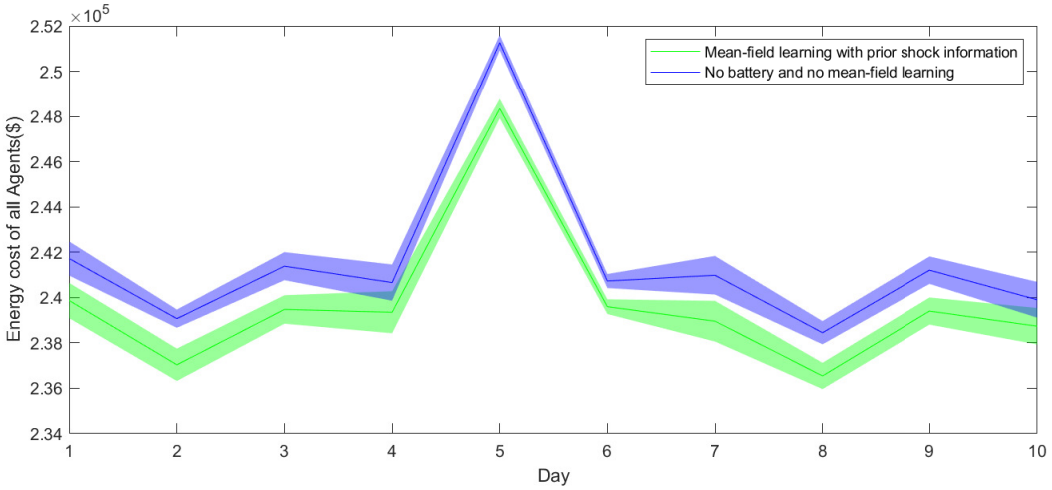
exhibits greater volatility compared to the other two scenarios. Unsurprisingly, the mean-field learning framework with prior shock information achieves the lowest volatility, owing to its capacity to mitigate price fluctuations during shock hours effectively.

Finally, we compare the daily energy costs of all agents over the last ten days across 10 independent simulation runs, focusing on the scenario with mean-field learning and prior shock information versus the scenario without mean-field learning, as shown in Figure 11. The results show a clear reduction in energy costs with mean-field learning.

8. Conclusion and Future Research

In this paper, we propose a mean-field game-based model and an algorithmic framework to enhance the participation of DER owners in wholesale energy markets. Our approach enables prosumers to make autonomous decisions based on real-time electricity prices while maintaining control over their assets. The mean-field approach is appropriate since all market information is reflected in

Figure 11 The total cost of all agents over the last 10 days: mean-field learning with prior shock information vs. no battery and no mean-field learning



the LMPs, which, along with signals from the system operator, are the primary data available to consumers and prosumers. We also proved the existence of a mean-field equilibrium for an infinite number of agents and the existence of an ϵ-Markov-Nash equilibrium for a finite but sufficiently large number of agents within this framework. Our numerical results indicate that, even with high renewable penetration or extreme weather conditions, the decentralized learning approach can help prevent extreme LMP fluctuations, contributing to a more stable energy market.

An immediate extension of this work is to investigate whether a system with a finite number of agents can converge to the mean-field equilibrium (MFE) as the number of agents approaches infinity. Additionally, developing a provably convergent algorithm to reach the MFE remains an important area for further research. Incorporating uncertainty in renewable generation and demand forecasts into prosumers' decision-making framework, and applying a reinforcement learning algorithm, could further enhance the robustness of the model under real-world conditions. Furthermore, if aggregators adopt a more active role, a promising direction is to apply mean-field control within each aggregator while modeling interactions among multiple aggregators as a mean-field game. Preliminary numerical results are provided in our related work (He and Liu 2024). We are currently working on establishing the theoretical foundations of this approach and will report our findings in a follow-up paper.

Appendix A: Proofs

A.1. Proof of Proposition 1 – Lipschitz continuity of LMPs

To prove the Lipschitz continuity of the LMPs with respect to the aggregate demand, we will need to resort to linear complementarity problems (LCPs) and a known result regarding the Lipschitz continuity of LCP solutions. An LCP with a given vector $u \in \mathbb{R}^n$ and a matrix $M \in \mathbb{R}^{n\times n}$, denoted by LCP$(u, M)$, seeks to find an $x \in \mathbb{R}^n$ such that $0 \le x \perp u + Mx \ge 0$, where the symbol $\perp$ denotes orthogonality; that is, $x^T(u + Mx) = 0$.

THEOREM 1. *(Theorem 3.2 in Mangasarian and Shiau (1987) – Lipschitz continuity of uniquely solvable LCPs). Let u^1 and u^2 be points in $\mathfrak{R}^n$ such that the* LCP$(u(\tau), M)$ *with $u(\tau) := (1-\tau)u^1 + \tau u^2$ has a unique solution for each $\tau \in [0, 1]$. Then the unique solutions x^1 of the LCP (u^1, M) and x^2 of (u^2, M) satisfy $||x^1 - x^2||_\infty \le \sigma_\beta(M)||u^1 - u^2||_\beta$, where $\sigma_\beta(M)$ is some constant derived from the matrix M.*

Proof of Proposition 1. Since the LMPs are determined by the dual variables of supply and demand balancing constraint and the transmission line constraints, as given in (5), under LICQ, the dual variables are unique, and hence, $P^n(\mathbf{B}_t)$ is single-valued with a given $\mathbf{B}_t \in \mathcal{F}_B$.

To utilize Theorem 1 to prove Lipschitz continuity of the LMPs with respect to energy demand, we write down the first-order optimality conditions (aka the KKT conditions) of the ED problem (1) – (4) at a given time t, with the quadratic cost function defined in 1:

$$
\begin{aligned}
0 \le g_t^n \perp \alpha^n g_t^n + \beta^n - \lambda + & \\
& \sum_{l=1}^{L} PTDF_l^n(\bar{\mu}_l - \underline{\mu}_l) + \bar{\eta}^n \ge 0 \\
0 \le \lambda \ \perp \sum_{n=1}^{N} g_t^n - \mathbf{1}^T \mathbf{B}_t \ge 0 & \\
0 \le \bar{\mu}_l \perp \widehat{F}_l - \sum_{n=1}^{N} PTDF_{l,n}(g_t^n - B_t^n) \ge 0 & \\
0 \le \underline{\mu}_l \perp \widehat{F}_l + \sum_{n=1}^{N} PTDF_{l,n}(g_t^n - B_t^n) \ge 0 & \\
0 \le \bar{\eta}_n \perp \widehat{G}_n - g_t^n \ge 0. &
\end{aligned}
$$

Since the objective function in (1) is assumed to be convex quadratic, and the constraints are all linear (and hence the linear constraint qualification holds everywhere), the KKT condition is a necessary and sufficient optimality condition. Let $\mathbf{g}_t, \bar{\boldsymbol{\mu}}, \underline{\boldsymbol{\mu}}, \bar{\boldsymbol{\eta}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \widehat{\boldsymbol{F}}$, and $\widehat{\boldsymbol{G}}$ represent vectors containing collections of their corresponding elements. Furthermore, let $PTDF \in \mathfrak{R}^{L\times N}$ be the matrix whose l-th row and n-th column element is $PTDF_l^n$. Furthermore, let $PTDF \in \mathfrak{R}^{L\times N}$ be the matrix whose l-th row and n-th column is $PTDF_l^n$, and $\Lambda = \text{Diag}(\boldsymbol{\alpha}) \in \mathfrak{R}^{N\times N}$ be a diagonal matrix with diagonal entries being the elements of the vector $\boldsymbol{\alpha}$. We can write the KKT conditions into the following LCP form:

$$0 \le \begin{pmatrix} \mathbf{g}_t \\ \lambda \\ \bar{\boldsymbol{\mu}} \\ \underline{\boldsymbol{\mu}} \\ \bar{\boldsymbol{\eta}} \end{pmatrix} \perp \begin{pmatrix} \boldsymbol{\beta} \\ -\mathbf{B}^t \\ \widehat{\boldsymbol{F}} + PTDF \times \mathbf{B}^t \\ \widehat{\boldsymbol{F}} - PTDF \times \mathbf{B}^t \\ \widehat{\boldsymbol{G}} \end{pmatrix} +$$

$$\begin{bmatrix} \Lambda & -\mathbf{1} & PTDF & -PTDF & I \\ \mathbf{1}^T & 0 & 0 & 0 & 0 \\ -PTDF & 0 & 0 & 0 & 0 \\ PTDF & 0 & 0 & 0 & 0 \\ -I & 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \mathbf{g}_t \\ \lambda \\ \bar{\boldsymbol{\mu}} \\ \underline{\boldsymbol{\mu}} \\ \bar{\boldsymbol{\eta}} \end{pmatrix} \ge 0,$$

where $\mathbf{1}$ denotes a vector of all 1's, I denotes the identity matrix, and 0 represents either a vector or a matrix, all of the appropriate dimensions. Let $\mathbf{x}$ denote the collection of all variables in the above LCP, $\mathbf{u}(\mathbf{B^t})$ represent the constant vector, and M be the big matrix. Then, the LCP above can be written in the following condensed form:

$$0 \le \mathbf{x} \perp \mathbf{u}(\mathbf{B^t}) + M\mathbf{x} \ge 0. \tag{25}$$

Under the assumptions of a strongly convex objective function and Assumption 1, for a given $\mathbf{B}^t$, the optimal primal and dual solutions are unique, and hence, the LCP (25) also has a unique solution. Consequently, Theorem 1 applies here, and since $\mathbf{u}(\mathbf{B^t})$ is a linear function with respect to B^t, it is straightforward to derive the LMPs, $P^n(\mathbf{B^t})$ as defined in (5), are Lipschitz continuous with respect to $\mathbf{B^t}$ for $n = 1, \ldots, N$. □

A.2. Proof of single-valuedness of a prosumer's optimal policy

To facilitate the derivation of theoretical results that follow, we need to endow $\mathcal{P}(\Xi)$ with the weak topology through the concept of weak convergence as follows.

DEFINITION 3. (Weak convergence (Aliprantis and Border 2006)) We say that a sequence of measures $\{p_n\} \in \mathcal{P}(\Xi)$ converges weakly to $p \in \mathcal{P}(\Xi)$ if, for all bounded and continuous functions $f : \Xi \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \int_{\Xi} f(x)\, p_n(dx) = \int_{\Xi} f(x)\, p(dx).$$

To prove Proposition 2, we need to first show that the expectation of the LMPs is continuous with respect to the population p_t^∞ at any location and at any time period t.

LEMMA 1. *Under Assumption 1 and the condition that at time t, the random noise of individual agent's demand $\zeta_{i,t}^\theta$, as defined in* (6)*, is i.i.d. the expected value of the LMP at each node $n = 1, \ldots, N$ at time t, as defined in Eq. (5), is a continuous function of the population profile p_t^∞ with respect to weak convergence, as defined in Definition 3.*

Proof As in Eq. (5), the LMPs at each node n are a function of the aggregated demand bids at all locations. Based on the definition of the bids in (9), when $I^\theta \to \infty$ for all $\theta \in \Theta$, since the total capacity of all type θ agents is assumed to be capped at $\overline{C}^\theta$, each individual agent's bid becomes infinitesimal, and the aggregate bids remain finite. We first characterize such aggregate bids using the Strong Law of Large Numbers (SLLN).

To sum over all the bids, for ease of notation, we use a function, $v_{i,t}^\theta(e_{i,t}, a_{i,t})$, to denote the second part of an agent's bid in (9):

$$v_{i,t}^\theta(e_{i,t}, a_{i,t}) := \begin{cases} \eta(a_{i,t}) \cdot \max\left\{-e_{i,t},\ a_{i,t}\right\},\ a_{i,t} < 0, \\ \dfrac{\min\left\{1 - e_{i,t},\ a_{i,t}\right\}}{\eta(a_{i,t})},\ a_{i,t} \geq 0. \end{cases} \tag{26}$$

Since both the state and action are random variables (due to the exogenous uncertainties in each agent's reward functions), whose joint distribution is exactly the population profile $p_t^{\infty,\theta}$ when

$I^\theta \to \infty$, $v^\theta_{i,t}(e_{i,t}, a_{i,t})$ is also a random variable. Since within the same type, all agents use the same optimal policy and are subject to the same weather conditions, we can assume that the series $\{v^\theta_{i,t}\}_{i=1}^{\infty}$ is i.i.d. By multiplying $\bar{e} = \overline{C}^\theta / I^\theta$ (to obtain the actual energy bids considering battery charging/discharging, as defined in the bid formulation (9) and by applying the SLLN, we have that

$$\begin{aligned}
&\lim_{I^\theta \to \infty} \sum_{i=1}^{I^\theta} \left[v^\theta_{i,t}(e_{i,t}, a_{i,t}) \times \bar{e} \right] \\
&= \lim_{I^\theta \to \infty} \left[\frac{\sum_{i=1}^{I^\theta} v^\theta_{i,t}(e_{i,t}, a_{i,t})}{I^\theta} \right] \overline{C}^\theta \\
&= \overline{C}^\theta \int_{\mathcal{E}\times\mathcal{A}} v^\theta_{i,t}(e_{i,t}, a_{i,t}) dp^{\infty,\theta}_t,
\end{aligned} \tag{27}$$

where the integration in the last equation represents the expected value of $v^\theta_{i,t}(e_{i,t}, a_{i,t})$.[6]

For the other part of an agent's bid, $q^\theta_{i,t}\bar{e}$, we have that

$$\begin{aligned}
&\lim_{I^\theta \to \infty} \sum_{i=1}^{I^\theta} q^\theta_{i,t} \bar{e} = \lim_{I^\theta \to \infty} \sum_{i=1}^{I^\theta} \left(\omega^\theta_t + \zeta^\theta_{i,t} \right) \frac{\overline{C}^\theta}{I^\theta} \\
&= \left(\omega^\theta_t + \lim_{I^\theta \to \infty} \frac{\sum_{i=1}^{I^\theta} \zeta^\theta_{i,t}}{I^\theta} \right) \overline{C}^\theta = \left(\omega^\theta_t + \bar{\zeta}^\theta_t \right) \overline{C}^\theta,
\end{aligned} \tag{28}$$

where the second equality holds because the random variable ω^θ_t represents weather-related uncertainties and does not depend on the agents (hence, no agent subindex i). The last equality directly follows from the SLLN.

[6] Note that in (28), when $I^\theta \to \infty$, it does not imply that the agents' actions (and states) correspond to a finite-agent game with I^θ agents. Instead, the agents' actions are derived from the optimal policy in the setting where the number of agents is already infinite. The limit in (28) simply represents the partial sum of an infinite series.

By (27) and (28), with a given population profile $p_t^{\infty,\theta}$, we can write out the aggregate bids of type θ as follows:

$$\begin{aligned} B_t^{\infty,\theta} &:= \sum_{i=1}^{\infty} b_{i,t}^{\theta}(e_{i,t}, a_{i,t}, q_{i,t}^{\theta}) \\ &= \lim_{I^\theta \to \infty} \sum_{i=1}^{I^\theta} \left[v_{i,t}^{\theta}(e_{i,t}, a_{i,t}) + \omega_t^{\theta} + \zeta_{i,t}^{\theta} \right] \times \bar{e} \\ &= \overline{C}^{\theta} \left(\int_{\mathcal{E}\times\mathcal{A}} v_{i,t}^{\theta}(e_{i,t}, a_{i,t}) dp_t^{\infty,\theta} + \omega_t^{\theta} + \bar{\zeta}_t^{\theta} \right). \end{aligned} \tag{29}$$

It can be seen that the aggregate bids for each type remain random variables due to the presence of the weather-related random variable ω^θ. Let ρ^ω denote the joint probability distribution of ω^θ for all $\theta \in \Theta$, and assume the joint distribution has a compact support Ω. Using the formulation in (5), the expected value of the LMP at time t at node $n = 1, \ldots, N$ can be written as:

$$\mathbb{E}[LMP_t^n] = \int_{\Omega} P^n \left(B_t^{\infty,1}, \cdots, B_t^{\infty,N} \right) \rho^\omega(d\omega). \tag{30}$$

Let $\{p_{t,k}^{\infty,\theta}\}_{k=1}^{\infty}$ be a sequence of population measures that weakly converge to $\{p_t^{\infty,\theta}\}$. By its definition in (26), the function $v_{i,t}^{\theta}(e_{i,t}, a_{i,t})$ is bounded and continuous. Hence, by Definition 3, we have

$$\begin{aligned} &\lim_{k\to\infty} \int_{\mathcal{E}\times\mathcal{A}} v_{i,t}^{\theta}(e_{i,t}, a_{i,t})\, dp_{t,k}^{\infty,\theta} \\ &= \int_{\mathcal{E}\times\mathcal{A}} v_{i,t}^{\theta}(e_{i,t}, a_{i,t})\, dp_t^{\infty,\theta}. \end{aligned}$$

As a result, by (29) and (30), $\mathbb{E}[LMP_t^n]$ is continuous with respect to the population profile $\{p_t^{\infty,\theta}\}$ since the LMP, P^n, is Lipschitz continuous with respect to the aggregated bids under Assumption 1. □

Proof of Proposition 2. The non-emptiness of the mapping in (16) follows from the existence of a stationary optimal policy, a well-established result in dynamic programming, as mentioned

earlier. Therefore, we omit the proof and proceed to show that the objective function in the Bellman equation (15) is strictly concave with respect to a. Together with the non-emptiness of the mapping, this implies the single-valuedness of the optimal policy.

To do so, we want to simplify the bid function (9) by removing the outer 'max' or 'min' operator first. By writing out the bid function explicitly and restricting the action a based on the current energy storage level $[-e, 1-e]$, we can equivalently re-write the Bellman equation as follows:

$$\begin{aligned}
&V^{\pi^{\theta^*}}(s, p^\infty) \\
&= \max_{a\in\mathcal{A}} \left\{ \bar{R}^\theta(s, a|p^\infty) + \beta V^{\pi^{\theta^*}}\left[Tr(s,a),\ p^\infty\right] \right\} \\
&= \max_{a\in[-e,\ 1-e]} \Bigg\{ \mathbb{E}_{q^\theta}\left[q^\theta \overline{e}^\theta\right] \\
&\quad - P^{n(\theta)}(p^\infty)\cdot\eta(a)\cdot\overline{e}^\theta\cdot\min(a,0) \\
&\quad - P^{n(\theta)}(p^\infty)\cdot\left(\overline{e}^\theta/\eta(a)\right)\cdot\max(a,0) \\
&\quad + \beta V^{\pi^{\theta^*}}\left[Tr(s,a), p^\infty\right] \Bigg\}.
\end{aligned} \tag{31}$$

We want to show that $\bar{R}^\theta(s, a|p^\infty)$ is strictly concave with respect to a. Since the first term in (31), $\mathbb{E}_{q^\theta}[q^\theta\overline{e}^\theta]$ is a constant, we only need to focus on the remaining terms. Additionally, for a given population profile p^∞, the LMP $P^{n(\theta)}(p^\infty)$ can also be treated as a constant for a given agent, which we simply denote it as P. Consider the following step-wise function:

$$\begin{aligned}
&u^\theta(a) \\
&:= \begin{cases} -\eta(a)\cdot a\cdot\bar{e}^\theta\cdot P & \text{if } a\in[-1,0], \\ -\dfrac{a}{\eta(a)}\cdot\bar{e}^\theta\cdot P & \text{if } a\in(0,1], \end{cases} \qquad (32) \\
&= \begin{cases} -(\alpha_0+\alpha_d\cdot a)\cdot a\cdot\bar{e}^\theta\cdot P & \text{if } a\in[-1,0], \\ -\dfrac{a}{\alpha_0-\alpha_c\cdot a}\cdot\bar{e}^\theta\cdot P & \text{if } a\in(0,1], \end{cases} \qquad (33)
\end{aligned}$$

where $\alpha_0\in(0,1)$, α_c, and $\alpha_d > 0$ are the parameters in defining battery charging/discharging efficiency in (8), with $\alpha_0-\alpha_c > 0$ and $\alpha_0-\alpha_d > 0$. It is straightforward to see that $u^\theta(a)$ is strictly

concave on either [-1, 0] or on (0, 1]. To show that $u(a)$ is strictly concave over the entire region [-1, 1], we construct two auxiliary functions $\overline{u}^{\theta}(a)$ and $\widetilde{u}^{\theta}(a)$ as follows:

$$\overline{u}^{\theta}(a) := \begin{cases} -(\alpha_0 + \alpha_d \cdot a) \cdot a \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in [-1, 0], \\ [(\frac{1}{2}\alpha_0 - \frac{1}{2\alpha_0}) \cdot a^2 - \alpha_0 a] \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in (0, 1], \end{cases} \tag{34}$$

and

$$\widetilde{u}^{\theta}(a) := \begin{cases} [(\frac{1}{2}\alpha_0 - \frac{1}{2\alpha_0}) \cdot a^2 - \frac{1}{\alpha_0} a] \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in [-1, 0], \\ -\dfrac{a}{\alpha_0 - \alpha_c \cdot a} \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad \text{if } a_h \in (0, 1]. \end{cases} \tag{35}$$

By taking the derivatives of the two functions, we get that

$$\frac{d\overline{u}^{\theta}(a)}{da} = \begin{cases} -(\alpha_0 + 2\alpha_d a) \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in [-1, 0], \\ [(\alpha_0 - \frac{1}{\alpha_0}) \cdot a - \alpha_0] \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in (0, 1], \end{cases} \tag{36}$$

and

$$\frac{d\widetilde{u}^{\theta}(a)}{da} = \begin{cases} [(\alpha_0 - \frac{1}{\alpha_0}) \cdot a - \frac{1}{\alpha_0}] \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in [-1, 0], \\ -\dfrac{\alpha_0}{(\alpha_0 - \alpha_c \cdot a)^2} \cdot \bar{e}^{\theta} \cdot \widetilde{P}, \\ \qquad\qquad \text{if } a \in (0, 1]. \end{cases} \tag{37}$$

Note that both functions are differentiable over the entire range of $[-1,1]$, as the left and right derivatives at $a=0$ are equal for both functions. For $\overline{u}^{\theta}(a)$, when $a \in [-1,0]$, clearly $d\overline{u}^{\theta}(a)/da$ is a strictly decreasing function since α_d, $\bar{e}^{\theta}$, and P are all positive. When $a \in (0,1]$, since $\alpha_0 \in (0,1)$, then $d\overline{u}^{\theta}(a)/da$ is also a strictly decreasing function. Hence, $d\overline{u}^{\theta}(a)/da$ is strictly decreasing over $[-1,1]$. By the well-known result for univariate functions (see Theorem 1.4 in Peajcariaac and Tong (1992)), $\overline{u}^{\theta}(a)$ is strictly concave on $[-1,1]$. Similarly, we can show that $\widetilde{u}^{\theta}(a)$ is also strictly concave on $[-1,1]$.

By the way of constructing $\overline{u}^{\theta}$ and $\widetilde{u}^{\theta}$, it is easy to see that $u^{\theta}(a) = \min\{\overline{u}^{\theta}(a), \tilde{u}^{\theta}(a)\}$. Hence, $u^{\theta}(a)$ is strictly concave on $[-1,1]$. Next, we show that the optimal value function $V^{\pi^{\theta^*}}(Tr(s,a), p^{\infty})$ is also strictly concave in a.

Let $\overline{\mathcal{J}}(\mathcal{E}\times\mathcal{H}\times\mathcal{P}(\Xi)^{|\Theta|})$ denote the space of all bounded functions on $\mathcal{E}\times\mathcal{H}\times\mathcal{P}(\Xi)^{|\Theta|}$, where $\mathcal{E}=[0,1]$ is the range of the energy storage state of charge, $\mathcal{H}$ is the discrete set of all times of day, and $\mathcal{P}(\Xi)^{|\Theta|}$ is the space of possible distributions of population profile p^{∞}. For a function $J^{\theta}(s,p^{\infty}) \in \overline{J}$ that is jointly continuous, define the Bellman operator $T : \overline{\mathcal{J}} \to \overline{\mathcal{J}}$ as follows:

$$\begin{aligned} &TJ^{\theta}(s,p^{\infty}) \\ &= \max_{a\in\mathcal{A}} \bar{R}^{\theta}(s,a|p^{\infty}) + \beta J^{\theta}\left(Tr(s,a), p^{\infty}\right). \end{aligned} \tag{38}$$

Although the state variable includes both the state of charge and the time of day, we can focus solely on the state of charge, as the time of day transition is discrete and deterministic, and it will not affect any of the discussion that follows. To simplify the transition function of the state of charge (7), we can let the feasible action space depend on the current state of charge, that is $a \in [-e, 1-e]$, then the state transition function (7) becomes $E(s,a) = e + a$. Let J^{θ} be any continuous function on $\overline{J}$ and concave with respect to s, then $J^{\theta}(E(s,a), p^{\infty})$ is also concave with respect a since $E(s,a)$ is a linear function in s and a. Now define the Bellman operator corresponding to the modified Bellman equation (31) as follows:

$$\begin{aligned} &TJ^{\theta}(s,p^{\infty}) \\ &= \max_{a\in[-e,\ 1-e]} \bar{R}^{\theta}(s,a|p^{\infty}) + \beta J^{\theta}\left(E(s,a), p^{\infty}\right). \end{aligned} \tag{39}$$

By reformulating the reward function as in (31) and expressing its explicit form in (33), the reward function $\bar{R}^\theta(s,a|p^\infty)$ does not explicitly depend on the state variable s. Since we have shown that it is concave in a, the term $\bar{R}^\theta(s,a|p^\infty)+\beta J^\theta\left(E(s,a),p^\infty\right)$ is jointly concave in (s,a). Additionally, the feasible region $a\in\mathcal{A}(e)\equiv[-e,1-e]$, considered as a point-to-set mapping, is hull concave over the percentage interval $\mathcal{E}=[0,1]$, meaning that the convex hull of $\mathcal{A}(e)$ is a concave mapping over $\mathcal{E}$. By a well-known result on the concavity of optimal value functions (see Proposition 3.2 in Fiacco and Kyparisis (1986)), TJ^θ is concave in s for a fixed p^∞. Consequently, the operator T preserves concavity, and T^kJ^θ remains concave in s for all $k=1,2,\ldots$. Furthermore, by the standard result from dynamic programming, the Bellman operator is a contraction mapping, ensuring that T^kJ^θ converges uniformly to $V^{\pi^{\theta^*}}$ (see Bertsekas and Shreve (1996)). Therefore, by a known result in convex analysis stating that the pointwise limit of a sequence of convex functions is also convex (Theorem 10.8 in Rockafellar (1997)), $V^{\pi^{\theta^*}}(s,p^\infty)$ is concave with respect to s, implying that $V^{\pi^{\theta^*}}[Tr(s,a),p^\infty]$ is concave with respect to a. Together with the strict concavity of the function $u^\theta(a)$ in (32) (and thus the strict concavity of $\bar{R}^\theta(s,a|p^\infty)$ in a), the 'argmax' mapping in (16) must be a singleton.

To show that the optimal policy mapping is continuous in (s,p^∞), we again rely on the Bellman operator in (38) with an arbitrary continuous function $J^\theta\in\overline{J}$. The one-stage reward function $\bar{R}^\theta(s,a\mid p^\infty)$ is the product of the LMP and bid quantity. By Proposition 1, the LMP is Lipschitz continuous with respect to p^∞. Since the bid function is jointly continuous in (s,a) (as can be seen in (9)), the reward function is jointly continuous in $[(s,p^\infty),a]$ in light of Lemma 1. Furthermore, the transition function $Tr(s,a)$, as defined in (7), is also jointly continuous in (s,a), making $J^\theta(Tr(s,a),p^\infty)$ jointly continuous as well, given that J^θ is a continuous function.

With the feasible action space $\mathcal{A}$ being compact, the Berge Maximum Theorem (Theorem 17.31 in Aliprantis and Border (2006) or Lemma 6.11.8 in Puterman (2014)) ensures that the optimal value function $TJ^\theta(s,p^\infty)$ is continuous in (s,p^∞). Since T^kJ^θ converges uniformly to $V^{\pi^{\theta^*}}$, the uniform limit theorem (Theorem 21.6 in Munkres (2014)) guarantees that $V^{\pi^{\theta^*}}(s,p^\infty)$ is jointly continuous. Finally, by the Berge Maximum Theorem again (or Lemma 6.11.9 in Puterman (2014)), the unique 'argmax' in (16) is continuous in (s,p^∞). □

A.3. Proof of MFE existence

As stated in the main text, proving the existence of an MFE in our context requires the Schauder-Tychonoff Fixed Point Theorem, stated below

PROPOSITION 5. *(Schauder-Tychonoff Fixed Point Theorem, Corollary 17.56, Aliprantis and Border (2006)) Let X be a nonempty, compact, convex subset of a locally convex Hausdorff space, and let $f : X \to X$ be a continuous function. Then the set of fixed points of f is compact and nonempty.*

Proof of Proposition 3. Given the uniform boundedness of the reward function (Remark 1) and the continuity result from Proposition 2, the existence proof follows directly from Theorem 3 in Light and Weintraub (2022), which applies the Schauder-Tychonoff Fixed Point Theorem. □

References

Adlakha S, Johari R (2013) Mean field equilibrium in dynamic games with strategic complementarities. Operations Research 61(4):971–989.

Aliprantis CD, Border KC (2006) Infinite dimensional analysis: A Hitchhiker's Guide (Springer).

Amoroso FA, Cappuccino G (2012) Advantages of efficiency-aware smart charging strategies for PEVs. Energy Conversion and Management 54(1):1–6.

Anderson EJ, Cau TD (2011) Implicit collusion and individual market power in electricity markets. European Journal of Operational Research 211(2):403–414.

Anderson EJ, Philpott AB (2002) Using supply functions for offering generation into an electricity market. Operations Research 50(3):477–489.

Anderson EJ, Xu H (2005) Supply function equilibrium in electricity spot markets with contracts and price caps. Journal of Optimization Theory and Applications 124(2):257–283.

Bagagiolo F, Bauso D (2014) Mean-field games and dynamic demand management in power grids. Dynamic Games and Applications 4:155–176.

Baldick R, Grant R, Kahn E (2004) Theory and application of linear supply function equilibrium in electricity markets. Journal of Regulatory Economics 25:143–167.

Bertsekas D, Shreve SE (1996) Stochastic optimal control: the discrete-time case, volume 5 (Athena Scientific).

Bertsekas DP, et al. (2007) Dynamic programming and optimal control: Volume 2. Belmont, MA: Athena Scientific .

Bunn DW, Oliveira FS (2001) Agent-based simulation-an application to the new electricity trading arrangements of England and Wales. IEEE transactions on Evolutionary Computation 5(5):493–503.

Du Y, Li F, Zandi H, Xue Y (2021) Approximating nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning. Journal of Modern Power Systems and Clean Energy 9(3):534–544.

Erev I, Roth AE (1998) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 848–881.

Fabra N, Toro J (2005) Price wars and collusion in the spanish electricity market. International Journal of Industrial Organization 23(3-4):155–181.

FERC (2020) Order No. 2222: Participation of Distributed Energy Resource Aggregations in Markets Operated by Regional Transmission Organizations and Independent System Operators. https://www.ferc.gov/ferc-order-no-2222-explainer-facilitating-participation-electricity-markets-distributed-energy, accessed: 2024-08-30.

Fiacco AV, Kyparisis J (1986) Convexity and concavity properties of the optimal value function in parametric nonlinear programming. Journal of Optimization Theory and Applications 48(1):95–126.

Fudenberg D, Tirole J (1991) Game Theory (The MIT Press).

Gu H, Guo X, Wei X, Xu R (2024) Mean-field multiagent reinforcement learning: A decentralized network approach. Mathematics of Operations Research .

Guerci E, Rastegar MA, Cincotti S (2010) Agent-based modeling and simulation of competitive wholesale electricity markets. Rebennack S, Pardalos PM, Pereira MVF, Iliadis NA, eds., Handbook of Power Systems II, 241–286 (Springer).

Guo X, Hu A, Xu R, Zhang J (2019) Learning mean-field games. Advances in Neural Information Processing Systems 32.

Guo X, Hu A, Xu R, Zhang J (2023) A general framework for learning mean-field games. Mathematics of Operations Research 48(2):656–686.

He J, Liu AL (2024) Evaluating the impact of multiple der aggregators on wholesale energy markets: A hybrid mean field approach. arXiv preprint arXiv:2409.00107.

Hobbs BF (1986) Network models of spatial oligopoly with an application to deregulation of electricity generation. Operations Research 34(3):395–409.

Holmberg P, Newbery D (2010) The supply function equilibrium and its policy implications for wholesale electricity auctions. Utilities Policy 18(4):209–226.

Krishnamurthy D, Li W, Tesfatsion L (2015) An 8-zone test system based on ISO New England data: Development and application. IEEE Transactions on Power Systems 31(1):234–246.

Light B, Weintraub GY (2022) Mean field equilibrium: uniqueness, existence, and comparative statics. Operations Research 70(1):585–605.

Liu AL (2010) Repeated games in electricity spot and forward markets-an equilibrium modeling and computational framework. 2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 66–71 (IEEE).

Liu AL, Hobbs BF (2013) Tacit collusion games in pool-based electricity markets under transmission constraints. Mathematical Programming 140:351–379.

Macal C, Thimmapuram P, Koritarov V, Conzelmann G, Veselka T, North M, Mahalik M, Botterud A, Cirillo R (2014) Agent-based modeling of electric power markets. Proceedings of the Winter Simulation Conference 2014, 276–287 (IEEE).

Mangasarian OL, Shiau TH (1987) Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. SIAM Journal on Control and Optimization 25(3):583–595.

Metzler C, Hobbs BF, Pang JS (2003) Nash-Cournot equilibria in power markets on a linearized DC network with arbitrage: Formulations and properties. Networks and Spatial Economics 3:123–150.

Mondal WU, Agarwal M, Aggarwal V, Ukkusuri SV (2022) On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC). Journal of Machine Learning Research 23(129):1–46.

Munkres J (2014) Topology (Pearson Education Limited), 2 edition.

Neuhoff K, Barquin J, Boots MG, Ehrenmann A, Hobbs BF, Rijkers FA, Vazquez M (2005) Network-constrained Cournot models of liberalized electricity markets: the devil is in the details. Energy Economics 27(3):495–525.

North M, Conzelmann G, Koritarov V, Macal C, Thimmapuram P, Veselka T (2002) E-laboratories : agent-based modeling of electricity markets. Proceedings of the 2002 American Power Conference.

Peajcariaac JE, Tong YL (1992) Convex functions, partial orderings, and statistical applications (Academic Press).

Powell WB (2011) Approximate Dynamic Programming: Solving the Curses of Dimensionality (John Wiley & Sons), 2 edition.

Price TC (1997) Using co-evolutionary programming to simulate strategic behaviour in markets. Journal of Evolutionary Economics 7:219–254.

Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming (John Wiley & Sons).

Ramchurn SD, Vytelingum P, Rogers A, Jennings N (2011) Agent-based control for decentralised demand side management in the smart grid. The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, 5–12 (International Foundation for Autonomous Agents and Multiagent Systems).

Ringler P, Keles D, Fichtner W (2016) Agent-based modelling and simulation of smart electricity grids and markets – a literature review. Renewable and Sustainable Energy Reviews 57:205–215.

Rockafellar RT (1997) Convex analysis, volume 28 (Princeton University Press).

Roozbehani M, Dahleh MA, Mitter SK (2012) Volatility of power grids under real-time pricing. IEEE Transactions on Power Systems 27(4):1926–1940.

Roth AE, Erev I (1995) Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. Games and economic behavior 8(1):164–212.

Rudkevich A (2005) On the supply function equilibrium and its applications in electricity markets. Decision Support Systems 40(3-4):409–425.

Saldi N, Basar T, Raginsky M (2018) Markov–nash equilibria in mean-field games with discounted cost. SIAM Journal on Control and Optimization 56(6):4256–4287.

Sensfuß F, Ragwitz M, Genoese M, Möst D (2007) Agent-based simulation of electricity markets: a literature review. Working Paper Sustainability and Innovation, No. S5/2007, Fraunhofer Institute for Systems and Innovation Research (ISI).

Shafie-khah M, Catalão JP (2014) A stochastic multi-layer agent-based model to study electricity market participants behavior. IEEE Transactions on Power Systems 30(2):867–881.

Sun J, Tesfatsion L (2007) Dynamic testing of wholesale power market designs: An open-source agent-based framework. Computational Economics 30:291–327.

Sutton RS, Barto AC (2018) Reinforcement learning: An introduction. The MIT Press .

Tajeddini MA, Kebriaei H (2018) A mean-field game method for decentralized charging coordination of a large population of plug-in electric vehicles. IEEE Systems Journal 13(1):854–863.

Visudhiphan P, Ilic MD (1999) Dynamic games-based modeling of electricity markets. IEEE Power Engineering Society Winter Meeting, volume 1, 274–281 (IEEE).

Willems B (2002) Modeling Cournot competition in an electricity market with transmission constraints. The Energy Journal 23(3):95–125.

Xie Q, Yang Z, Wang Z, Minca A (2021) Learning while playing in mean-field games: Convergence and optimality. International Conference on Machine Learning, 11436–11447 (PMLR).

Ye Y, Papadaskalopoulos D, Yuan Q, Tang Y, Strbac G (2022) Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets. IEEE Transactions on Smart Grid 14(2):1541–1554.

Zhao Z, Liu AL, Chen Y (2018) Electricity demand response under real-time pricing: A multi-armed bandit game. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 748–756 (IEEE).

Zhu Z, Lambotharan S, Chin WH, Fan Z (2016) A mean field game theoretic approach to electric vehicles charging. IEEE Access 4:3501–3510.