

UC Berkeley

Proceedings of the PowerUp Conference

Title

A Hybrid Mean Field Framework for Aggregators Participating in Wholesale Electricity Markets

Permalink

<https://escholarship.org/uc/item/5dk226n2>

Journal

Proceedings of the PowerUp Conference, 1(1)

Authors

He, Jun

Liu, Andrew L.

Publication Date

2026-09-10

DOI

None/powerup_proceedings.69144

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nd/4.0/>

Peer reviewed



PowerUp 2026

BOULDER, COLORADO
SEPTEMBER 9 – 11, 2026

A Hybrid Mean Field Framework for Aggregators Participating in Wholesale Electricity Markets

Jun He, Andrew L. Liu

Purdue University

Suggested Citation

J. He and A. L. Liu, “A Hybrid Mean Field Framework for Aggregators Participating in Wholesale Electricity Markets,” in *Proceedings of the PowerUp Conference*, 2026.

BibTeX Entry

```
@inproceedings{he2026hybridmean,  
  author = {He, Jun and Liu, Andrew L.},  
  title = {A Hybrid Mean Field Framework for Aggregators Participating in Wholesale  
    Electricity Markets},  
  booktitle = {Proceedings of the PowerUp Conference},  
  year = {2026},  
}
```

A Hybrid Mean Field Framework for Aggregators Participating in Wholesale Electricity Markets

Jun He and Andrew L. Liu

Abstract—The rapid growth of distributed energy resources (DERs) is reshaping wholesale electricity markets, where aggregators coordinate large populations of prosumers, end users who own DERs and can both consume and produce electricity, to participate at scale. Existing models typically either optimize a single aggregator under exogenous prices or analyze a small number of strategic agents; neither captures the scale, structure, and price feedback that arise when large DER populations interact through wholesale market clearing.

We address this gap with a hybrid mean field framework in which each aggregator coordinates a large prosumer population through its infinite agent limit, while a finite number of aggregators participate in wholesale markets. Each aggregator maximizes the collective payoff of its prosumers, a cooperative structure represented by mean field control. Across aggregators, interactions are non-cooperative; as a starting point, we model aggregators as non-strategic, price-responsive learners, while market prices remain endogenous through aggregate prosumer actions, leading to a mean field equilibrium. We establish conditions under which this equilibrium exists and is unique.

To handle uncertainty in demand, renewable generation, and market prices, we design a two-phase reinforcement learning algorithm that enables aggregators to learn optimal control policies from price feedback. Numerical experiments on the Oahu power system show that the proposed framework reduces price volatility, flattens demand profiles, improves storage utilization, and lowers costs for both consumers and prosumers.

Index Terms—Mean-field equilibrium; reinforcement learning; wholesale energy market; energy storage; solar PVs; aggregators.

I. INTRODUCTION

Distributed energy resources (DERs), such as rooftop solar and energy storage, are growing rapidly in adoption, enabling end-users to act as prosumers who both produce and consume electricity. Through bi-directional power flows, these prosumers are shifting the grid from a centralized to a more decentralized structure.

To accommodate this transition, FERC Order 2222 enables DER participation in wholesale markets. Because individual DERs are typically too small to meet market requirements, aggregators, also known as virtual power plants (VPPs), serve as intermediaries. Companies such as OhmConnect and Tesla already aggregate large numbers of households or devices into wholesale market resources [1], [2]. Integrating many heterogeneous, small-scale DERs at this scale raises new challenges in understanding their collective impact on market prices, volatility, and system reliability.

Most existing work studies how a single aggregator optimally manages a DER portfolio under exogenous wholesale

electricity prices, known as locational marginal prices (LMPs) [3]–[5]. While useful for operational insights, these models overlook that under large-scale DER participation, aggregators’ collective actions affect price formation. At scale, this creates a closed-loop feedback structure between aggregator actions and LMPs that must be captured to design algorithms in shaping market outcomes under high DER penetration.

To address this need, we develop a prescriptive, learning-based framework grounded in mean-field games (MFGs), where each aggregator optimizes its own portfolio decisions in response to dynamic market outcomes. The goal is not to mimic or study existing market behavior, but to engineer a scalable and decentralized system in which aggregators learn to optimize their DER portfolios under endogenously determined prices. Reinforcement learning (RL) enables each aggregator to adapt its strategy over time. We establish conditions for the existence and uniqueness of a mean-field equilibrium (MFE), which characterizes the steady-state outcome of this learning process. By explicitly modeling the feedback between aggregator actions and LMPs, the framework provides a foundation for AI-enabled, automated participation of DER aggregators in wholesale markets.

While MFGs capture interactions across aggregators, each aggregator must still manage a large population of DERs internally. To ensure scalability, we adopt a mean-field control (MFC) formulation in which the aggregator maximizes the collective payoff of its prosumers. Energy storage introduces intertemporal coupling, and variability in weather, load, and market prices creates significant uncertainty. Classical optimization methods rely on centralized coordination or accurate probabilistic models, whereas the MFC approach learns optimal policies directly from interaction with the environment. By representing prosumers through a representative agent in the infinite-population limit, the approach substantially reduces computational complexity while enabling adaptive, forward-looking control.

The contributions of this paper are twofold. First, we develop a hybrid MFC–MFE framework integrated with RL that enables decentralized DER aggregators to participate in wholesale electricity markets under uncertainty. Second, we design a two-phase RL algorithm for the hybrid MFC–MFE model and demonstrate its effectiveness on the Oahu power system, showing reduced price volatility, stabilized aggregator behavior, and lower customer costs.

The remainder of the paper is organized as follows. Section II reviews related literature. Section III presents the wholesale market model. Section IV formulates the aggregator’s problem under the hybrid MFC–MFE framework. Section V introduces the two-phase RL algorithm. Section VI reports numerical experiments, and Section VII concludes.

Jun He is with Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN, USA, email: he184@purdue.edu.

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

II. LITERATURE REVIEW

A growing body of work examines DER participation in wholesale electricity markets using multi-agent decision making and game-theoretic methods. Early studies generally treat each participant as an individual decision maker and rely on optimization- or learning-based formulations. For instance, [6] proposes a two-level optimization model for a single aggregator, and [7] applies deep RL for individual bidding. Related multi-agent approaches include [8], which uses linear programming for demand response, and [9], which develops a multilayer agent-based model. Institutional designs enabling aggregator participation are analyzed in [10]. Learning-based approaches, particularly multi-agent reinforcement learning (MARL), have also been used in energy systems, though mostly in peer-to-peer or small-scale contexts [11].

Although MARL is related to game-theoretic notions such as Nash equilibria, convergence guarantees are rare in stochastic games, and equilibrium computation becomes intractable as the number of agents increases [12]. These challenges motivate mean-field approximations, where interactions in a large population are modeled through a representative agent interacting with the population distribution. The resulting MFE framework greatly improves tractability while approximating behavior in large but finite systems.

Beyond tractability arguments, MFC has also been applied to large populations of flexible loads in power systems. For example, [13] develops a continuous-time MFC architecture for coordinating electric water heating loads to track desired aggregate targets. More closely related, [14] studies state estimation in MFC with an application to demand dispatch, where randomized local control enables aggregate load to provide grid services with limited information exchange. In these settings, coordination is guided by an externally specified aggregate target; in our setting, aggregators learn from LMPs formed endogenously through wholesale market clearing.

Recent work explores mean-field ideas under both known and unknown system dynamics. With known dynamics, [15] solves MFGs via dynamic programming. Under unknown dynamics, [16] and [17] develop RL-based MFG algorithms for competitive agents, although these typically require either full transition knowledge or direct access to the mean field. The sandbox method in [18] avoids a mean-field oracle by learning from a single sample path, but it operates in generic single-population environments without market-clearing feedback. Meanwhile, [19] applies MFC to approximate cooperative MARL settings, but does not model endogenous price formation or interactions across multiple aggregators.

In contrast, our work develops a mean-field framework tailored to DER aggregators participating in a wholesale electricity market. Each aggregator solves a cooperative MFC problem over a large DER population, aggregators interact indirectly through market clearing, and learning proceeds without access to the true mean field or uncertainty distributions, relying only on observed LMPs. This structure captures the closed-loop price effects absent in prior MFG/MFC RL approaches while preserving scalability.

III. WHOLESALE MARKET MODEL AND LOCATIONAL MARGINAL PRICES

This section presents a standard wholesale electricity market model along with the formulation of LMPs. We consider a wholesale electricity market operating over a transmission network with N buses, L transmission lines, and G bulk generators. Each bus $n \in \mathcal{N} := \{1, \dots, N\}$ serves a group of M_n agents, comprising M_n^p prosumers and M_n^c pure consumers without any DERs. Each bulk generator $g \in \mathcal{G} := \{1, \dots, G\}$ is characterized by a cost function $C_g(\cdot)$ that captures the cost of electricity production. Let $G_n \subseteq \mathcal{G}$ denote the set of generators connected to bus n . We require the sets to be disjoint and collectively exhaustive; that is, $\cup_{n=1}^N G_n = \mathcal{G}$ and $G_n \cap G_{n'} = \emptyset$ for any $n \neq n'$.

On the demand side, let d_{it}^n denote the net demand of prosumer i at bus n , defined as total energy consumption minus local solar generation. If $d_{it}^n > 0$, the prosumer is a net consumer; if $d_{it}^n < 0$, the prosumer exports energy to the grid. This bidirectional interaction allows prosumers to participate flexibly in the energy market. Similarly, we use $d_{jt}^n \geq 0$ to denote the demand of consumer j at bus n . The total demand at bus n and time t is:

$$D_t^n = \sum_{i=1}^{M_n^p} d_{it}^n + \sum_{j=1}^{M_n^c} d_{jt}^n. \quad (1)$$

To avoid technical complications from infeasible supply-demand imbalances, we assume that total net demand is non-negative at each timestep t : $\sum_{n=1}^N D_t^n \geq 0$.

The wholesale market operates over discrete time intervals, indexed by $t \in \{1, 2, \dots\}$, where each timestep corresponds to a fixed duration, such as an hour. During each timestep, the independent system operator (ISO) of a wholesale electricity market collects aggregate supply and demand bids and solves an optimization problem to determine the least-cost dispatch that satisfies system constraints. In this paper, we let the ISO solve a simplified version of the economic dispatch (ED) problem (see detailed formulation in Appendix A). The marginal cost of supplying one additional unit of electricity at each node, exactly the LMPs, is derived from the dual variables associated with the power balance and transmission constraints. We let λ_t^n denote the LMP at bus $n \in \{1, \dots, N\}$ and time t . For notation simplicity, let $\mathbf{D}_t := (D_t^1, \dots, D_t^N)$ denote the vector of all demand bids; then the LMP λ_t^n is a function of $\mathbf{D}_t$, which we write as $\lambda_t^n(\mathbf{D}_t)$.

IV. AGGREGATORS' PROBLEM AND THE MEAN-FIELD FRAMEWORK

In this section, we first formulate the decision-making problem for each aggregator and then introduce the notion of an MFE. Acting on behalf of a large population of prosumers equipped with solar PVs and energy storage, each aggregator must determine control strategies for the prosumers under multiple sources of uncertainty, including solar energy output, real-time energy demand, and market prices. A key challenge lies in the presence of energy storage, which links decisions across time and leads to a high-dimensional stochastic dynamic optimization problem. To address this complexity in a scalable manner,

we adopt an MFC formulation. In the infinite-population limit, the behavior of many similar prosumers converges to that of a representative agent, allowing the aggregator's collective control problem to be expressed through this limiting agent. This approximation is standard in MFC and enables decentralized learning without tracking each prosumer. It does not force identical behavior: since a policy maps a prosumer's state to an action and prosumers occupy different states, a single population-level policy still induces heterogeneous actions; the learned policies are moreover stochastic, so even prosumers in the same state need not act alike. The detailed descriptions of states and policies will be introduced later in this section.

A. Components of An Aggregator's Problem

We begin by introducing the key components of the game. Throughout, we use $\mathcal{P}(\mathcal{X})$ for the set of Borel probability measures on a space $\mathcal{X}$.

Time: Demand and solar generation vary strongly across hours yet follow recurring daily patterns. To capture both intra-day variation and inter-day similarity, we define two mappings over the global timestep $t = 0, 1, 2, \dots$: with H timesteps per day, $T_{\text{hour}}(t) = t \bmod H$ and $T_{\text{day}}(t) = \lfloor t/H \rfloor$ return the hour of day and the day index. This encodes hourly dynamics while treating each day as structurally similar.

Household agents (prosumers and consumers): At each bus n , there are M_n^p prosumers, each equipped with rooftop PV and energy storage. The total aggregated storage capacity at bus n is capped at $\bar{E}^n$. To capture heterogeneity, we partition the M_n^p prosumers into K types indexed by $k \in \{1, \dots, K\}$, reflecting differences in storage capacity and associated PV size. Let $b_k^n \in [0, 1]$ denote the fraction of type- k prosumers at bus n , with $\sum_{k=1}^K b_k^n = 1$. Each type k is assigned a relative (unnormalized) capacity parameter $\theta_k^n > 0$. For example, $\theta_2^n = 2\theta_1^n$ implies that type-2 prosumers have twice the storage capacity of type-1 prosumers. The storage capacity of a type- k prosumer i at bus n is then defined as $\bar{E}_{ik}^n := \theta_k^n \bar{E}^n / M_n^p \sum_{\kappa=1}^K \theta_{\kappa}^n b_{\kappa}^n$. This scaling ensures that the total storage capacity at each bus remains finite in the mean-field limit, where the number of agents grows large and the impact of any individual agent becomes infinitesimal.

We express each prosumer's net load over a time interval in normalized form. Let Δt denote the fixed duration of one market interval (such as an hour). For notational simplicity, we omit the type- k subscript unless needed. At time t , let q_{it}^n denote the ratio between prosumer i 's net energy consumption over interval t , excluding storage charging or discharging, and its storage energy capacity $\bar{E}_i^n$. Thus, q_{it}^n is a dimensionless ratio in $[-1, 1]$, where positive values indicate net consumption and negative values indicate net export from local PV. Equivalently, the corresponding average net power over the interval is $q_{it}^n \bar{E}_i^n / \Delta t$. The normalized net load is stochastic; we denote its unknown distribution by $\mathcal{Q}_n^p$. Neither prosumers nor aggregators are assumed to know $\mathcal{Q}_n^p$ a priori. Instead, decisions are learned through repeated interaction with the environment. At each bus n , there are also M_n^c pure consumers. To enable a unified treatment, we assign each consumer a notional reference storage capacity $\bar{E}_j^n$ and define

q_{jt}^n as the ratio of demand to this reference capacity. This quantity is also stochastic, with distribution denoted by $\mathcal{Q}_n^c$.

Aggregators: Although individual prosumers could in principle learn their own policies, this is impractical in practice, as most do not meet the 100 kW minimum threshold required by FERC Order No. 2222 for wholesale market participation. We therefore assume that, at each bus n , a single aggregator represents all prosumers at that location. While prosumers may differ in PV and storage capacities, they face the same LMPs and experience similar weather conditions at a given bus, which justifies modeling them with a common policy. We assume that PV generation is first used to meet local demand, with any surplus automatically charged into storage. The only decision variable is the storage action: how much energy to charge from or discharge to the market. Each aggregator learns a policy that maps system states (defined below) to charging or discharging actions. This policy is updated over time and broadcast to all prosumers at the bus, who use the policy to choose actions. Aggregators act as the learning agents, while prosumers are passive followers. As a starting point, aggregators in this work are modeled as non-strategic, price-responsive learners rather than as entities learning to exercise market power. Strategic behavior modeling and analysis of aggregators are deferred to future research.

Actions: As noted above, the only decisions in our setup involve charging or discharging energy storage. Let $\mathcal{A} \subseteq [-1, 1]$ denote the action space for each aggregator, where each action $a_t^n \in \mathcal{A}$ represents the proportion of storage capacity to charge (if $a_t^n > 0$) or discharge (if $a_t^n < 0$) at time t .¹ Actions are drawn according to the policy learned by the aggregator at bus n and executed by individual prosumers. To ensure feasibility, we implement *action masking*, a common technique in RL that eliminates invalid actions based on the current state.

States: In our model, each aggregator at bus n is characterized by three state variables at time t : (i) storage level $e_t^n \in [0, 1]$ (as a percentage of capacity), (ii) net load q_t^n (unrelated to storage charging/discharging), and (iii) the current hour of the day, $T_{\text{hour}}(t)$. Let the state space be $\mathcal{S} \subseteq \mathbb{R}^3$, where each element is a tuple of the form $s_t^n := (e_t^n, q_t^n, T_{\text{hour}}(t)) \in \mathcal{S}$.

Among the state variables, the net load q_t^n is an exogenous random variable, mainly driven by solar irradiance and consumption behavior, and is unaffected by aggregators' actions. Time of day is deterministic. The only state variable affected by decisions is the storage level, which evolves according to the following rule after action a_t^n :

$$e_{t+1}^n := \max\{\min\{e_t^n + a_t^n, 1\}, 0\}. \quad (2)$$

Mean Field: With states and actions defined, the next step in standard RL (and Markov games) setups is to specify each agent's reward function. In our setting, however, rewards do not depend on the individual actions of other agents but only on the aggregate behavior of the population. Since we consider a large (infinite) number of agents, the influence of

¹For prosumer i at bus n , the corresponding average charging or discharging power over the interval is $a_t^n \bar{E}_i^n / \Delta t$, up to efficiency losses. Explicit storage power limits can be incorporated by restricting the feasible action set, for example through $|a_t^n| \bar{E}_i^n / \Delta t \leq \bar{P}_i^n$, where $\bar{P}_i^n$ is the power capacity of prosumer i at bus n .

any single agent is negligible. Instead, each agent responds to the population-wide distribution of states and actions, known as the mean field. For simplicity and tractability, we begin by assuming that the state space $\mathcal{S}$ and action space $\mathcal{A}$ are discrete and finite. Let $\mathcal{P}(\mathcal{S} \times \mathcal{A})$ denote the space of probability distributions over the joint state–action space $\mathcal{S} \times \mathcal{A}$; the formal definition of a mean field at bus n and time t , $\mathcal{L}_t^n \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, is as follows:

$$\mathcal{L}_t^n(s, a) := \lim_{M_n^p \rightarrow \infty} \frac{1}{M_n^p} \sum_{i=1}^{M_n^p} \mathbb{1}_{(s_{it}^n, a_{it}^n)=(s,a)}, \quad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

where $\mathbb{1}_{(s_{it}^n, a_{it}^n)=(s,a)} = 1$ if prosumer i is in state s and takes action a at time t , and 0 otherwise. Intuitively, the mean field is the limit of the histogram of joint state–action pairs as the number of prosumers goes to infinity. For notation simplicity, let $\mathcal{L}_t := (L_t^1, \dots, L_t^N)$ denote the vector of all buses' mean-field; then our formulation allows us to equivalently treat the LMP as a function of $\mathcal{L}_t$, which we write as $\lambda_t^n(\mathcal{L}_t)$.

Reward: We now define the single-period reward function $r_n : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^H \rightarrow \mathbb{R}$ for each aggregator n . Let s_t^n and a_t^n be the state and action at time t , and let $\lambda^n := (\lambda_1^n, \dots, \lambda_H^n)$ denote the LMP profile at bus n . The reward is given by

$$r_n(s_t^n, a_t^n, \lambda^n) = -\lambda_{T_{\text{hour}}(t)}^n \cdot \bar{E}_n \cdot (\Phi(e_t^n, a_t^n, \eta_n) + q_t^n), \quad (3)$$

where $\eta_n \in (0, 1]$ is the storage efficiency (assumed uniform across prosumers at bus n), and $\Phi(\cdot)$ adjusts the action for efficiency losses:

$$\Phi(e, a, \eta) = \begin{cases} \max\{-e, a\} \cdot \eta, & \text{if } a < 0 \text{ (discharging)}, \\ \min\{1 - e, a\} / \eta, & \text{if } a \geq 0 \text{ (charging)}. \end{cases} \quad (4)$$

The reward represents net profit, computed as negative price multiplied by total energy demand. To promote exploration and improve learning stability, we incorporate *entropy regularization* into the reward function. Specifically, the single-period reward is modified as:

$$r_n^{\text{REG}}(s, a, \lambda^n) = r_n(s, a, \lambda^n) - \Omega(\pi(\cdot | s, \lambda^n)), \quad (5)$$

where $\Omega(\cdot)$ is a ρ -strongly convex regularization function. A common choice is the negative entropy:

$$\Omega(\pi(\cdot | s, \lambda^n)) = \beta \sum_{a \in \mathcal{A}} \pi(a | s, \lambda^n) \log \pi(a | s, \lambda^n),$$

with $\beta > 0$ controlling the strength of regularization.

Policy: At each decision time t , aggregator n learns a policy $\pi_t^n : \mathcal{S} \times \mathbb{R}^H \rightarrow \mathcal{P}(\mathcal{A})$, which maps each state and LMP profile to a distribution over actions. Given the LMP profile $\lambda^n \in \mathbb{R}^H$ and initial state s_0^n , the *regularized value function* under policy π_t^n is defined as the expected sum of discounted rewards:

$$V_n^{\text{REG}}(s_0^n, \pi_t^n, \lambda^n) = \mathbb{E} \left[\sum_{\tau=0}^{\infty} \gamma_n^\tau r_n^{\text{REG}}(s_\tau^n, a_\tau^n, \lambda^n) \right], \quad (6)$$

where $\gamma_n \in (0, 1]$ is the discount factor, and the expectation is over the trajectory induced by the policy and external uncertainties. Each aggregator seeks to learn an optimal policy π_t^{n*} that maximizes V_n^{REG} at each time t .

B. Mean Field Equilibrium

An MFE is a fixed point in a dynamic system of infinitely many interacting agents, characterized by two key properties: (i) **optimality**, where each agent's strategy is optimal given the mean field, and (ii) **consistency**, where the mean field coincides with the distribution of states and actions induced by all agents following that strategy. Let $\Pi := \{\pi | \pi : \mathcal{S} \times \mathbb{R}^H \rightarrow \mathcal{P}(\mathcal{A})\}$ denote the space of state-dependent stochastic policies. Let $\pi = (\pi^1, \dots, \pi^N)$ denote the policy profile of all aggregators. The formal definition of an MFE is given as follows:

Definition 1 (MFE). A profile $\{(\pi^{n*}, \lambda^{n*})\}_{n=1}^N$ is a mean field equilibrium if, for each aggregator n and any state $s^n \in \mathcal{S}$ at time t , the following conditions hold:

- 1) **Optimality:** Given λ^{n*} , each aggregator finds the optimal policy π^{n*} , such that for any policy $\pi^n \in \Pi$,

$$V_n^{\text{REG}}(s^n, \pi^{n*}, \lambda^n) \geq V_n^{\text{REG}}(s^n, \pi^n, \lambda^n).$$

- 2) **Consistency:** If all aggregators follow π^* , then the mean field distribution evolves according to: $\mathcal{L}_{t+1}^n = \Gamma^n(\mathcal{L}_t^n, \pi^{n*})$, where Γ^n is the consistency operator defined in the following form:

$$\begin{aligned} \Gamma^n(\mathcal{L}^n, \pi^n)(s', a') &:= \zeta \frac{1}{|\mathcal{S}||\mathcal{A}|} \\ &+ (1 - \zeta) \sum_{s, a} \mathcal{L}^n(s, a) \cdot P^n(s' | s, a') \cdot \pi^n(a' | s, \lambda), \end{aligned} \quad (7)$$

for all $s' \in \mathcal{S}, a' \in \mathcal{A}$, and $\zeta \in (0, 1)$ is the probability of a uniform noise into the MF update. $P^n(s' | s, a')$ denotes the state transition probability for the aggregator at bus n . The corresponding LMP obtained through the ED problem (11) – (14), $\lambda_{t+1}^n(\mathcal{L}_{t+1})$, must satisfy $\lambda_{t+1}^n(\mathcal{L}_{t+1}) = [\lambda^{n*}]_{T_{\text{hour}}(t+1)}$.

The parameter ζ in (7) is a regeneration probability that gives each prosumer a small chance of resetting to a uniformly sampled state. Economically, this models prosumer turnover at a bus, new prosumers entering and existing ones exiting, and keeps the environment dynamic even at steady state. This regeneration mechanism prevents the mean field from becoming static and supports continual learning.

We establish in Appendix B the conditions under which such an MFE exists and is unique.

Given the definition above, a natural fixed-point iteration for approximating an MFE proceeds as follows: (i) Fix $\mathcal{L}_t^n$ and λ_t^n . For a given state s^n , aggregator n computes its optimal policy π_t^{n*} . (ii) Update the mean field using $\mathcal{L}_{t+1}^n = \Gamma^n(\mathcal{L}_t^n, \pi_t^{n*})$ and compute the updated LMP $\lambda_{t+1}^n(\mathcal{L}_{t+1})$ by solving the ED problem (11)–(14). If $t + 1$ marks the start of a new day, update λ^n using the most recent H LMPs. (iii) Repeat steps (i)–(ii) until λ_{t+1}^n stabilizes.

The key idea behind this framework is that information about the mean field is embedded in the LMPs. Aggregators update their policies based on observed LMPs, and the mean field evolves in response to these updated policies. We also note that an MFE is not generally a Nash equilibrium, as the

latter is typically defined for games with a finite number of agents. Whether an MFE arises as the limit of an M -agent Nash equilibrium as $M \rightarrow \infty$ is a nontrivial question. [20] provides sufficient conditions under which the infinite-agent MFE policy yields an ϵ -Nash equilibrium for the corresponding finite- M game. In this work, we focus on the properties of the MFE in the infinite-agent setting and leave its connection to finite-agent Nash equilibria to future research.

V. THE TWO-PHASE RL ALGORITHM

We propose a two-phase, distributed mean-field RL algorithm executed at each time period t . **Phase 1 (Training):** each aggregator independently trains its policy using an RL algorithm. **Phase 2 (Execution):** prosumers act according to their aggregator's trained policy. Since policies are stochastic, identical prosumers may take different actions. These actions are aggregated into bids and submitted to the market clearing process, after which LMPs for time period t are determined by solving the ED problem.

Aggregators do not have direct knowledge of external uncertainty distributions, such as the net load. However, an important feature of electricity markets is that LMP fluctuations implicitly reflect both market dynamics and exogenous uncertainties. Following the approach in [15], we allow each aggregator to maintain a belief over the LMPs for each hour of the day, represented by a vector of length H . These beliefs are formed from historical observations and guide the RL-based policy training in Phase 1. After market clearing, aggregators update their beliefs using the realized LMPs.

All prosumers at the same bus share the belief vector maintained by their aggregator, since they face identical LMPs. Let $\hat{\lambda}_t^n \in \mathbb{R}^H$ denote the belief vector of aggregator n at time t . Once the ISO solves the ED problem and returns LMP λ_t^n at bus n , the aggregator updates its belief as

$$\hat{\lambda}_{t+1}^n := \hat{\lambda}_t^n - \frac{\delta_n}{\sqrt{T_{\text{day}}(t) + 1}} \left((\hat{\lambda}_t^n)^\top \mathbf{1}_{T_{\text{hour}}(t)} - \lambda_t^n \right) \mathbf{1}_{T_{\text{hour}}(t)}, \quad (8)$$

where $\delta_n \in [0.5, 1]$ is a learning-rate parameter and $\mathbf{1}_{T_{\text{hour}}(t)} \in \mathbb{R}^H$ is the unit vector with a 1 at entry $T_{\text{hour}}(t)$ and zeros elsewhere. Thus, at each time t , only the belief corresponding to hour $T_{\text{hour}}(t)$ is updated.

Training Phase: At the beginning of each time period t , each aggregator fixes its LMP belief and trains a policy using an RL algorithm, referred to generically as *Alg* (such as Proximal Policy Optimization (PPO) [21], Trust Region Policy Optimization (TRPO) [22], or Soft Actor-Critic (SAC) [23]), for T_{train} steps.

Execution Phase: After each aggregator has learned a policy, this policy is distributed to its prosumers. At bus n , prosumer i takes an action from the policy $a_{it}^n \sim \pi_t^{n*}(\cdot | s_{it}^n, \lambda_t^n)$. Also, each prosumer i and consumer j 's original net demand are realized following the distributions $\mathcal{Q}_n^p$ and $\mathcal{Q}_n^c$, respectively. The net demand in the quantity of energy now has the following forms for prosumer i at bus n at time t : $d_{it}^n = (\Phi(e_{it}^n, a_{it}^n, \eta_n) + q_t^n) \bar{E}_i^n$, and similarly, for each consumer j at bus n at time t : $d_{jt}^n = q_{jt}^n \bar{E}_j^n$.

The transition of the state of charge for each prosumer i at bus n from time t to $t + 1$ is defined as follows:

$$e_{i,t+1}^n = \begin{cases} \text{Uniform}(0, 1), \text{ with probability } \zeta, \\ \max \{ \min \{ e_{it}^n + a_{it}^n, 1 \}, 0 \}, \text{ with probability } 1 - \zeta, \end{cases} \quad (9)$$

where ζ is the regeneration probability introduced in (7). The aggregator at n updates its storage level as:

$$e_{t+1}^n = \frac{1}{\bar{E}_n} \sum_{i=1}^{M_n^p} e_{i,t+1}^n \bar{E}_i^n, \quad (10)$$

which is a weighted average of all its prosumers storage level. This storage level is then used as the initial storage level state to begin the RL training for time t for T_{train} steps. The pseudo-code is presented in Algorithm 1 in Appendix C.

It is important to note that the RL training in Phase 1 is performed offline and does not operate on real-time system timescales. Although historical LMPs are used to initialize each aggregator's belief vector, the training itself does not rely on historical prices. Instead, during offline training the aggregator interacts with a simulated environment in which LMPs are generated endogenously by solving the ED model in response to the aggregator's actions and the evolving mean field. This preserves the closed-loop feedback between aggregator behavior and market prices while avoiding any real-time computational burden. Aggregators periodically update their policies by retraining them offline using newly observed LMPs, updated belief vectors, and accumulated historical trajectories, and these computations can be carried out on dedicated cloud or local servers without interfering with market operations. Once an updated policy is obtained, it is deployed in real time by simply broadcasting it to prosumers. During real-time operation, only Phase 2 is executed, and this phase requires negligible computation. Thus, the proposed framework does not impose any real-time computational burden.

Remark on distribution network constraints. The baseline formulation models aggregators at transmission-level buses and therefore does not explicitly include distribution network constraints. However, such constraints can be incorporated into our framework without modifying its overall structure. In particular, voltage or line-flow limits at the distribution level can be enforced through negative reward terms during RL training. This approach follows the same principle used in peer-to-peer energy trading in distribution networks [24]. Such extensions are not pursued in this work, as our focus is on presenting the framework for wholesale-level market interactions.

VI. NUMERICAL EXPERIMENT

We evaluate the proposed framework on the 37-bus synthetic Oahu network from [25]. Each Hawaiian Electric generating plant is mapped to its nearest bus using data from [26], resulting in 26 generators: 4 oil, 2 biomass, 17 utility-scale solar (distinct from household PV), and 3 wind units. Oil and biomass units use quadratic costs $C(p) = ap^2 + bp$, where p is MW and $C(p)$ is \$/h, with parameter ranges from [27], [28]. For oil, $a \in [0.0059, 0.0342]$ \$/MW²h and

$b = 19.98 \text{ \$/MWh}$; for biomass, $a \in [0.001, 0.002] \text{ \$/MW}^2\text{h}$ and $b \in [28.45, 52.65] \text{ \$/MWh}$, sampled uniformly. Solar and wind have zero marginal cost. Their capacity factors follow historical hourly profiles, solar from [29], wind from [30], and are multiplied by triangular noise: $\Delta(0.8, 1.2, 1)$ for solar and $\Delta(0.5, 1.5, 1)$ for wind.

Each bus hosts 650 prosumers and 2,000 consumers. Prosumers consist of 500 small, 100 medium, and 50 large types with storage capacities of 10, 20, and 30 kWh. Demand profiles follow [31]. Larger storage capacities imply proportionally scaled net-load levels in the dataset. To introduce uncertainty, each agent's mean daily demand is scaled by $\Delta(0.8, 1.2, 1)$.

We implement the PPO RL algorithm with $T_{\text{train}} = 1,200$, $H = 12$ (2-hour intervals), $\delta_n = 0.9$, $\zeta = 0.01$, and entropy regularization coefficient $\beta = 0.01$. Each simulation runs 50 days and is repeated five times using different seeds. Experiments are run on a Windows 11 machine with an Intel i7 CPU and an NVIDIA RTX 4070 GPU.

Figure 1 shows hub prices over the first and last 3 days of training. We compare 3 scenarios: the proposed MFE-based framework, a decentralized heuristic algorithm (DHA; see Appendix D), and a baseline without storage (grid-tied PV only). Initially, MFE prices behave similarly to the no-storage case, while DHA exhibits the lowest volatility. As learning progresses, prices under MFE become less volatile and converge to a stable daily pattern, indicating convergence to a steady state.

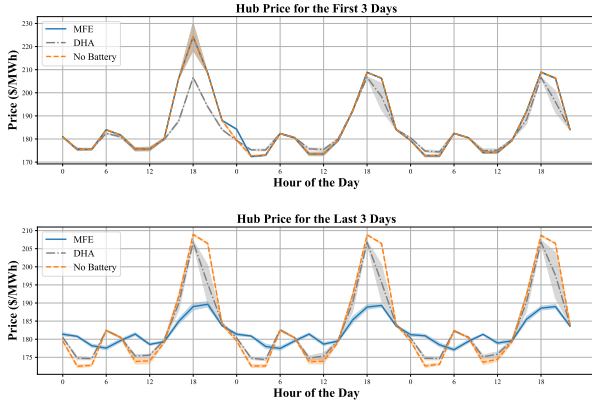


Figure 1. LMP hub price Comparisons (shaded areas show one-standard-deviation error bounds across simulations).

To quantify price volatility, we adopt the incremental mean volatility (IMV) metric from [32]: $\text{IMV} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T |\lambda_{t+1} - \lambda_t|$. Figure 2 reports IMV values over the final 3 days. The MFE scenario yields significantly lower IMV, indicating reduced price volatility.

We also compute the *ex-post* daily cost for prosumers and consumers, defined as the sum over time of realized LMPs multiplied by cleared energy quantities. Figure 3 shows that total daily costs are lowest under MFE for both groups.

Finally, we examine the impact on the “duck curve,” that is, the net demand curves obtained by subtracting solar generation from the gross system load. Figure 4 compares the net demand

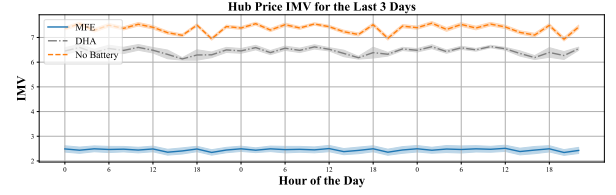


Figure 2. Comparison of IMV over the last 3 days across scenarios. Shaded areas show one-standard-deviation error bounds across simulations.

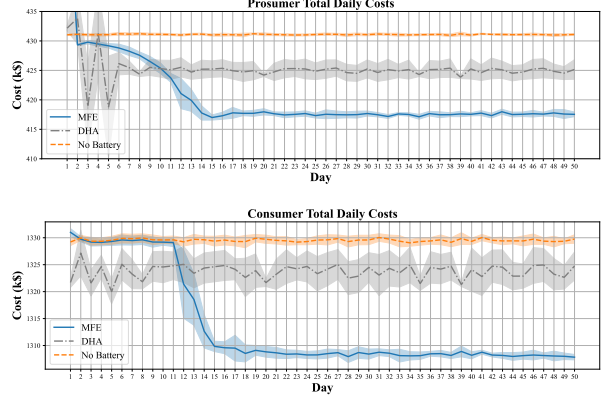


Figure 3. Comparison of daily costs for prosumers and consumers. Shaded areas show one-standard-deviation error bounds across simulations.

under the no-storage, DHA, and MFE cases, averaged over the final ten days. The MFE policy produces the greatest degree of load shifting, charging during periods of abundant midday sunshine and reducing the evening peak, thereby smoothing the aggregate load profile.

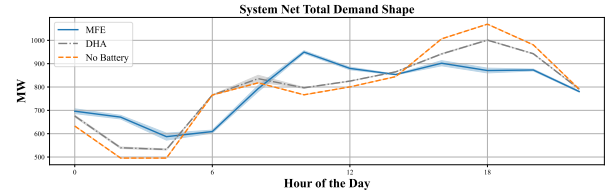


Figure 4. Comparison of net load curves.

VII. CONCLUSION

This paper develops an algorithmic framework for integrating DERs into wholesale electricity markets through decentralized, RL-enabled aggregators. Aggregators learn storage charging and discharging strategies for prosumers under uncertainty in renewable output, demand, and market prices. A central idea is to treat LMPs as a mean-field signal that reflects aggregate supply-demand interactions and network constraints. Within this hybrid MFC–MFE setting, we propose a two-phase RL algorithm that enables fully decentralized policy learning.

We also establish sufficient conditions for the existence and uniqueness of an MFE and show convergence of the resulting three-step learning process in the infinite-agent limit. Numerical experiments demonstrate that coordinated storage control within this framework substantially reduces LMP volatility and

lowers costs for both prosumers and consumers, highlighting the value of market-integrated DER coordination.

Future work will proceed in three directions. First, although Algorithm 1 provides a natural fixed-point-style approximation of an MFE, each policy update requires an evolving LMP belief, making the method heuristic in practice. An important direction is to analyze the belief-update dynamics and determine conditions under which LMP beliefs, and consequently Algorithm 1, converge. Second, we aim to characterize the relationship between the infinite-population MFE and the Markov-Nash equilibrium of a finite but large number of aggregators, identifying regimes where the latter converges to the former. Third, we plan to incorporate strategic aggregators who internalize their price impact, enabling a systematic study of market power and strategic interactions within the mean-field framework.

AI USAGE DISCLOSURE

The authors used ChatGPT solely for language polishing and editorial refinement. All research ideas, modeling, analysis, and results were developed independently by the authors.

APPENDIX A ECONOMIC DISPATCH (ED)

To provide the specific formula of LMPs, we present a basic economic dispatch (ED) problem solved by an ISO. Although actual dispatch problems are considerably more complex, this formulation retains the essential constraints: power balance, generation capacity, and transmission limits, and is sufficient to demonstrate how prices are formed and how they relate to system demand. Importantly, our framework is not restricted to this simplification. A key strength of the framework is that it requires no modification to ISO operations: system operators can continue using their full-scale market-clearing processes without adjustment. All learning and coordination occur at the aggregator level, ensuring full compatibility with existing market structures. The simplified ED problem is given as follows:

$$\min_{p_{1t}, \dots, p_{Gt}} \sum_{g=1}^G C_g(p_{gt}) \quad (11)$$

$$\text{s.t. } \sum_{g=1}^G p_{gt} = \sum_{n=1}^N D_t^n, \quad (\lambda_t^{\text{HUB}}) \quad (12)$$

$$-\bar{F}_l \leq \sum_{n=1}^N \text{PTDF}_{ln} \left(\sum_{g \in G_n} p_{gt} - D_t^n \right) \leq \bar{F}_l, \quad (\underline{\mu}_{lt}, \bar{\mu}_{lt}) \quad (13)$$

$$0 \leq p_{gt} \leq \bar{p}_g, \quad (\underline{\nu}_{gt}, \bar{\nu}_{gt}) \quad \forall g \in \{1, \dots, G\}. \quad (14)$$

Here, the variables p_{gt} represent the average power output of generator g in time interval t . The parameter PTDF_{ln} denotes the power transfer distribution factor for line l and bus n , which encodes how nodal injections distribute across lines under a DC power-flow approximation, that is, Kirchhoff's current and voltage laws in a linearized network; $\bar{F}_l$ represents

the maximum allowable power flow on line l , and $\bar{p}_g$ represents the generation capacity of generator g . In addition, the dual variables are listed on the right-hand-side of each constraint. The dual of the power balance constraint (12), λ_t^{HUB} , represents the so-called hub price, and $\underline{\mu}_{lt}, \bar{\mu}_{lt}, \underline{\nu}_{gt}, \bar{\nu}_{gt}$ are the dual variables corresponding to line flow limits of each line l , and generator output limits of each generator g , respectively. The LMP at bus $n \in \{1, \dots, N\}$ and time t , denoted λ_t^n , is derived as follows:

$$\lambda_t^n := \frac{\partial \mathcal{L}_t}{\partial D_t^n} = \lambda_t^{\text{HUB}} - \sum_{l=1}^L \text{PTDF}_{ln} (\underline{\mu}_{lt} - \bar{\mu}_{lt}), \quad (15)$$

where $\mathcal{L}_t$ represents the Lagrangian function of the ED problem.

APPENDIX B EXISTENCE AND UNIQUENESS OF AN MFE

In establishing the existence and uniqueness of an MFE, a key technical condition needed is the Lipschitz continuity of the mapping $\lambda_t^n(\mathbf{D}_t)$. The following result, established in [15], provides sufficient conditions under which this property holds. For completeness, we restate it here, along with the required constraint qualification assumption.

Assumption 1. (LICQ) Let $X(\mathbf{D}_t)$ denote the feasible region of the economic dispatch problem defined in (11)–(14). We assume that, for all t and all $\mathbf{D}_t$ such that $X(\mathbf{D}_t)$ is non-empty, the linear independence constraint qualification (LICQ) holds at every feasible point in $X(\mathbf{D}_t)$.

Proposition 1. (Lipschitz Continuity of LMPs) [15] Given Assumption 1 and that each generator's cost function $C_g(\cdot)$ is strongly convex and quadratic. Then, for all buses $n = 1, \dots, N$, the LMP $\lambda_t^n(\mathbf{D}_t)$ is single-valued and Lipschitz continuous with respect to $\mathbf{D}_t$; that is, there exists a constant $L_\lambda > 0$ such that for any $\mathbf{D}_t$ and $\tilde{\mathbf{D}}_t \geq 0$, $|\lambda_t^n(\mathbf{D}_t) - \lambda_t^n(\tilde{\mathbf{D}}_t)| \leq L_\lambda \|\mathbf{D}_t - \tilde{\mathbf{D}}_t\|$.

We now prove that the proposed 3-step framework in Section IV-B defines a contraction mapping over the space of policy and price profiles, guaranteeing convergence to a unique fixed point. This fixed point corresponds exactly to an MFE, thereby establishing its existence and uniqueness. Unless otherwise noted, all norms here are ℓ_1 -norms.

Theorem 1 (Lipschitz Optimal Solution). Under the same assumptions of Proposition 1, in the regularized MFG, for any belief $\lambda^n, \lambda^{n'}$, there exists $L_1 \geq 0$ such that:

$$\sup_{s \in \mathcal{S}} \|\pi^{n*}(s, \lambda^n) - \pi^{n*}(s, \lambda^{n'})\| \leq L_1 \|\lambda^n - \lambda^{n'}\|. \quad (16)$$

Proof Sketch. We first note that under regularization, V^{REG} is a strongly concave function and therefore admits unique optimal solution. We then proceed in 3 steps. (i) For a fixed policy π , the difference between the optimal Q -functions Q_λ^* and $Q_{\lambda_2}^*$ are bounded by $L_r \|\lambda - \lambda_2\|$, due to the Lipschitz continuity of the reward function. This implies that Q_λ^* is Lipschitz continuous with respect to λ . (ii) Define $q_\lambda^s := r(s, \cdot, \lambda) + \gamma \sum_{s'} Q_\lambda^*(s') P(s'|s, \cdot)$, the action-value

vector before regularization. This inherits the Lipschitz property with constant $(L_r + \frac{\gamma L_r}{1-\gamma})$. (iii) By Fenchel duality, the optimal policy is given by $\pi^{n*}(s, \lambda) = \nabla \Omega^*(q_\lambda^s)$, where Ω^* is the conjugate of the ρ -strongly convex regularizer Ω . Since Ω^* is $1/\rho$ -smooth, π^{n*} is Lipschitz with constant $L_1 = \frac{1}{\rho}(L_r + \frac{\gamma L_r}{1-\gamma})$. $\square$

Theorem 2 (Lipschitz Continuity of Γ^n). *Under the same assumptions of Proposition 1, for any MF $\mathcal{L}^n, \mathcal{L}^{n'}$ and policies $\pi^n, \pi^{n'}$, given the belief λ^n , there exists $L_2, L_3 \geq 0$ such that*

$$\|\Gamma^n(\mathcal{L}^n, \pi^n) - \Gamma^n(\mathcal{L}^{n'}, \pi^n)\| \leq L_2 \|\mathcal{L}^n - \mathcal{L}^{n'}\|, \quad (17)$$

$$\begin{aligned} \|\Gamma^n(\mathcal{L}^n, \pi^n) - \Gamma^n(\mathcal{L}^n, \pi^{n'})\| \\ \leq L_3 \sup_{s \in S} \|\pi^n(s, \lambda^n) - \pi^{n'}(s, \lambda^n)\|. \end{aligned} \quad (18)$$

Proof Sketch. We expand the no-regeneration transition operator as

$$\tilde{\Gamma}^n(\mathcal{L}^n, \pi^n)(s', a') = \sum_{s, a} \mathcal{L}^n(s, a) P^n(s'|s, a) \pi^n(a'|s).$$

Fixing π^n , the difference $\tilde{\Gamma}^n(\mathcal{L}^n, \pi^n) - \tilde{\Gamma}^n(\mathcal{L}^{n'}, \pi^n)$ can be bounded linearly in $\|\mathcal{L}^n - \mathcal{L}^{n'}\|$ since the sum is a linear operator in $\mathcal{L}^n$ with probability weights. Similarly, fixing $\mathcal{L}^n$, the difference in π^n can be bounded by $\sup_s \|\pi^n(\cdot|s) - \pi^{n'}(\cdot|s)\|$. Including regeneration with probability ζ scales both terms by $(1 - \zeta)$, yielding Lipschitz constants $L_2 = L_3 = 1 - \zeta$. $\square$

We now present the main theorem regarding the MFE existence and uniqueness.

Theorem 3 (Existence and Uniqueness of MFE). *Under the same LICQ and strongly convex quadratic cost function assumptions as in Proposition 1, given that $L_1 L_{MF} L_3 + L_2 < 1$, there exists a unique MFE following the 3-step procedure.*

Proof Sketch. We first show that LMPs are Lipschitz continuous with respect to the mean-field (MF) profile $\mathcal{L}_t := (\mathcal{L}_t^1, \dots, \mathcal{L}_t^N)$. Fixing the consumer demand, the total demand at each bus n is a weighted sum over $\mathcal{L}_t^n(s, a)$ scaled by the storage rating $\bar{E}_n$, implying

$$\|D_t - D_{t'}\|_1 \leq \max_n \bar{E}_n \|\mathcal{L}_t - \mathcal{L}_{t'}\|_1.$$

By Proposition 1, LMPs are Lipschitz in total demand with constant L_λ . Combining the two gives

$$|\lambda_t^n - \lambda_{t'}^n| \leq L_{MF} \|\mathcal{L}_t - \mathcal{L}_{t'}\|_1, \quad \text{with } L_{MF} := L_\lambda \max_n \bar{E}_n.$$

Now consider the update from time t to $t + H$ (same hour on consecutive days). We then have:

$$\begin{aligned} \|\Gamma^n(\mathcal{L}_t^n, \pi^{n*}(s^n, \lambda_t^n)) - \Gamma^n(\mathcal{L}_{t+H}^n, \pi^{n*}(s^n, \lambda_{t+H}^n))\|_1 \\ \leq L_3 \|\pi^{n*}(s^n, \lambda_t^n) - \pi^{n*}(s^n, \lambda_{t+H}^n)\|_1 + L_2 \|\mathcal{L}_t^n - \mathcal{L}_{t+H}^n\|_1 \\ \leq L_1 L_3 \|\lambda_t^n - \lambda_{t+H}^n\|_1 + L_2 \|\mathcal{L}_t^n - \mathcal{L}_{t+H}^n\|_1 \\ \leq L_1 L_{MF} L_3 \|\mathcal{L}_t^n - \mathcal{L}_{t+H}^n\|_1 + L_2 \|\mathcal{L}_t^n - \mathcal{L}_{t+H}^n\|_1. \end{aligned}$$

Thus, the full update operator is a contraction with constant $(L_1 L_{MF} L_3 + L_2)$ under the ℓ_1 norm. If this constant is strictly less than 1, Banach's fixed-point theorem guarantees existence and uniqueness of the MFE. $\square$

Remark 1. *This sufficient contraction condition may fail when small changes in aggregate DER demand lead to large LMP changes, such as near-binding generation or transmission constraints, or when aggregate storage capacity is large, policy responses are sharp, regularization is weak, or regeneration is small. In such cases, uniqueness and convergence are not guaranteed by the theorem, although the learning procedure may still perform well computationally.*

APPENDIX C

THE TWO-PHASE RL ALGORITHM

We present the pseudo-code of the two-phase RL algorithm in Algorithm 1 based on Section V.

Algorithm 1: A two-phase distributed mean-field RL algorithm with LMP beliefs and entropy regularization

Input: Initial battery states $e_0^n \in [0, 1]$, initial LMP beliefs $\hat{\lambda}_0^n \in \mathbb{R}^H$ with learning rates $\delta_n \in [0.5, 1]$, demand shapes $\mathcal{Q}_n^p, \mathcal{Q}_n^c$; training step T_{train} ; random regeneration probability ζ ; an RL algorithm Alg ; time functions $T_{\text{hour}}(\cdot), T_{\text{day}}(\cdot)$.

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

Training phase

for Bus $n = 1, \dots, N$ **do**

 Train the aggregator for T_{train} steps using Alg with initial storage e_t^n under $\hat{\lambda}_t^n$ to get π_t^{n*} ;

end

Execution phase

for Bus $n = 1, \dots, N$ **do**

foreach Prosumer $i = 1, \dots, M_n^p$ **do**

 Get net demand $q_{it}^n \sim \mathcal{Q}_n^p$;

 Take actions $a_{it}^n \sim \pi_t^{n*}$;

 Storage state transition as in (9);

end

for Consumer $j = 1, \dots, M_n^c$ **do**

 Get demand $q_{jt}^n \sim \mathcal{Q}_n^c$;

end

 Compute next storage state e_{t+1}^n as in (10);

end

 Solve an ED problem to get LMPs λ_t^n for all n ;

 Update the LMP belief as in (8);

end

APPENDIX D

DECENTRALIZED HEURISTIC ALGORITHM (DHA)

To serve as a benchmark to compare with the proposed MFE learning framework, we propose a decentralized heuristic algorithm (DHA). At each timestep t , each aggregator n determines a storage action by first using the same LMP belief vector as in (8). Then each aggregator n defines the low- and high-price windows using pre-determined and fixed thresholds λ_{low}^n and λ_{high}^n (where $\lambda_{\text{high}}^n > \lambda_{\text{low}}^n$) over the next day:

$$\mathcal{T}_{\text{low}}^n(t) := \left\{ h : \left[\hat{\lambda}_t^n \right]_h \leq \lambda_{\text{low}}^n \right\}, \quad (19)$$

$$\mathcal{T}_{\text{high}}^n(t) := \left\{ h : \left[\hat{\lambda}_t^n \right]_h \geq \lambda_{\text{high}}^n \right\}, \quad (20)$$

where $\left[\hat{\lambda}_t^n \right]_h$ indicates the h -th entry of the vector $\hat{\lambda}_t^n$. The cardinality of the two sets can be computed as $N_{\text{low}}^n(t) := |\mathcal{T}_{\text{low}}^n(t)|$ and $N_{\text{high}}^n(t) := |\mathcal{T}_{\text{high}}^n(t)|$. DHA spreads charging across the low-price window rather than charging aggressively at a single hour. If $N_{\text{low}}^n(t) > 0$, define the target battery level at the end of the low-price window as

$$e_{\text{ch},t}^{n*} := \min \{1, e_t^n + N_{\text{low}}^n(t)(1 - e_t^n)\}, \quad (21)$$

and define the planned per-step charging amount as

$$\bar{a}_{nt}^C := \frac{e_{\text{ch},t}^{n*} - e_t^n}{N_{\text{low}}^n(t)}. \quad (22)$$

Similarly, if $N_{\text{high}}^n(t) > 0$, define the target battery level at the end of the high-price window, and the planned per-step discharging amount as follows:

$$e_{\text{dis},t}^{n*} := \max\{0, e_t^n - N_{\text{high}}^n(t)e_t^n\}, \quad (23)$$

$$\bar{a}_{nt}^D := \frac{e_t^n - e_{\text{dis},t}^{n*}}{N_{\text{high}}^n(t)}. \quad (24)$$

Now, given the current level e_t^n , the deterministic action $a_t^{n*} \in [-1, 1]$ is

$$a_t^{n*} = \begin{cases} -\min\{\bar{a}_{nt}^D, e_t^n\}, & \text{if } T_{\text{hour}}(t) \in \mathcal{T}_{\text{high}}^n(t), \\ \min\{\bar{a}_{nt}^C, 1 - e_t^n\}, & \text{if } T_{\text{hour}}(t) \in \mathcal{T}_{\text{low}}^n(t), \\ 0, & \text{otherwise,} \end{cases} \quad (25)$$

where we prioritize discharging if both windows are active. To reduce synchronized behavior, each aggregator draws from an i.i.d. random factor $\xi_t^n \sim \text{Uniform}(\alpha, 1)$, where $\alpha \in (0, 1)$, and implements the final action $a_t^n := \xi_t^n a_t^{n*}$, where in our numerical setup, we choose $\alpha = 0.8$. Finally, the battery level is updated via (2) and the resulting aggregate net demand bids are cleared by the standard ED (11)–(14).

REFERENCES

- [1] OhmConnect, “OhmConnect paid members \$2.7M and saved 1.5 GWh of energy during recent California heat wave,” October 2022. Accessed: 2025-06-23.
- [2] Tesla, Inc., “Tesla virtual power plant,” June 2025. Accessed: 2025-06-23.
- [3] J. E. Contreras-Ocana, M. A. Ortega-Vazquez, and B. Zhang, “Participation of an energy storage aggregator in electricity markets,” *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1171–1183, 2017.
- [4] Z. Gao, K. Alshehri, and J. R. Birge, “On efficient aggregation of distributed energy resources,” in *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 7064–7069, IEEE, 2021.
- [5] M. Khoshjahan and M. Kezunovic, “Robust bidding strategy for aggregation of distributed prosumers in flexiramp market,” *Electric Power Systems Research*, vol. 209, p. 107994, 2022.
- [6] J. Iria, F. Soares, and M. Matos, “Optimal bidding strategy for an aggregator of prosumers in energy and secondary reserve markets,” *Applied Energy*, vol. 238, pp. 1361–1372, 2019.
- [7] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, “Deep reinforcement learning for strategic bidding in electricity markets,” *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343–1355, 2020.
- [8] A. Fallahi, J. M. Rosenberger, V. C. Chen, W.-J. Lee, and S. Wang, “Linear programming for multi-agent demand response,” *IEEE Access*, vol. 7, pp. 181479–181490, 2019.
- [9] M. Shafie-khah and J. P. Catalão, “A stochastic multi-layer agent-based model to study electricity market participants behavior,” *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 867–881, 2014.
- [10] C. Chen, S. Bose, T. D. Mount, and L. Tong, “Wholesale market participation of DERAs: DSO-DETA-ISO coordination,” *IEEE Transactions on Power Systems*, vol. 39, no. 5, pp. 6605–6614, 2024.
- [11] A. Liu and Z. Zhao, “Multi-agent learning in repeated double-side auctions for peer-to-peer energy trading,” in *Proceedings of the 54th Hawaii International Conference on System Sciences*, p. 3121, 2021.
- [12] K. Zhang, Z. Yang, and T. Başar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” *Handbook of reinforcement learning and control*, pp. 321–384, 2021.
- [13] A. C. Kizilkale and R. P. Malhamé, “Collective target tracking mean field control for markovian jump-driven models of electric water heating loads,” *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 1867–1872, 2014.
- [14] Y. Chen, A. Bušić, and S. P. Meyn, “State estimation for the individual and the population in mean field control with application to demand dispatch,” *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1138–1149, 2017.
- [15] C. Feng and A. L. Liu, “Decentralized integration of grid edge resources into wholesale electricity markets via mean-field games,” *arXiv preprint arXiv:2503.07984*, 2025.
- [16] X. Guo, A. Hu, R. Xu, and J. Zhang, “Learning mean-field games,” *Advances in neural information processing systems*, vol. 32, 2019.
- [17] Q. Xie, Z. Yang, Z. Wang, and A. Minca, “Learning while playing in mean-field games: Convergence and optimality,” in *Proceedings of the 38th International Conference on Machine Learning (M. Meila and T. Zhang, eds.)*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 11436–11447, PMLR, 18–24 Jul 2021.
- [18] M. A. U. Zaman, A. Koppel, S. Bhatt, and T. Basar, “Oracle-free reinforcement learning in mean-field games along a single sample path,” in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (F. Ruiz, J. Dy, and J.-W. van de Meent, eds.)*, vol. 206 of *Proceedings of Machine Learning Research*, pp. 10178–10206, PMLR, 25–27 Apr 2023.
- [19] W. U. Mondal, V. Aggarwal, and S. Ukkusuri, “Mean-field control based approximation of multi-agent reinforcement learning in presence of a non-decomposable shared global state,” *Transactions on Machine Learning Research*, 2023.
- [20] N. Saldi, T. Basar, and M. Raginsky, “Markov-Nash equilibria in mean-field games with discounted cost,” *SIAM Journal on Control and Optimization*, vol. 56, no. 6, pp. 4256–4287, 2018.
- [21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [22] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, PMLR, 2015.
- [23] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, PMLR, 2018.
- [24] C. Feng and A. L. Liu, “Peer-to-peer energy trading of solar and energy storage: A networked multiagent reinforcement learning approach,” *Applied Energy*, vol. 383, p. 125283, 2025.
- [25] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, “Grid structural characteristics as validation criteria for synthetic networks,” *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3258–3265, 2017.
- [26] Hawaiian Electric, “Power facts,” 3 2024.
- [27] D. Krishnamurthy, W. Li, and L. Tesfatsion, “An 8-zone test system based on ISO New England data: Development and application,” *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 234–246, 2016.
- [28] R. Tidball, J. Bluestein, N. Rodriguez, and S. Knoke, “Cost and performance assumptions for modeling electricity generation technologies,” tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States), 2010.
- [29] A. Robinson, “Solar PV Analysis of Honolulu, United States,” 2024.
- [30] D. Argüeso and S. Businger, “Wind power characteristics of oahu, hawaii,” *Renewable Energy*, vol. 128, pp. 324–336, 2018.
- [31] M. Coffman, P. Bernstein, S. Wee, and A. Arik, “Estimating the opportunity for load-shifting in Hawaii,” https://uhero.hawaii.edu/REPEc/hae/wpaper/WP_2016-10.pdf, 2016.
- [32] M. Roozbehani, M. A. Dahleh, and S. K. Mitter, “Volatility of power grids under real-time pricing,” *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 1926–1940, 2012.

I. REVIEW #265A

A. Paper summary

The paper proposes a mean field control approach for multiple aggregators coordinating consumer energy storage devices responding to locational marginal prices. A reinforcement learning approach leverages LMPs to help the system converge to equilibrium.

B. Comments for authors

The paper focuses on the case of multiple aggregators that affect the electricity price and behave unstrategically to coordinate distributed storage to reduce customer energy costs. The approach is interesting in how it combines mean field control and reinforcement learning.

Here are my major comments:

Mean field games have computational advantages but also limitations in that theoretical and analytical results often rely on unrealistic assumptions. It would be nice if the paper included more discussion of these limitations, and perhaps even an explicit list of the assumptions needed to make the approach work.

The paper seems to mix (or not be clear about) its use of energy vs power variables. For example q is the ratio of demand to DER capacity (in energy) so its units are kW/kWh? Here, DERs (storage) are modeled by their energy capacity alone; however, in practice they also have physical power capacities, which seem to be neglected here? The storage action space (charge/discharge) is in terms of energy (i.e., the proportion of total storage energy to charge/discharge) but power markets use power (or equivalent total energy consumed/produced over a fixed time period), not energy to schedule/dispatch resources. Why use storage energy vs storage power (charging/discharging power) as the state/action variable? Does using energy instead complicate the formulation?

The DHA approach shown in the results seems to appear out of nowhere. It hasn't been mentioned that the proposed approach will be compared to a decentralized approach before it is shown in the figures. The reader is referred to the appendix for a description of it. Does it need to be included in this paper at all?

In addition to the cost for consumers and prosumers, it would also be interesting to look at the (change in) profit for the generators, and the overall change in social welfare.

The economic dispatch approach uses PTDFs. It is not clear to me why these are used in this setting vs just using the DC optimal power flow formulation.

C. Summarize your recommendation in one to two sentences

Overall this is an interesting paper that will likely lead to good discussion at the conference; however, some points should be clarified/corrected.

— Johanna Mathieu

II. REVIEW #265B

A. Paper summary

This paper addresses the assumption that DER aggregators are price takers, when in fact DERs can influence market outcomes in aggregate. It does so by building a hybrid mean field equilibrium (MFE) framework that models a DER aggregation in its "infinite agent" limit. The paper also introduces a two phase RL algorithm for aggregators to learn policies.

B. Comments for authors

One of the key points of the model setup is the regularization of the reward function, which makes it strongly concave, which eventually enables the proof of existence and uniqueness of the MFE. In my opinion this critical parameter is insufficiently motivated in practical terms. What additional cost does it represent for agents/aggregators, and do they actually pay this cost or just include it in their algorithmic process. If they do not pay it, then how would one create incentives to keep the regularization in their algorithm? I also believe the paper would be much stronger with an exploration of the implications of the size of α (the strength of the regularization function, which is not actually specified in the numerical experiments section) on the performance of the algorithm. (There is another α in appendix D, but this is a different parameter.)

Beyond this concern, I find that the paper is very well written, rigorous, and a valuable contribution to the field. As with other MFG research, the notation is rather cumbersome but I appreciate the effort the authors have taken to make it as manageable as possible. The numerical experiments are convincing of the value of the algorithm for reducing price volatility; I appreciate that the authors ran multiple experiments and used a real world system.

C. Summarize your recommendation in one to two sentences

This paper makes advances in the area of DER aggregation and control using rigorous mean field game and reinforcement learning methods. Results require an important assumption about regularization that merits further investigation.

— *Duncan Callaway*

III. REVIEW #265C

A. Paper summary

The paper presents a framework for integrating DERs into the wholesale electricity markets through a hybrid mean-field control and game-theoretic approach, and supported by RL. The results are demonstrated on the Oahu Power System.

B. Comments for authors

The paper is well-written and easy to follow, and presents a scalable method for energy aggregators to optimize storage under uncertainty. A few comments

- 1) Though the authors consider PTDF, there is no consideration of voltage or reactive power limits, which will play a role in a full-blown ACPF setting. Can the authors comment on how this affects the reliability?
- 2) Have the authors considered/studied the effects of communication delays or noise between the ISO and the aggregator on updating the belief (via LMPs)? If so, what is the impact?
- 3) Theorem 3 relies on the constant $(L_1 L_{MF} L_3 + L_2) < 1$. Can the authors elaborate on when this relation may not hold?

C. Summarize your recommendation in one to two sentences

The paper presents a scalable framework for integrating decentralized DERs in wholesale markets via aggregators. It will be good contribution to the conference.

— *Bala Kameshwar Poola*

IV. BRIEF RESPONSE TO REVIEWERS (OPTIONAL)

We thank the reviewers for their thoughtful and constructive feedback. The comments helped us improve the final version. In particular, we revised the paper to clarify the role of aggregators, the interpretation of energy and power quantities, the use of entropy regularization, and the assumptions behind the mean-field equilibrium result. We also strengthened the literature positioning relative to prior mean-field control work in power systems and added standard references for the reinforcement learning algorithms discussed in the paper.

Consistent with the final submission instructions, we limited the revision to clarifying edits rather than adding new experiments or substantially restructuring the paper. Suggestions that require broader extensions, such as additional welfare metrics, communication delay analysis, or more detailed distribution network modeling, are valuable directions for future work.

V. BRIEF SUMMARY OF CHANGES MADE TO THE FINAL PAPER

We made several targeted clarifying edits in the final version. First, we revised the abstract and model description to clarify that aggregators are modeled as non-strategic, price-responsive learners, rather than strict price takers, while LMPs remain endogenous through aggregate prosumer actions and wholesale market clearing.

Second, we expanded the literature review to better position the paper relative to prior mean field control work on flexible loads and demand dispatch, including work by Kizilkale and Malhamé and by Chen, Bušić, and Meyn. We clarified that, unlike these target tracking and demand dispatch settings, our setting studies aggregators that learn from LMPs formed endogenously through wholesale market clearing.

Third, we clarified the energy versus power interpretation of the model. Net load and storage actions are defined over fixed market intervals, and the corresponding average charging or discharging power can be recovered by dividing the interval energy quantity by the interval length. We also noted how explicit storage power limits can be incorporated through the feasible action set.

Fourth, we revised the notation for entropy regularization, specified the regularization coefficient used in the numerical experiments, and added standard references for PPO, TRPO, and SAC.

Finally, we added discussion of distribution network constraints, clarified the use of the PTDF formulation as a DC power flow representation, and added a remark explaining when the sufficient contraction condition in Theorem 3 may fail.