Harder *et al. Energy Informatics* 2023, **6**(Suppl 1):41
https://doi.org/10.1186/s42162-023-00293-0

Energy Informatics

RESEARCH Open Access

Finding individual strategies for storage units in electricity market models using deep reinforcement learning

Nick Harder[1*], Anke Weidlich[1] and Philipp Staudt[2]

From The 12th DACH+ Conference on Energy Informatics 2023
Vienna, Austria. 4-6 October 2023. https://www.energy-informatics2023.org/

*Correspondence:
nick.harder@inatech.uni-freiburg.de

[1] Institute for Sustainable Systems Engineering, University of Freiburg, Freiburg, Germany
[2] Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Abstract

Modeling energy storage units realistically is challenging as their decision-making is not governed by a marginal cost pricing strategy but relies on expected electricity prices. Existing electricity market models often use centralized rule-based bidding or global optimization approaches, which may not accurately capture the competitive behavior of market participants. To address this issue, we present a novel method using multi-agent deep reinforcement learning to model individual strategies in electricity market models. We demonstrate the practical applicability of our approach using a detailed model of the German wholesale electricity market with a complete fleet of pumped hydro energy storage units represented as learning agents. We compare the results to widely used modeling approaches and demonstrate that the proposed method performs well and can accurately represent the competitive behavior of market participants. To understand the benefits of using reinforcement learning, we analyze overall profits, aggregated dispatch, and individual behavior of energy storage units. The proposed method can improve the accuracy and realism of electricity market modeling and help policymakers make informed decisions for future market designs and policies.

Keywords: Agent-based modeling, Electricity markets, Energy storage, Multi-agent reinforcement learnin, Reinforcement learning

Introduction

With the increasing share of intermittent renewable generation in energy systems, the complexity of electricity markets and the number of market participants have increased. Such increase, in turn, presents a new challenge to energy system modelers as more actors and more complex market interactions need to be modeled, and their decision-making needs to be encoded (Hache and Palle 2019). This decision-making modeling is especially challenging for energy storage units, as storage operators need to anticipate how the price of electricity will behave and will act according to market expectations (McConnell et al. 2015).

The challenge of behavior representation is often approached in two distinct ways: global optimization or agent-based simulation with rule-based decision-making (for details, see "Related work").

The optimization approach usually assumes some form of foresight or even perfect foresight, which is unattainable in practice. Even stochastic optimization, generally based on a probability distribution of events, fails to explain individual behavior as it either assumes a common expectation of the probability of events or becomes complex very quickly and is similarly based on a broad set of assumptions. These models help derive general statements about the system's state or needed capacity expansion but need improvement to derive or explain the actual strategies of operators. On the other hand, the rule-based approach typically uses one single, often heuristic strategy for all actors of a specific type. While such a strategy might be (close to) optimal for the individual that does not influence the price with its behavior, it might no longer be optimal in the aggregate case if the marketed capacity can affect the price.

A promising approach to this modeling challenge is Multi-agent deep reinforcement learning (MADRL) (Gronauer and Diepold 2022). Algorithms from this class allow us to model competing agents that adapt to the market and the strategies of other agents. This way, changing circumstances in the market, such as the entry of additional agents, are incorporated into the strategy design. At the same time, this class of algorithms only requires the design of a feedback loop with the environment, reducing the necessary heuristic or empirical assumptions for strategy development. One possible reason why MADRL is currently not applied on a large scale to energy market models is the challenge of training in a high-dimensional environment with many simultaneous learning agents. Such simulations are highly unstable and rarely converge, making these models intractable for large networks with many agents.

Therefore, this paper presents the design of an MADRL algorithm that can be applied to large energy system models and demonstrate its applicability. We use the multi-agent variation of the Twin Delayed Deep Deterministic Policy Gradients (TD3) algorithm (Fujimoto et al. 2018a) following the centralized training and decentralized execution paradigm. We use a specific state design to reduce the overall complexity while increasing the scalability of the learning method. In addition, we demonstrate that such representation can simulate competing strategic behavior in realistic (complex) market environments. As a case study, we use a high-resolution German wholesale electricity market model to demonstrate how energy storage units adapt their strategies when they compete and show how they pursue specific niches with their strategy. We assume the energy storage units as profit-seeking market participants and compare the results to widely used modeling approaches, i.e., unit commitment optimization and rule-based heuristic strategies. The proposed method will be helpful to researchers in refining electricity market models and better understanding decision-making in energy markets. Such improvements would increase the realism of electricity market models and help stakeholders make better decisions, benefiting both consumers and energy providers.

Table 1 Overview of existing electricity market models

Model family	Behavior modeling method	Use case	Limitations	Examples
Optimization-based	Optimization	System expansion Investment planning	System optimal behavior Assumption of perfect foresight	ELMOD (Leuthold et al. 2012) PyPSA (Brown et al. 2018) Oemof (Hilpert et al. 2018)
Equilibrium-based	Game theory	Emerging strategies Emerging market dynamics	Mathematically challenging Computationally expensive Inflexible to market changes	Orgaz et al. (2017) Liu and Conejo (2021) Huppmann and Egging (2014)
	Static rule-based	Static market dynamics Market results	No strategy adaptation to changing environment	flexABLE (Qussous et al. 2022) AMIRIS (Deissenroth et al. 2017)
Agent-based	Classic RL	Emerging strategies Emerging market dynamics	Limited performance Discrete action space No multi-agent learning	AMES (Tesfatsion and Battula 2020) Weidlich (2008) MASCEM (Pinto et al. 2014)
	Novel DRL	Emerging strategies Emerging market dynamics	Computationally expensive Only applied to small cases	Du et al. (2021) Ye et al. (2019)

Related work

To understand the challenges of modeling electricity markets and representing the behavior of their participants, we first give a short overview of existing approaches, including those using Reinforcement learning (RL). Three main types of electricity market models exist: optimization-based, equilibrium-based, and simulation-based (Ventosa et al. 2005). In the following, we discuss these types with model examples and the representation of energy storage in these models. We also look closer at models utilizing RL and Deep reinforcement learning (DRL). Table 1 briefly summarizes existing electricity market models, their use cases, and their main limitations.

Optimization-based models are usually constructed using linear or mixed-integer linear programming with various technical constraints (Ventosa et al. 2005). These models use a specific objective function, usually defined as welfare maximization (Leuthold et al. 2012), system cost minimization, investment cost minimization, or a combination of these (Brown et al. 2018). This family of models is best suited when the research question involves long-term investment or short-term optimal dispatch planning. Such models assume a perfect market and perfect competition between market participants. These assumptions make them unsuitable when the research question focuses on market design abuse, particular agents' behavior, or detailed strategic market analysis.

Open-source optimization-based electricity market models such as ELMOD (Leuthold et al. 2012), PyPSA (Brown et al. 2018), Oemof (Hilpert et al. 2018), and PowerFlex (Koch et al. 2015) are some of the best-known examples. These models have different architectures and purposes but commonly solve an optimization problem with an

operational cost minimization objective. They include technical parameters of power plants, storage units, and transmission networks in the form of optimization constraints and rely on short-run marginal and long-term investment costs for optimization. Storage units in these models are modeled as pure network supplements and are dispatched to minimize total system cost. Such models fail to reflect imperfect competition and individually optimal storage behavior and typically assume perfect foresight.

Equilibrium-based models also use optimization techniques but rely on a two-level problem formulation, where the lower-level problem is specific to each market participant, for example, profit maximization (Boland 2017). An objective function, similar to the optimization-based models, is used for the upper-level problem (Boland 2017). The main purpose of such models is to investigate the market equilibrium and emerging market dynamics and analyze the behavior of market participants.

Examples of such models include (Orgaz et al. 2017), where the authors analyze the power exerted by different market players, and (Liu and Conejo 2021), which indicates that equilibrium models can help to comprehend market participants' behavior better. Another example is (Huppmann and Egging 2014), which accounts for the global energy market, including fuel substitution and power markets. Unfortunately, none of these previous studies include energy storage units. According to Böhringer and Rutherford (2006), the high mathematical and computational complexity of equilibrium-based models makes them difficult to use for large market simulations. As suggested by Niu (2005), these drawbacks lead to the creation of hybrid models, where the market is modeled using a simulation model, while individual agents can still use optimization methods for their bidding.

Simulation-based models and their most prominent subcategory of agent-based models use individual bidding strategies for the market participants. Such strategies are typically defined using manually derived rule-based or optimization-based algorithms (Bonabeau 2002). Models from this family, which employ fixed bidding strategies, are primarily used to investigate market dynamics. If adaptive bidding strategies are implemented using learning techniques, these models also allow for analyzing emerging market dynamics (Bonabeau 2002).

Exemplary open-source agent-based models for electricity markets, such as AMIRIS Deissenroth et al. (2017) and flexABLE Qussous et al. (2022), use a fixed set of manually derived non-linear rules as bidding policies for both power plants and energy storage units. Such fixed policies are typically fine-tuned to produce results in line with historical data. Due to the absence of adaptation capabilities by the agents, they are not intended to analyze emerging dynamics in future electricity markets. In addition, such rules usually depend on an external signal such as a price forecast (Qussous et al. 2022), which leads to almost identical behavior of homogeneous agents, for example, simultaneous charging of all storage units when the forecast price is below a certain threshold. Therefore, accurate analysis of emerging market dynamics of competitive storage behavior cannot be performed with such models.

Using RL algorithms in electricity market simulation has been the subject of extensive research (Weidlich 2008). The AMES framework (Tesfatsion and Battula 2020) is an example of a market simulation framework that implements learning using an Erev-Roth RL algorithm (Erev and Roth 1998). Another example is the MASCEM

model (Pinto et al. 2014), which employs value iteration and an Erev-Roth learning algorithm. While these models enable learning, the built-in RL algorithms only allow for a small discrete set of actions. However, recent advances in RL, particularly in DRL, can represent more complex behavior and promise better performance and stability levels. Moreover, DRL models support simultaneous multi-agent learning and can be used to train storage units with their complex market strategies.

Numerous studies have focused on using RL and DRL to develop bidding strategies for storage units in electricity markets, but only for a single agent. For instance, Wang and Zhang (2018) derive a bidding policy for an energy storage unit bidding on a real-time market using a Q-Learning algorithm (Watkins and Dayan 1992). Meanwhile, Dong et al. (2021) use a function approximation-based RL algorithm to develop a bidding strategy for a battery storage unit on a day-ahead and control reserve market. Similarly, Anwar et al. (2022) concentrate on creating a bidding strategy for an energy storage unit on both energy and capacity markets. They use a Proximal Policy Optimization algorithm (Schulman et al. 2017) and demonstrate excellent performance. Verdaasdonk et al. (2022) employ a TD3 algorithm (Fujimoto et al. 2018a) for bidding of storage units on the continuous intra-day market and compare the results with an intrinsic rolling method. Although all these studies demonstrate a comparable performance of DRL algorithms, they assume that storage units are small enough not to influence the market price and therefore are modeled as price takers. This price-taking assumption is likely no longer true when the storage capacity and the bid size are considerable. Applying the learned policy from a single agent to multiple similar agents in an electricity market model would result in a behavior similar to manually-derived rules, where all storage units act alike, which is not advisable, as we demonstrate in this study.

Furthermore, applying single-agent RL algorithms to multiple simultaneous learning agents is not possible due to the actions of other agents and the impact of their actions on the environment. For instance, their effect on the market clearing price makes the environment non-stationary, which violates the Markov property. This property is required for the convergence guarantees of single-agent RL algorithms (Hernandez-Leal et al. 2017). As suggested in Cao et al. (2020), an MADRL approach using centralized training and a decentralized execution approach can mitigate this issue and enable simultaneous learning of multiple agents. Such an approach is utilized in Du et al. (2021), where the authors develop an MADRL approach using a multi-agent variation of the Deep Deterministic Policy Gradient (DDPG) algorithm (Lillicrap et al. 2015), which was first presented in Lowe et al. (2017) to simulate an electricity market. The authors study a market setup with nine agents, where three are learning agents. The authors conclude that MADRL can approximate a Nash equilibrium on the market and that the agents can exploit the market during grid congestion. Similar results are presented in Ye et al. (2019), where the MADRL approach with ten learning agents using a version of the policy gradient algorithm simulates a simplistic electricity market. The results also closely approximate the results of an equilibrium model. In both studies, the authors focus only on power plants and do not consider energy storage units. Additionally, the considered test cases by the authors are small, and no further studies on the application of MADRL techniques for large-scale market simulations were found.

In conclusion, electricity market models are based on different modeling approaches. Many of these models consider energy storage units, whose behavior is modeled either through a centralized optimization or manually derived bidding heuristics. The optimization approach does not consider the strategic behavior of agents, and rule-based approaches do not adapt to the market environment and depend on an external signal, thus resulting in similar (and sometimes unprofitable) behavior for all agents. Applying DRL can generate adapting emergent bidding strategies. However, single-agent DRL algorithms, which have been extensively studied, cannot be directly used in a multi-agent environment. Using MADRL approaches for modeling agents on electricity markets is currently limited in the number of learning agents and, in the past, has not included storage units. This study closes this gap and provides an MADRL method applicable to large-scale electricity market simulations. This method can help derive individual bidding strategies for multiple energy storage units simultaneously and is thus well-suited for electricity market modeling.

Methodology

This section presents a detailed overview of the proposed MADRL method for modeling storage units in a complex electricity market environment. It starts with a short introduction to general RL concepts and some details on MADRL approaches in Section Introduction to RL and MADRL. Section Detailed model architecture provides a detailed description of the algorithm and modeling framework, with the description of the used state, action, and reward designs stated in Section State, action and rewards. Finally, Section Conventional modeling baselines presents the applied modeling approaches used as comparison baselines of the proposed method.

Introduction to RL and MADRL

RL algorithms generally aim to develop a policy for Markov decision process (MDP), mainly finite MDPs, where state and action spaces are finite. A finite MDP is defined using a state space S, an action space A, and a one-step transition probability function of the environment. The transition probability function $\mathcal{P}$ defines the probability of transitioning to each next state $s_{t+1} \in S$ given the current state $s_t \in S$ and action $a_t \in A$ at time-step $t \in T$.

An agent chooses actions according to some policy π, which can be deterministic or stochastic. After each transition, the agent receives a reward $r_t = \mathcal{R}(s_t, a_t, s_{t+1})$, where $\mathcal{R}$ is the reward function. The main target of an agent is to maximize the total reward $R = \sum_{t=0}^{t=T} r_t$. Given the transition probabilities we can derive the expected return $\mathcal{J}$ for policy π:

$$\mathcal{J}(\pi) = \int_t^T \mathcal{P}(s_{t+1}|s_t, a_t)\mathcal{R}(s_t, a_t, s_{t+1}) = \mathbb{E}_\pi\{R\} \tag{1}$$

From Eq. 1 we can derive an action-value function $Q^\pi(s, a)$, based on which we can calculate the Q-value:

$$Q^\pi(s, a) = \mathbb{E}_\pi[R|s_0 = s, a_0 = a] \tag{2}$$

The Q-value represents the expected return starting from state s_0, taking action a_0 and following the policy $\pi(s)$ afterward. An optimal action-value function $Q^*(s,a)$ gives the expected return if taking action a in state s and acting according to an optimal policy afterward. Knowing the optimal action-value function $Q^*(s,a)$, we can derive the optimal policy $\pi^*(s)$, which chooses action $a^*(s)$ that maximizes the expected return $\pi^*(s) = \arg\max_a Q^*(s,a)$. The technique of learning the underlying but unknown action-value functions is the main idea behind the *Q-Learning* family of RL algorithms with its most well-known examples the *Q-Learning* (Watkins and Dayan 1992) and the *Deep Q-Networks* (Mnih et al. 2013).

A different approach is to explicitly learn the policy as $\pi(a|s,\theta)$, where θ is a set of parameters. The policy can then directly determine action a when in state s without referring to the action-value function. The parameters θ can be optimized using gradient ascent on some performance measure such as $\mathcal{J}(\theta)$ from Eq. 1 as follows:

$$\theta_{t+1} = \theta_t + \alpha \nabla \widehat{\mathcal{J}(\theta_t)} \tag{3}$$

Here, α is the learning rate, and $\nabla \widehat{\mathcal{J}(\theta_t)}$ is the estimate of the gradient of the performance measure with respect to θ. Due to the gradient procedure, this family of RL algorithms is referred to as *policy gradient methods*. The most well-known examples are *A2C* by Mnih et al. (2016) and *PPO* by Schulman et al. (2017).

It is also possible to combine the strengths of both families, in which the action-value function is used as a baseline in policy gradient methods. In this case, the action-value function serves as a *critic* to the actions chosen by the policy. The policy is termed an *actor*, and such methods are referred to as *actor-critic* methods. Most such algorithms are trained off-policy, similar to Q-Learning, which improves the sample efficiency and optimizes the policy directly, improving stability. Examples of such algorithms include DDPG (Lillicrap et al. 2019), its successor TD3 (Fujimoto et al. 2018b) and *Soft Actor-Critic* (Haarnoja et al. 2018).

The algorithms mentioned above provide stable performance, which can only be guaranteed for single-agent setups. In multi-agent cases, the system's non-stationarity caused by the actors' changing actions causes instability. Such non-stationarity violates the Markov property, which states that the future state must depend only on the current state and action of the agent. As a result, memory-based algorithms such as Deep Q-Networks or TD3 are no longer applicable, and convergence guarantees are no longer valid (Hernandez-Leal et al. 2017). Therefore, we have to apply specific algorithms from the family of multi-agent DRL.

The issue of non-stationarity can be overcome by one of the following approaches: improved experience replay buffer, parameter sharing, and centralized training and decentralized execution (Cao et al. 2020). In our work, we use the centralized training and decentralized execution approach, which is compatible with actor-critic algorithms, and has an existing research foundation. When using this approach, during the training phase, each agent receives information about the states and actions of other agents, making the environment stationery. While access to information about other market participants is unrealistic for real-life applications, it can be successfully used for market modeling. Alternatively, an estimation of the states of other market participants could be used, representing a more realistic approach.

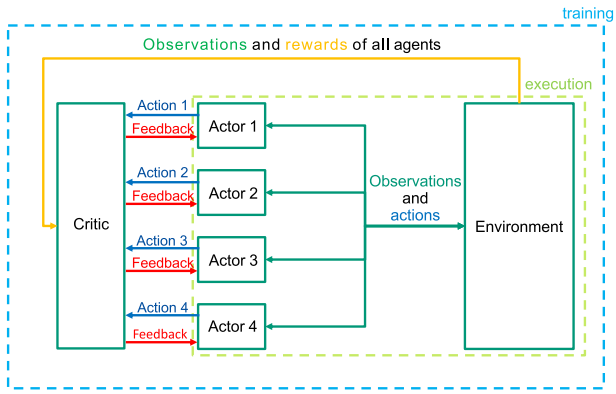


Fig. 1 Graph of centralized training and decentralized execution algorithms

In Fig. 1, a structure of a Multi-agent deep Deterministic Policy Gradient (MADDPG) algorithm (Lowe et al. 2017), utilizing the idea of centralized training and decentralized execution, is presented. It is built to extend the actor-critic DDPG algorithm. In this algorithm, a centralized critic receives the observations and actions of all agents and provides informed feedback to the actors during the training phase. This approach allows the actors to learn policies based on local information and no longer rely on the centralized critic during the execution phase. It is also possible to define an individual centralized critic per agent, which increases the performance at the cost of computational complexity (Lowe et al. 2017). While this approach overcomes the non-stationarity issue, it suffers from the dimensionality curse. With many agents, the centralized critic needs to deal with a high number of state and action values from all agents. Such a high number of inputs can cause stability issues and might be one reason only a few simultaneous learning agents have been used for electricity market modeling until now (Du et al. 2021; Ye et al. 2019).

Detailed model architecture

In this study, we rely on the framework of the Markov game (Littman 1994) — a generalization of finite MDPs to accommodate multiple interacting agents. It is represented by a finite number of agents N, a state space S, action spaces A_i for each agent $i \in [1, N]$ and a transition probability function $\mathcal{P}$:

$$\mathcal{P}(s_{t+1}|s_t, a_{t,1}, ..., a_{t,N}) = Pr\{s_{t+1}|s_t, a_{t,1}, ..., a_{t,N}\} \tag{4}$$

Compared to a single-agent setup, the state transition in the Markov game depends on the actions of all N agents. At time-step t, all agents observe state s_t and simultaneously decide on an action $a_{t,i}$. These actions form an action profile $\mathcal{A}_t$, which affects the transition to the state s_{t+1}. After the transition, each agent i receives a reward $r_{t,i}$ according to reward function $\mathcal{R}$. The reward function can be global $\mathcal{R}$ or specific to each agent $\mathcal{R}_i$. In real applications, agents typically cannot observe the complete state of the system. Therefore, we use the partially observable MDP setup to model the electricity market. Instead of the entire state s_t, each agent receives a partial observation dependent on the complete state $o_{t,i} \sim s_t$ where $o \in S$. Each agent i aims to find the optimal policy $\pi_i^*(o)$ that maximizes its total reward R_i.

We utilize the Multi-agent twin Delayed Deep Deterministic Policy Gradients (MATD3) algorithm, a variation of the MADDPG algorithm (Lowe et al. 2017) with several modifications from the TD3 algorithm (Fujimoto et al. 2018b). For each agent i, we define two centralized critic neural networks (NNs) $Q^i_{\theta_{1,2}}$, two target critic NNs $Q^i_{\theta'_{1,2}}$, an actor NN π^i_ϕ and an actor target NN $\pi^i_{\phi'}$. Here, θ, θ' and ϕ, ϕ' are corresponding weights of underlying NNs. For simplicity, we discard the agent i notion in the following definitions since the same operations are performed for each agent.

The critic networks $j = 1, 2$ are updated using gradient descent on the loss function $L(\theta)$ defined by Eq. 5.

$$L(\theta_j) = \frac{1}{B}\sum_{k=1}^{B}[y_k - Q_{\theta_j}(\mathcal{O}_k, \mathcal{A}_k)]^2 \tag{5}$$

In Eq. 5, B is the size of the minibatch, y is the target value defined by Eq. 6, $\mathcal{O}$ and $\mathcal{A}$ are a collection of all observations and actions during the transition k in the minibatch.

$$y_k = r_k + \gamma \min_{j=1,2} Q_{\theta'_j}(\mathcal{O}'_k, \tilde{\mathcal{A}}'_k) \tag{6}$$

Here, γ is the reward discount factor and $\mathcal{O}'$ is the collection of the next observations during the transition k, and $\tilde{\mathcal{A}}'$ is the collection of target actions. The target action for each agent is defined as $\tilde{a}_k = \pi_{\phi'}(o'_k) + \epsilon$, where $\epsilon = \text{clip}(\mathcal{N}(0, \sigma), -c, c)$ is a clipped Gaussian noise with σ and c being hyper-parameters.

The actor-network parameters ϕ are updated using gradient ascent in the direction of actions maximizing the action-value function and is defined by Eq. 7. Here, $\mathcal{A}_k^{\mathcal{A}\setminus[i]}$ is the collection of the next actions excluding the action of agent i, whose actor-network is being updated.

$$\nabla_\phi J(\phi) = \frac{1}{B}\sum_{k=1}^{B} \nabla_\phi \pi_\phi(o_k) \nabla_a Q_{\theta_1}\left(\mathcal{O}_k, \mathcal{A}_k^{\mathcal{A}\setminus[i]}, \pi_{\phi_i}(o_k)\right) \tag{7}$$

The target critic parameters $\theta'_{1,2}$ and target actor parameters ϕ' are updated in the direction of critic parameters $\theta_{1,2}$ and actor parameters ϕ using the soft-update given by Eq. 8 only every d updates to improve training stability. In Eq. 8, τ is a hyper-parameter between 0 and 1.

$$\theta'_{1,2} = (1-\tau)\theta'_{1,2} + \tau\theta_{1,2} \tag{8a}$$

$$\phi' = (1-\tau)\phi' + \tau\phi \tag{8b}$$

For the market simulation, we use the flexABLE model developed by Künzel (2019) and further extended in Qussous et al. (2022). It is an agent-based market simulation model featuring heuristic bidding strategies for conventional power plants and storage units and several electricity market implementations, including the energy only market, which is an abstraction of day-ahead and intraday markets. This framework has been validated and produces results in line with historical data (Qussous et al. 2022).

The code and model inputs are published along the paper[1]. The complete flow of the algorithm, the architecture, hyper-parameters for NNs, and training parameters of the MATD3 algorithm are provided as supplementary material along the code. The model is implemented in Python, with the learning algorithms also implemented using Python and PyTorch (Paszke et al. 2019). The training was performed on a workstation with an i9-13700K CPU, 128GB of RAM, and an NVidia RTX 3090 24GB GPU.

State, action and rewards

One of the main challenges of using MADRL with a centralized critic approach is the dimensionality curse, as indicated by Lowe et al. (2017). The dimensionality curse refers to the fact that as the size of the inputs to the neural networks grows, learning stability suffers, leading to instabilities. These instabilities can limit the number of simultaneous learning agents and thus the applicability for large-scale models. To overcome this issue, we have engineered a set of observations common to all agents, which does not depend on the number of agents and is sufficient for good learning performance. This common state and the actions of all learning agents are shared with the centralized critic during the training phase. To further improve the agents' performance, we have derived a small set of additional observations specific to each agent, which is not shared with the central critic and is only used by the actor. While this decision may appear to violate centralized training and introduce non-stationarity, it does not. Non-stationarity primarily stems from changing bidding strategies and resulting market clearing prices, which affect agent rewards. However, as the central critic receives agents' actions, it can account for these changes in rewards, ensuring that non-stationarity is addressed. During the development phase, we tried both approaches and did not observe any decrease in performance with the smaller state. This design allows us to perform simulations with a large number of simultaneous learning agents (tested with up to 140 learning agents).

At each time-step, t, we construct a global observation o_t^{global} available to all agents, which consists of past $[L_{t-N}^{\text{h}} : L_t^{\text{h}}]$ and forecast $[L_t^{\text{f}} : L_{t+N}^{\text{f}}]$ residual load time series and past $[M_{t-N}^{\text{h}} : M_t^{\text{h}}]$ and forecast $[M_t^{\text{f}} : M_{t+n}^{\text{f}}]$ market clearing price time series. The past and forecast values are provided for $N = 24$ steps. In addition to the global state, each energy storage agent receives its past state of charge $[SOC_{t-6} : SOC_t]$ for the last six time steps, and its current energy charge cost ec_t, defined using Eq. 9.

$$ec_{t+1} = \frac{ec_t SOC_t - P_t M_t \Delta t}{SOC_{t+1}} \tag{9}$$

Here, P_t is the power value of the storage unit at time-step t in MW, which is positive when charging and negative when discharging; Δt is the market time-step, used to convert capacity to energy; the initialization ec_0 of this iterative value is defined as $ec_0 = M_0$, assuming that the initial energy content was purchased at the market clearing price at initialization M_0.

Only the global observation is passed to the centralized critics. As this global observation does not depend on the number of agents, the critic input does not grow with the

[1] https://github.com/INATECH-CIG/storage-flexRL.git

number of agents, which improves stability and enables scalability. Also, as the global observation information is available to all market participants in real applications, such a setup closely approximates real electricity markets. All inputs are normalized to improve stability further.

In our setup, an action of a storage unit at time-step t is defined as $a_t = (ep_t, B_t^{\text{dir}})$, where ep_t is the bid price and B_t^{dir} is the direction of the bid (charging/discharging). This action is then transformed into a market bid $B_t = (P_t, ep_t)$, where P_t is the power of the bid in MW and ep_t is the bid price in EUR/MW. P_t is defined using Eq. 10. The power value is positive when it is a supply bid and negative when it is a demand bid. $P_{\text{ch}}^{\max}, P_{\text{dis}}^{\max}$ represent maximal charging and discharging power and $SOC^{\min}, SOC^{\max}$ — minimal and maximal SOC. Such formulation for the P_t was chosen over a continuous formulation $P_t \in [-P_{\text{ch}}^{\max}, P_{\text{dis}}^{\max}]$ as it proved to produce better results and converge faster.

$$P_t = \begin{cases} \min\left[\frac{(SOC_t - SOC^{\min})}{\Delta t}\eta_{\text{dis}}; P_{\text{dis}}^{\max}\right] & \text{if } B_t^{\text{dir}} \geq 0 \\ -\min[\frac{SOC^{\max} - SOC_t}{\Delta t}/\eta_{\text{ch}}; P_{\text{ch}}^{\max}] & \text{otherwise} \end{cases} \tag{10}$$

As a reward function, we use a straightforward implementation of economic profit or loss of the storage unit, defined using Eq. 11. $P^{\text{conf.sup}}, P^{\text{conf.dem}}$ represent the confirmed supply and demand powers, $vc_{\text{ch}}, vc_{\text{dis}}$ — variable charging and discharging cost and β is a scaling factor for better stability equal $\beta = \left(10P_{\text{dis}}^{\max}\right)^{-1}$.

$$r_t = \beta\left[(P_t^{\text{conf.sup}} - P_t^{\text{conf.dem}})M_t - P_t^{\text{conf.sup}}vc_{\text{dis}} - P_t^{\text{conf.dem}}vc_{\text{ch}}\right]\Delta t \tag{11}$$

The reward function allows for positive and negative values, such that the agent needs to learn to make losses to achieve future profits and compensate for the conversion losses. While this reward function is simplistic, it naturally represents the profits of energy storage units on real electricity markets. It avoids artificial reward functions such as comparison to profit maximization solutions.

During the initial phase of the training (10 episodes), we include a small reward of 0.01 when the agent submits a demand bid when $SOC_t = SOC^{\min}$ and a supply bid when $SOC_t = SOC^{\max}$. It was observed that such additional reward significantly reduces the time required for training.

Conventional modeling baselines

We compare the result of the MADRL to the most common conventional modeling approaches, namely an optimization-based model and an agent-based model, using two types of heuristic bidding strategies. Unfortunately, as mentioned in Section Related work, we found no open-source equilibrium models applicable to our use case. Therefore, we do not compare our results to these types of models.

For the optimization-based model, we use a unit commitment model implemented using the PyPSA framework by Brown et al. (2018). A complete and clear formulation of a unit commitment problem is provided by Conejo and Baringo (2018).

The first type of the heuristic strategy is a rule-based bidding strategy, based on work Weidlich et al. (2018), which was later refined and validated on historical data in

Qussous et al. (2022). This strategy depends on the average past and forecast price in a period $[t-N : t+N]$ and is defined using Eq. 12.

$$ep_t^{\mathrm{av}} = \frac{1}{N}\sum_{j=1}^{N} M_{t-j}^{\mathrm{h}} + \frac{1}{N}\sum_{j=1}^{N} M_{t+j}^{\mathrm{f}} \tag{12}$$

Next, the bid $B_t^{\mathrm{rule}} = (P_t, ep_t)$ is defined, where P_t is calculated as in Eq. 13 and $ep_t = ep_t^{\mathrm{av}}$.

$$P_t = \begin{cases} \min\left[\frac{(SOC_t - SOC^{\min})}{\Delta t}\eta_{\mathrm{dis}}; P_{\mathrm{dis}}^{\max}\right] & \text{if } M_t^{\mathrm{f}} \geq ep_t^{\mathrm{av}}/\eta_{\mathrm{dis}} \\ -\min[\frac{SOC^{\max} - SOC_t}{\Delta t}/\eta_{\mathrm{ch}}; P_{\mathrm{ch}}^{\max}] & \text{if } M_t^{\mathrm{f}} \leq ep_t^{\mathrm{av}}\eta_{\mathrm{ch}} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Such a rule-based strategy translates into a band around the anticipated average market clearing price with a width of η, compensating for the conversion losses when traded at prices below and above this band. The storage units purchase energy when the anticipated price is below the band and sell energy when above the band.

The second type of heuristic strategy is an optimization-based bidding strategy, whose performance was compared to the validated rule-based strategy and demonstrated comparable results regarding unit dispatch and profits for a single energy storage unit. This strategy uses a rolling window optimization approach, where at each time-step t a profit R maximization problem over a given foresight window $N = 48$ using the price forecast M^{f} and current SOC_t as initial SOC_0 is solved. The complete optimization problem is presented in Eq. 14.

$$\max R = \sum_{j=1}^{N}\left[(P_j^{\mathrm{sup}} - P_j^{\mathrm{dem}})M_j^{\mathrm{f}} - P_j^{\mathrm{sup}} vc_{\mathrm{dis}} - P_j^{\mathrm{dem}} vc_{\mathrm{ch}}\right]\Delta t \tag{14a}$$

$$\text{s.t.} \quad 0 \leq P^{\mathrm{sup}} \leq P_{\mathrm{dis}}^{\max}, \tag{14b}$$

$$0 \leq P^{\mathrm{dem}} \leq P_{\mathrm{ch}}^{\max}, \tag{14c}$$

$$SOC^{min} \leq SOC \leq SOC^{max}, \tag{14d}$$

$$SOC_0 = SOC_t, \tag{14e}$$

$$SOC_{j\neq 0} = SOC_{j-1} + P_j^{\mathrm{dem}}\eta_{\mathrm{ch}} - P_j^{\mathrm{sup}}/\eta_{\mathrm{dis}} \tag{14f}$$

After solving the optimization problem, the bid $B_t^{\mathrm{rule}} = (P_t, ep_t)$ is formulated using Eq. 15 to define the P_t, where the demand or supply power from the optimal solution is used as a power bid.

Table 2 Data sources used for input data

Type	Data description	References
Plants	Storage units Power plant characteristics	Giesecke and Mosonyi (2009); Bundesnetzagentur (2021a); Firm (2017); Umwelt Bundesamt (2020)
Prices	Natural gasCO_2 EU-ETS certificatesOther energy carriers	EEX (2019a); EEX (2019b); Statistische Bundesamt (2021)
Net load	VRE generation Inelastic demand Net load forecast	Bundesnetzagentur (2021b); Bundesnetzagentur (2021b); Bundesnetzagentur (2021b)
Exchanges	Scheduled commercial exchanges	Bundesnetzagentur (2021b)

$$P_t = \begin{cases} P_0^{\text{sup}} & \text{if } P_0^{\text{sup}} > 0 \\ -P_0^{\text{dem}} & \text{if } P_0^{\text{dem}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

To define the bid price ep_t, we use Eq. 16, which uses the forecast price M^{f} and the average profit during the optimization horizon.

$$ep_t = \begin{cases} M_t^{\text{f}} - R/\sum_{j=1}^{N} P_j^{\text{sup}} & \text{if } P_0^{\text{sup}} > 0 \\ M_t^{\text{f}} + R/\sum_{j=1}^{N} P_j^{\text{sup}} & \text{if } P_0^{\text{dem}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

This approach ensures that a storage agent would temporarily operate at a loss if the forecast price sequence allows compensating for this loss overall. In our trial, such bid formulation resulted in higher yields than a simple bid of the forecast price.

Use cases

We employ a validated large-scale energy system model for the simulations that accurately represents the German wholesale electricity market, as expounded in Qussous et al. (2022). We utilize the fleet of pumped hydro plants in Germany to depict energy storage units, which we obtained from Giesecke and Mosonyi (2009) and the Federal Network Agency Bundesnetzagentur (2021a). For a comprehensive overview of the storage systems and their geographical distribution, kindly refer to Giesecke and Mosonyi (2009). The bidding strategies of conventional power plants are defined by the flexABLE model Qussous et al. (2022), which are based on the short-run marginal costs and are contingent on the fuel and CO_2 certificate costs. The power plant list and their techno-economic parameters are based on the data provided by the World Electric Power Plants Database (Firm 2017), the German Environmental Agency (Umwelt Bundesamt 2020), and the Federal Network Agency (Bundesnetzagentur 2021a).

During the development phase, we used data from the entire year of 2019. Each episode has a length of 30 days of simulation, and each episode starts on a random day in a year. Our results are based only on the data from March to April 2019 to allow for a more detailed analysis. The Variable renewable energy (VRE) and the inelastic demand time series for Germany, both actual and forecast, are obtained from SMARD (Bundesnetzagentur 2021b). For the bidding zone import and export data, we used time series of total cross-border scheduled commercial exchanges from SMARD (Bundesnetzagentur 2021b). For a complete overview of the data, please refer to Table 2.

Similar to Lillicrap et al. (2019) and as is common in the literature on DRL, we do not distinguish between training and test data sets for performance evaluation. If the proposed method were to be applied to real-life applications, such as developing bidding strategies for unit operators and their energy storage units, a test data set would test the agent's ability to generalize and act in previously unseen situations. However, such test data sets are not required in the given case. Instead, validation procedures that check how well the agents perform compared to theoretical benchmarks are more important.

As described in Section Conventional modeling baselines, both heuristic rule-based and optimization-based bidding strategies rely on price forecasts, and better price forecasts are expected to generate better results. Therefore, we differentiate the simulations using naive and simulated forecasts. A simple merit order model, where all conventional power plants bid their marginal cost, is used to generate the naive price forecast. On the other hand, the simulated price forecast is based on the simulation model used, with included rule-based bidding strategies. Both forecasts are produced in the absence of the storage units. The DRL method used in our approach relies on a state estimator in the form of an actor's neural network, which is responsible for estimating the current state of the environment. Therefore, the input features of the DRL approach are based on the naive forecast, as this state estimator can intrinsically construct its price forecast. In this case, the naive forecast serves more as a system indicator for the state estimator than a direct price forecast. In our model, any agent can act as the price setter, including the energy storage units. While this further increases the complexity of the environment, it better represents reality.

We use two cases to analyze the proposed algorithm's performance, investigate the emerging strategies, and compare them to conventional modeling approaches. The first case, Case 1, examines the proposed algorithm's performance with a single learning agent acting in an environment of high complexity. In contrast, Case 2 demonstrates the results for a multi-agent setup in the same environment.

Case 1 represents a complex market environment based on the German wholesale electricity market. In this case, the complete set of conventional power plants in the German market is simulated (257 units) based on bidding strategies described by Qussous et al. (2022). In addition, the environment features VRE generation, cross-border exchange, and forecast errors. In this case, only one exemplary storage unit with 500 MW of power for charging and discharging and a 5 GWh energy capacity represented as a learning agent is included in the simulation.

Case 2 is identical to Case 1 with one difference—here, the complete fleet of pumped hydro storage units with a total capacity of 6 GW is represented using individual learning agents, resulting in 25 learning agents. This case demonstrates the applicability of the proposed learning and algorithm architecture to a multi-agent context, both in terms of the problem and learning performance. It also allows us to observe the differences in results compared to a single-agent setup. As typical rule-based bidding strategies are designed with the idea of good performance for a single unit, our main goal with this case is to demonstrate that bidding strategies that produce good results for a single agent are not necessarily effective when used in a multi-agent setup.

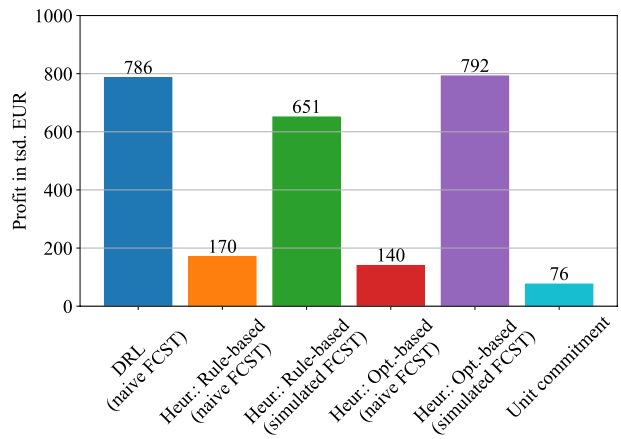


Fig. 2 Total profits in Case 1 using DRL based, heuristic rule-based and optimization-based bidding strategies, and unit commitment model. The results are presented using both naive and simulated forecasts (FCST)

Results

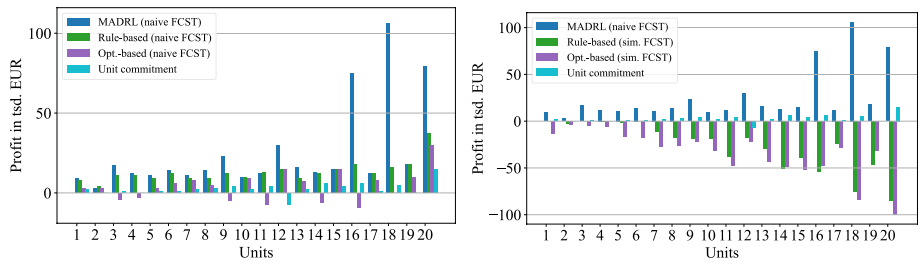
In this section, we present the findings of the case studies outlined in Section Use cases, utilizing the designed MADRL architecture. Specifically, we delve into Case 1 in a single-agent setup in Section Single agent performance while examining Case 2 and its emerging strategies in a multi-agent design in Section Emerging strategies. These results demonstrate the feasibility and applicability of the proposed method while providing valuable insights into its performance compared to heuristic bidding strategies and unit commitment models, which are presented in Section Conventional modeling baselines.

Single agent performance

Figure 2 depicts the total profit of the storage unit during the investigated period from March to April 2019, utilizing different bidding algorithms and forecasts for Case 1. The profits achieved using DRL are slightly below those generated by heuristic bidding strategies employing a simulated forecast. Conversely, the results of the heuristic strategies utilizing a naive forecast are lower but remain positive in all cases. The profits derived from the unit commitment problem are the lowest, which can be explained by noting that unit commitment does not aim to maximize the profits of each unit but rather treats the storage units as system supplements and can operate them at a loss to avoid activating more expensive technologies, which would lower the overall system cost. However, profit-seeking market participants would not operate this way, and higher profits suggest a more likely behavior in practice.

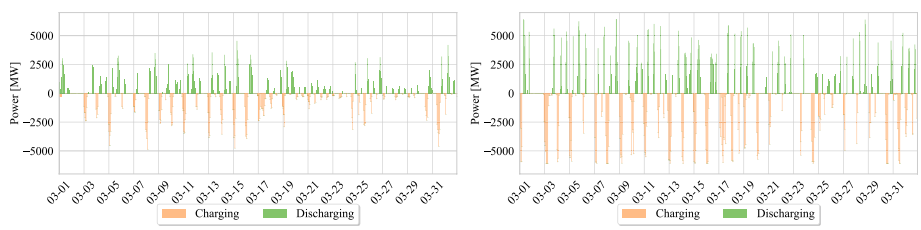
It is worth noting that optimization-based heuristic bidding strategies are not necessarily optimal. They generate a strategy based on a potentially flawed forecast using optimization and thus do not achieve the maximum possible profit, as the agent's actions can influence the market clearing price, which, in turn, can lower the total profit of the unit.

Based on the above results, we conclude that the proposed DRL algorithm produces bidding strategies comparable in performance to validated heuristic bidding strategies.



(a) using naive forecast (b) using simulated forecast

Fig. 3 Total profits in Case 2 using MADRL, heuristic rule-based and optimization-based bidding strategies, and unit commitment model



(a) MADRL (b) heuristic strategy

Fig. 4 Total charge and discharge power in Case 2 during March 2019. In (**a**), MADRL is used for bidding, while in (**b**), a heuristic optimization-based bidding strategy using the simulated forecast is used

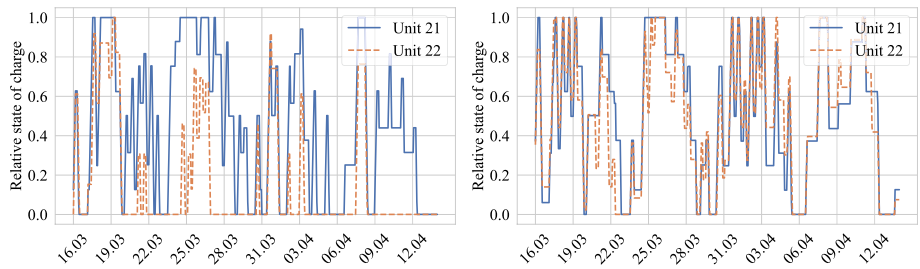
Additionally, it demonstrates that the proposed method can efficiently operate in a complex environment.

Emerging strategies

To analyze the performance and benefits of the proposed MADRL approach in a multi-agent setup, we analyze the total profits of the individual storage units in Case 2. Units 1–20 are presented in Fig. 3. The results for larger units 21–25 are presented in supplementary material, as they demonstrate similar results and do not add to the overall understanding.

When utilizing the naive forecast in Fig. 3a, the profits of units utilizing the rule-based heuristic bidding strategy remain positive. However, the profits of several units utilizing the optimization-based heuristic bidding strategy result in negative returns. In the case of a unit commitment problem, one unit also experiences losses. On the other hand, when implementing the MADRL approach, the profits of individual units are comparable to those of the rule-based bidding strategy for some units but surpass them for most units.

As we have discovered from the previous subsection, a more accurate forecast leads to improved profits for a single agent. However, upon examining Fig. 3b, we observe that this is no longer the case in a multi-agent setup, where a more accurate forecast that ignores the storage behavior results in losses for all units utilizing both heuristic bidding strategies. Nevertheless, the MADRL approach relies on its internal forecast and employs the naive forecast as a system indicator, which allows all units to benefit from profitable outcomes.



(a) Using MADRL

(b) Using optimization-based bidding with simulated forecast

Fig. 5 Relative state of charge of units 21 and 22 using MADRL (**a**) and heuristic optimization-based bidding with simulated forecast (**b**)

To better understand the reasons behind these effects, we present a plot of the aggregate charge and discharge of all storage units in Fig. 4. As can be observed, the optimization-based heuristic bidding strategy utilizing the simulated forecast (Fig. 4b) results in significant supply (positive) and demand (negative) peaks with a magnitude of up to 6 GW. This outcome stems from following a common centralized bidding strategy. Conversely, when implementing MADRL (Fig. 4a), the peaks are noticeably lower, reaching only around 2.5 GW, and the profile is more evenly distributed over time. Such behavior leads to less system disturbance and a more balanced market operation.

Upon closer examination of the emerging storage unit strategies, we have analyzed the behavior of individual units 21 and 22, utilizing their state of charge. These two units possess comparable energy capacity and power characteristics, with unit 21 boasting a maximum discharge power of 350 MW, while unit 22 can discharge up to 440 MW. Their respective capacities are 2 GWh and 3.4 GWh. In Fig. 5, we present the relative state of charge, which is obtained by dividing the current state of charge by the maximum state of charge, for both units using both MADRL and optimization-based bidding, with the aid of a simulated forecast. Fig. 5b indicates that the charging profiles of both units are quite similar, with notable differences arising only when physical constraints, such as energy capacity, come into play. When comparing the MADRL profiles in Fig. 5a, we can observe that the profiles differ notably, with the charging and discharging times for the two units mostly not aligning. For example, from March 23rd to March 25th, unit 21 (blue) starts charging earlier than unit 22 (orange).

Furthermore, from April 9th, unit 21 is active, while unit 22 is passive. Overall, this demonstrates the ability of MADRL to identify niche strategies for individual units, allowing the agents to follow unique strategies and generate profit while avoiding economically damaging overlap in the market. If this were to be modeled individually without MADRL, it would require numerous assumptions and prove difficult for a high number of storage units.

Conclusion

Electricity market models are crucial for system operators to ensure adequate capacity and for generation and demand unit operators to plan future investments. These models must account for the competitive and strategic behavior of market participants to produce realistic market results. While traditional approaches to modeling market

strategies, such as rule-based approaches, optimization, or equilibrium models, have been effective in the past, they are becoming increasingly complex to handle in light of ongoing electrification and decarbonization. The emergence of new and diverse actors in power markets, such as providers of demand-side flexibility or energy storage, presents a challenge as they are dispersed across the system and do not follow straightforward marketing strategies.

In this paper, we highlight the limitations of existing models and explore the potential of MADRL to address these challenges. However, the computationally large environments of power system models, combined with a continuous state and action space, make learning computationally expensive and can hinder model convergence. To address this, we propose a new architecture for the state space of a centralized critic that receives a limited amount of broadcast information from individual agents. This architecture enables computational feasibility[2] for conducting simulations with a large number of learning agents within the available resources. Moreover, this approach ensures scalability and facilitates the generation of models with larger environments.

We demonstrate the applicability of our model using two case studies and compare our results against rule-based and optimization-based heuristic bidding strategies and optimization approaches. As demonstrated in Results, our model performed better than utilized baselines regarding profits for the multi-agent setup, which shows its ability to produce competitive profit-driven behavior for energy storage units. This approach also solved the "avalanche" effect, in which applying similar behavior policies to multiple agents leads to high disturbances in the system (very high charge or discharge power in this case). It did so by generating emerging niche strategies of competing storage systems that allow them to integrate well within the same power market.

It is also important to acknowledge some limitations of our study. First, we do not compare the resulting market clearing prices and historical values. Additionally, we do not investigate the impact of storage unit behavior on market prices. Exploring these aspects represents an avenue for future research.

Second, we do not include the interpretability aspect of DRL methods. The black-box nature of these algorithms poses challenges to their widespread application. Comprehending the decision-making process of DRL algorithms is difficult, so addressing this limitation for real-life applications like bidding strategies is crucial. Techniques from explainable RL, as reviewed by Puiutta and Veith (2020), can improve the trustworthiness and interpretability of these models, enhancing their practical applicability.

In summary, we contribute a method for electricity market modelers to simulate the strategic market behavior of heterogeneous market actors and for power market analysts to derive market strategies for their power system resources. Our proposed approach provides a more elegant and sophisticated solution to the challenges faced by traditional modeling techniques.

Abbreviations

RL	Reinforcement learning
DRL	Deep reinforcement learning

[2] The learning did not converge and a large number of agents could not fit into the available memory with the initially designed states.

MADRL	Multi-agent deep reinforcement learning
MDP	Markov decision process
TD3	Twin delayed deep deterministic policy gradients
MATD3	Multi-agent twin delayed deep deterministic policy gradients
DDPG	Deep deterministic policy gradient
MADDPG	Multi-agent deep deterministic policy gradient
VRE	Variable renewable energy

Author contributions
NH: Conceptualization, Methodology, Software, Visualization, Investigation, Writing—Original draft; AW: Conceptualization, Supervision, Writing—Reviewing and Editing, Funding acquisition; PS: Conceptualization, Supervision, Writing—Reviewing and Editing, Project administration.

About this supplement
This article has been published as part of Energy Informatics Volume 6 Supplement 1, 2023: Proceedings of the 12th DACH+ Conference on Energy Informatics 2023. The full contents of the supplement are available online at https://energyinformatics.springeropen.com/articles/supplements/volume-6-supplement-1.

Funding
This work has been conducted in the context of the project "ASSUME: Agent-Based Electricity Markets Simulation Toolbox", which is funded by the the German Federal Ministry for Economic Affairs and Climate Action under Grant number BMWK 03EI1052A.

Availability of data and materials
Source code and input data used for the scenarios are available on GitHub and can be accessed using the following link: https://github.com/INATECH-CIG/storage-flexRL.git

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI and AI-assisted technologies in the writing process
During the preparation of this work the authors used ChatGPT in order to improve the readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Published: 19 October 2023

References

Anwar M, Wang C, de Nijs F, Wang H (2022) Proximal policy optimization based reinforcement learning for joint bidding in energy and frequency regulation markets. In: 2022 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5
Böhringer Christoph, Rutherford Thomas F (2006) Combining top-down and bottom-up in energy policy analysis: a decomposition approach. SSRN Electron J
Boland Lawrence A (2017) Equilibrium models in economics. Oxford University Press
Bonabeau E (2002) Agent-based modeling: methods and techniques for simulating human systems. Proc Natl Acad Sci 99(suppl_3):7280–7287
Brown T, Hörsch J, Schlachtberger D (2018) Pypsa: Python for power system analysis. J Open Res Softw 6
Bundesnetzagentur: BNetzA List of Power Plants, Online (2021)
Bundesnetzagentur: SMARD—Download Market Data (2021). https://www.smard.de/en/downloadcenter/download-market-data Accessed 2022-05-11
Cao D, Hu W, Zhao J, Zhang G, Zhang B, Liu Z, Chen Z, Blaabjerg F (2020) Reinforcement learning and its applications in modern power and energy systems: a review. J Modern Power Syst Clean Energy 8(6):1029–1042
Cobbe K, Klimov O, Hesse C, Kim T, Schulman J (2019) Quantifying Generalization in Reinforcement Learning. arXiv:1812.02341
Conejo AJ, Baringo L (2018) Power System Operations vol. 11. Springer
Deissenroth M, Klein M, Nienhaus K, Reeg M (2017) Assessing the plurality of actors and policy interactions: agent-based modelling of renewable energy market integration. Complexity 2017
Dong Y, Dong Z, Zhao T, Ding Z (2021) A strategic day-ahead bidding strategy and operation for battery energy storage system by reinforcement learning. Electric Power Syst Res 196: 107229
Du Y, Li F, Zandi H, Xue Y (2021) Approximating nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning. J Modern Power Syst Clean Energy 9(3):534–544
EEX—Market Data: EEX Group DataSource—Power—Natural Gas (2019). https://www.eex.com/en/market-data Accessed January 2020
EEX: Market Data—Environmental Markets—Auction Market (2019). https://www.eex.com/en/market-data/environmental-markets Accessed January 2020

Erev I, Roth AE (1998) Predicting how people play games: reinforcement learning in experimental games with unique. Mixed strategy equilibria. Am Econ Rev 88(4):848–881
(Firm), S.G.M.I.: World Electric Power Plants Database, March 2017. Harvard Dataverse (2017). https://doi.org/10.7910/DVN/OKEZ8A
Frew BA, Jacobson MZ (2016) Temporal and spatial tradeoffs in power system modeling with assumptions about storage: An application of the power model. Energy 117:198–213
Fujimoto S, Hoof H, Meger D (2018a) Addressing function approximation error in actor-critic methods. In: International Conference on Machine Learning, pp. 1587–1596. PMLR
Fujimoto S, van Hoof H, Meger D (2018b) Addressing function approximation error in actor-critic methods. 35th International Conference on Machine Learning, ICML 2018 4: 2587–2601. arXiv:1802.09477
Giesecke J, Mosonyi E (2009) Wasserkraftanlagen. Springer, Berlin Heidelberg, Berlin, Heidelberg
Gronauer S, Diepold K (2022) Multi-agent deep reinforcement learning: a survey. Artif Intel Rev 1–49
Haarnoja T, Zhou A, Abbeel P, Levine S (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning, pp. 1861–1870. PMLR
Hache E, Palle A (2019) Renewable energy source integration into power networks, research trends and policy implications: a bibliometric and research actors survey analysis. Energy Policy 124:23–35
Hernandez-Leal P, Kaisers M, Baarslag T, de Cote EM (2017) A survey of learning in multiagent environments: dealing with non-stationarity. arXiv. arXiv:1707.09183
Hilpert S, Kaldemeyer C, Krien U, Günther S, Wingenbach C, Plessmann G (2018) The open energy modelling framework (oemof)--a new approach to facilitate open science in energy system modelling. Energy Strategy Rev 22:16–25
Huppmann D, Egging R (2014) Market power, fuel substitution and infrastructure--a large-scale equilibrium model of global energy markets. Energy 75:483–500
Koch M, Bauknecht D, Heinemann C, Ritter D, Vogel M, Tröster E (2015) Modellgestützte Bewertung von Netzausbau im europäischen Netzverbund und Flexibilitätsoptionen im deutschen Stromsystem im Zeitraum 2020–2050. Zeitschrift für Energiewirtschaft 39(1):1–17
Künzel T (2019) Entwicklung eines agentenbasierten Marktmodells zur Bewertung der Dynamik am deutschen Strommarkt in Zeiten eines steigenden Anteils erneuerbarer Energien. PhD thesis, Karlsruher Instituts für Technologie
Leuthold FU, Weigt H, von Hirschhausen C (2012) A large-scale spatial optimization model of the European electricity market. Netw Spatial Econ 12(1):75–107
Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2015) Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971
Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2019) Continuous control with deep reinforcement learning. arXiv:1509.02971
Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. In: Machine Learning Proceedings 1994, pp. 157–163. Elsevier,
Liu X, Conejo AJ (2021) Single-level electricity market equilibrium with offers and bids in energy and price. IEEE Trans Power Syst 1
Lowe R, Wu YI, Tamar A, Harb J, Pieter Abbeel O, Mordatch I (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. Adv Neural Inf Proc Syst 30
McConnell D, Forcey T, Sandiford M (2015) Estimating the value of electricity storage in an energy-only wholesale market. Appl Energy 159:422–432
Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602
Mnih V, Badia AP, Mirza L, Graves A, Harley T, Lillicrap TP, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. 33rd International Conference on Machine Learning, ICML 2016 4: 2850–2869
Niu H (2005) Models for electricity market efficiency and bidding strategy analysis. Dissertation, The University of Texas at Austin, Texas. http://hdl.handle.net/2152/1643
Orgaz A, Bello A, Reneses J (2017) Multi-area electricity market equilibrium model and its application to the European case. In: 2017 14th International Conference on the European Energy Market (EEM), pp. 1–6. IEEE
Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
Pinto T, Vale Z, Sousa TM, Praça I, Santos G, Morais H (2014) Adaptive learning in agents behaviour? A framework for electricity markets simulation. Integr Comput-Aided Eng 21(4):399–415
Puiutta E, Veith EM (2020) Explainable reinforcement learning: A survey. In: Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4, pp. 77–95. Springer
Qussous R, Harder N, Weidlich A (2022) Understanding power market dynamics by reflecting market interrelations and flexibility-oriented bidding strategies. Energies 15(2):494
Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347
Statistische Bundesamt: Data on energy price trends—Long-time series from January 2005 to June 2021 (2021). https://www.destatis.de/EN/Themes/Economy/Prices/Publications/Downloads-Energy-Price-Trends/energy-price-trends-pdf-5619002.pdf Accessed 08/10/2021
Tesfatsion L, Battula S (2020) Analytical SCUC/SCED Optimization Formulation for AMES V5.0: ISU General Staff Papers. https://ideas.repec.org/p/isu/genstf/202007020700001108.html
Umwelt Bundesamt: Datenbank: Kraftwerke in Deutschland (2020). https://www.umweltbundesamt.de/dokument/datenbank-kraftwerke-in-deutschland Accessed 01/05/2020
Ventosa M, Baillo A, Ramos A, Rivier M (2005) Electricity market modeling trends. Energy Policy 33(7):897–913

Verdaasdonk F, Demir S, Paterakis NG (2022) Intra-day electricity market bidding for storage devices using deep reinforcement learning. In: 2022 International Conference on Smart Energy Systems and Technologies (SEST), pp. 1–6
Wang H, Zhang B (2018) Energy storage arbitrage in real-time markets via reinforcement learning. In: 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5
Watkins CJ, Dayan P (1992) Q-learning. Machine Learning 8(3):279–292
Weidlich A (2008) Engineering interrelated electricity markets: an agent-based computational approach. PhD thesis
Weidlich A, Künzel T, Klumpp F (2018) Bidding strategies for flexible and inflexible generation in a power market simulation model 532–537
Ye Y, Qiu D, Li J, Strbac G, Member S (2019) Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets? A novel multi-agent deep reinforcement learning approach. IEEE Access 7:130515–130529