Energy 214 (2021) 118870

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Modeling competition of virtual power plants via deep learning

Markus Löschenbrand

Sintef Energy Research, Sem Sælands Vei 11, 7034, Trondheim, Norway

ARTICLE INFO

Article history:
Received 5 June 2020
Received in revised form
13 August 2020
Accepted 17 September 2020
Available online 28 September 2020

Keywords:
Neural networks
Nash game
Virtual power plants
Renewables
DC-OPF

ABSTRACT

Traditionally, models pooling flexible demand and generation units into Virtual Power Plants have been solved via separated approaches, decomposing the problem into parts dedicated to market clearing and separate parts dedicated to managing the state-constraints. The reason for this is the high computational complexity of solving dynamic, i.e. multi-stage, problems under competition. Such approaches have the downside of not adequately modeling the direct competition between these agents over the entire considered time period. This paper approximates the decisions of the players via 'actor networks' and the assumptions on future realizations of the uncertainties as 'critic networks', approaching the tractability issues of multi-period optimization and market clearing at the same time. Mathematical proof of this solution converging to a Nash equilibrium is provided and supported by case studies on the IEEE 30 and 118 bus systems. Utilizing this approach, the framework is able to cope with high uncertainty spaces extending beyond traditional approximations such as scenario trees. In addition, the paper suggests various possibilities of parallelization of the framework in order to increase computational efficiency. Applying this process allows for parallel solution of all time periods and training the approximations in parallel, a problem previously only solved in succession.

Credit author statement

Markus Löschenbrand is the single author of the paper and developer of the methods.

1. Introduction

Uncertainty is becoming an essential characteristic of modern power systems. Increasing renewable generation will impose further challenges as well as increase the future need for flexibility in power systems under liberalized markets, such as the European mainland [1].

Provision of flexible capacity from the consumer side imposes further questions on the regulatory design side [2]. As a result of non-standardized regulations in addition to vast geographical differences in topology of such flexibility providers, analysis of market power impacts of flexible demand is gaining in its importance as a research topic.

In addition to the complications of regular electricity market competition models (i.e. the high computational complexity of the interactions between several players under consideration of grid constraints) such problems also encounter high dimensions of uncertainty (i.e. stochastic load and consumption patterns). This comes in addition to the computational complexity of the dynamic nature of such Bellman problems.

To cope with this complexity, bi-(or higher)level problems are applied to solve problems for competition with demand aggregators. For the sake of clarification it has to be mentioned that the two levels referred to here are **the problem of market clearing (the competitive aspect)** and **the problem of solving the state constraints (the temporal, or: dynamic, aspect)** as illustrated in Fig. 1. This might be in contrast to traditional literature on market games that usually solely refer to problems such as Stackelberg competition when utilizing the term "bi-level".

Dealing with such a problem of implementing strategic bidding in problems under multiple periods (i.e. dynamic or sequential games) specifically for operation of storage and flexible demand has been attempted in previous literature for problems of small size [3].

Another example for profit-making demand aggregators operating storage units and flexible demand is presented in Ref. [4] which uses discrete dynamic programming and a power trajectory baseline to obtain a tractable problem. Ref. [5] models a strategic aggregator of flexible demand as a bi-level problem, where the aggregator acts a single Stackleberg-price maker in a day-ahead

E-mail address: markus.loschenbrand@sintef.no.

https://doi.org/10.1016/j.energy.2020.118870

Nomenclature

Index

S, D	Supply, Demand indicators
$t \in T$	time periods [h]
$i, i_2 \in I$	generation units/consumers
$j \in J$	players
$n \in N$	bus node
ξ	uncertainty indicator
n^{slack}	slack bus
$*$	optimal/clearing value

Variable

$q^S \in \mathbb{R}$	generated quantity [MWh]
$q^D \in \mathbb{R}^+$	purchased quantity [MWh]
$\delta \in \mathbb{R}$	voltage angle [$\circ$]
$p^* \in \mathbb{R}$	market clearing price [€/MWh]
$\Delta q^S_{i,t} \in \mathbb{R}$	generation rescheduling [MWh]
$\Delta q^D_{i,t} \in \mathbb{R}$	load shedding [MWh]

Stochastic Parameters

$\underline{q}^S_{i,t}, \overline{q}^S_{i,t}$	generation capacity [MW]
$\overline{s}_{i,t}$	storage capacity [MW]
$\overline{l}_{i,t}$	load shift capacity [MW]
d	consumption [MWh]

Deterministic Parameters

η^s, η^l	storage, load shift efficiencies [%]
b	number of minibatches in a batch
B	transmission line susceptance [siemens]
$\overline{F}$	line flow capacity [MW]
VLL	value of lost load [€/MWh]
VCC	value of corrective control [€/MWh]

Functions

Π	profit [€]
c	generation cost [€/MWh]
p	market price [€/MWh]
$\mathbb{O}$	feasible space
π	policy function
φ	value function
f	generic function denominator

Others

$r \in R$	replay memory batch

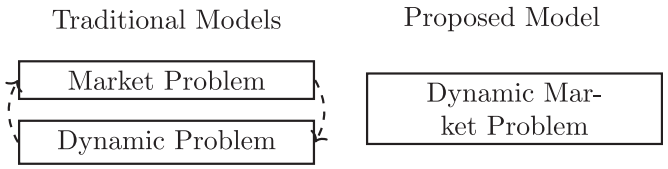

Fig. 1. Bi-Level vs. Single-Level Model.

market. The paper shows that a risk-taking aggregator is able to bid under a profit.

Another paper analyzing strategic aggregators of demand response as a Stackelberg model is presented in Ref. [6]. In this paper, the first level is the equilibrium amongst generators and the second level the decision of the demand aggregator.

Ref. [7] introduces another bi-level problem to approach to reach an equilibrium between operational decisions in the first level and the market clearing in the second level. In specific, this paper focuses on the analysis of long-term effects such as investments in generation capacity and demand response. In similar manner, Ref. [8] formulates a transmission expansion planning and combines the bi-level problem of Ref. [7] with the transmission problem to formulate a tri-level problem. However, both of those papers also simplify the dynamic aspect of shifting demand from one period to a later period and instead focus on the investment decision problem.

As shown, the literature suggests separation of the problems of aggregation of demand/flexibility and generation, only focusing on the effects of one form on the other instead of the bidirectional effects that single-level competition models offer. In real world systems, both parties participate in the same market. To underline this, the publication will hereon refer to flexible demand and storage aggregated with generation as 'Virtual Power Plants' (VPPs). Building on this literature, this paper intends to address the following question: assumed VPPs stand in direct competition to each other and to generation - how can the impact on markets and grid be modeled efficiently? In specific, the model below will focus on large uncertainty spaces as found in e.g. systems with high shares of renewables.

Utilizing such uncertainty spaces in problems under storage is of high computational complexity. This is even the case for single-player problems [9]. Thus, problems generally attempt to reduce the decision space, e.g. by considering only two stages such as in classical stochastic programming [10]. However, considering temporal decisions such as storage or demand shifts connects all future time periods and thus does not allow for such a two-stage decomposition. In practical applications, this problem can be solved via ignoring this connection over time [11]. However, as shown in e.g. Ref. [3] or Ref. [12], this temporal shift can itself provide a strategic asset, that would be therefore also ignored.

For dynamic problems without competition, approximations from the field of machine learning have already been established as state-of-the-art [13]. In power systems, these methods have also been introduced to non-competitive electricity storage and flexible demand [14].

Unrelated to power systems, similar approximations have been shown to efficiently solve agent-based games of competition [15].

Such approximation schemes to solve dynamic competition have also been introduced into power system literature. Ref. [16] deals with high uncertainty space by applying linear approximations in form of Benders cuts. Ref. [17] utilizes reinforcement learning to solve a dynamic game in power system security and assesses the systems robustness to attacks. Ref. [18] applies polynomial approximations in order to generate supply function approximations and yield a Nash game amongst approximated agents. Even though the model allows yielding multiple Nash equilibria, scalability to practical sizes is only able via problem decomposition, and thus only allowing for nodal and not system-wide competition.

In order to efficiently model the market effects of problems under demand aggregators, or VPPs in general, modeling such

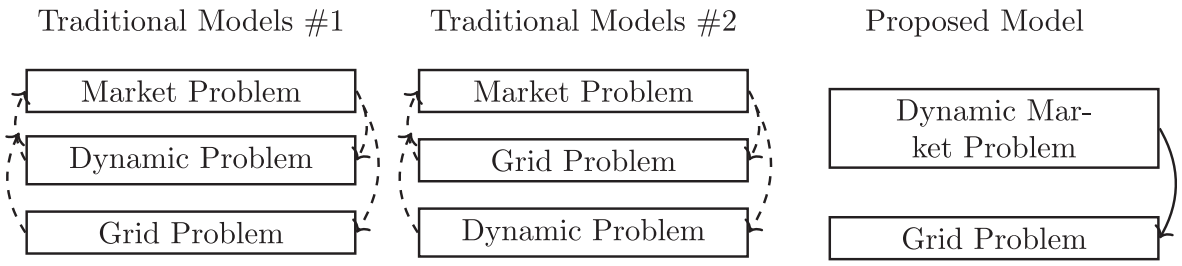

Fig. 2. Tri-Level vs. Bi-Level Model.

large-area coordination is, however, at the core of the problem. To approach this topic, the here proposed model framework builds on work from the intersection of reinforcement learning and game theory, most notably Ref. [19]. Even though proposals using deep-learning techniques to achieve agent-games of competition exist for smaller examples of dynamic games in electricity systems [20], scalability still requires the attention of researchers. This is supported by the bulk of current day literature in power markets [21], which show a clear focus on single-agent optimization problems and approximations on the decision space (e.g. Q-Learning requiring discretized decisions from agents in a problem that is continuous). Nonetheless, recent publications have approached continuous decisions. As such, Ref. [22] applied an actor-critic architecture on a single-period electricity market game and found superior convergence compared to traditional market clearing methods. Focusing only on a single entity, Ref. [23] showed actor-critic methods to be able to optimize bidding decisions of an electricity generator under consideration of non-linear constraints and showed such methods to outperform the traditional machine learning methods. In the here presented paper, agents using such actor-critic algorithms are utilized in a multi-period market clearing problem, a necessity to implement the storage and demand shift constraints that are at the core of VPPs.

In summary it can be said that a scalable algorithm to solve multi-market clearing under demand shift, storage and associated uncertainty is needed for accurately modeling the competition between VPPs. This is what this paper presents, based on an agent-critic formulation for agent based market modeling. In addition to solving this problem of scalable continuous dynamic competition, the model presented here also extends to grid problems via a transformation of area based market clearing to physical transmission. A case study based on standardized test data demonstrates how this allows showing the grid impact of coordination of supply and demand resources distributed within the grid. The study is aimed specifically to be a representation of the Northern European power system, which has area-localized market clearing nested within a centralized transmission problem. The resulting model principle is shown in Fig. 2.[1]

In summary, this paper offers four novel insights:

I A single-level model for competition between aggregators of flexible demand, generators and hybrids combining both.
II A versatile and computationally efficient framework to derive Nash equilibria for stochastic multistage models.
III A decomposition concept to parallelize training of agents and time periods.
IV A bi-level model to transform area based market clearing results into physical transmission grid results.

[1] dotted lines represent that this connection might not exist.

The structure of this paper is the following: Section 2 approaches the question of what deep learning contributes to the problem. Section 3 illustrates the model and thus present contribution I. Section 4 shows the solution framework used to solve this problem. It thus provides contribution II. Further, the section shows how to decompose the problem and increase computational efficiency, thus also providing contribution III. To extend the yielded solution to the grid, Section 5 introduces contribution IV. Sections 6 and 7 present and solve case studies of respectively small and practical size with high uncertainty space, outperforming previously presented small-scale problems from literature. Finally, Section 8 concludes the paper.

The implementation of the proposed model and solution framework requires a multitude of tools and techniques to work in unison. For the sake of transparency, the utilized tools and methods in the case study presented in this paper are listed here:

Training the Actor/Critic Networks: Pytorch [24].

Solving the Market Clearing Problem (5f): COBYLA ('Constrained Optimization BY Linear Approximations') introduced in Ref. [25] and extended on in Pagmo [26].

Solving the Transmission Problem (7): Linear Programming in Pyomo [27].

2. Why deep learning?

The **first main reason** for applying deep learning in the given context of competition amongst virtual power plants can be found in the model of the competition and the related approximations of the agents involved.

Within machine learning, deep learning refers to the utilization of layered functions (networks) that are used to approximate given problems by adjusting weights and biases via a training process. These then trained networks can then be utilized to supply approximations of (expected) outputs for a given unknown input. Such approximations have been utilized within energy markets before the recent surge of deep learning, which came as a result of specialized hardware based on GPUs which allowed for more efficient training. In the context provided in this paper the approximation is that of agent decisions to supply functions that allow for finding market equilibria amongst them. Based on traditional economics, in power system this topic has historically been discussed mainly in regards to linear approximations which allow for convex games and therefore simplify the search for equilibria [28]. Nonetheless, non-linear approximations have been proposed. However, most of these approximations require convexity, an example is provided by Ref. [29] which proposes quadratic supply functions. A non-linear approximation that does not require convexity is presented in Ref. [18] that uses a non-deep machine learning approach. Instead of neural networks polynomials are here utilized for the supply functions. Albeit promising in large-scale applications, this method still has significant downsides compared to deep learning approaches. The main issue is that for

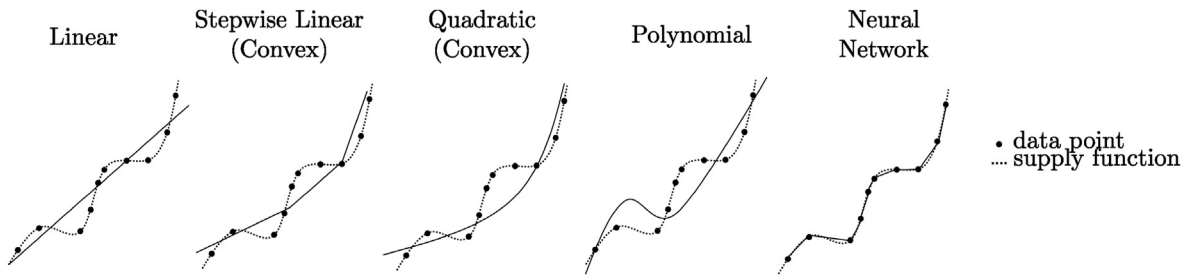

Fig. 3. Comparison of supply function approximations.

non-univariate approximations (i.e. any agent that has more than a single decision to make, e.g. a vector of decisions) such a polynomial may become intractable for the equilibrium problem very rapidly. The solution might be to adjust the polynomial (the referenced paper uses a heuristic to set certain coefficients of the polynomial to 0) in order to simplify it, which can help to reduce the scale of the problem. However, this simplification is also not scalable to a large degree, as it is still constrained by the computationally demanding market clearing algorithm. In the referenced paper this issue is approached by slicing the problem into smaller pieces, which allows scalability but only is able to implement local (i.e. nodal) competition and not direct competition of all units over the entire network. The reason for this is that the mechanism utilized to conduct the non-linear market clearing, the Gröbner basis, is not able to tractably cope with large polynomials. As a result, the mechanism is not able to scale well to problems with large decision spaces (such as a large number of flexible demand, storage and generation units that make up VPPs).

Providing a mechanism that is able to do so is the aim of the here presented paper. As discussed above, this is made possible by the non-linear but yet hardware efficient approximation capabilities of neural networks. A graphical representation of this can be found in Fig. 3.

In addition to additional advantages such as using replay memory for training, use various data points in parallel via batch gradient descent and the use of dropout to robustly train the weights and biases, the reason of using deep learning related to the competition is the high non-convexity that approximations can be efficiently trained on. This is ensured via activation functions, which are non-linear layers that are wrapped around the linear layers.[2] These activation functions represent non-convexity in the output approximation, but are convexified in order to allow for gradients and thus training of the network. With these non-linear layers, the agents are able to adjust their expectations via non-linear supply functions. Using a learning approach and specifically learning the reactions of the other agents, the agents manage to converge to an equilibrium state. This can be done via self-play, i.e. letting the agents adjust themselves based on other agents continuously until an equilibrium is reached. However, the proposed non-linear approximations are not without downside. Traditional machine learning approaches require large amounts of data to train large networks. As in the here proposed example a data point is the result of a players optimization problem, yielding sufficient data might not be possible. As a result, the here proposed application might require a lesser amount of layers in the neural network (e.g. a low two-digit number instead of a high three digit-number as in traditional machine learning applications).

The **second main reason** for applying deep learning in the given context of competition amongst virtual power plants can be found in the model of future outcomes. This is because virtual power plants have to make scheduling decisions on their storage and flexible load shifts that carry on into the future. Such dynamic programming problems, i.e. optimization problems with state constraints, fulfill the Bellman equation. In other words, the current values are connected to all potential values of the next period, of which in turn all are connected to all of the second-next and so forth. Problems like this can rapidly become intractable with small numbers of periods, even for discrete distributions of future uncertainties with low amounts of scenarios. Solutions for this problem of tractability are, again, approximations. And, similar to the modeling of competition, in dynamic optimization a popular approach of approximations is linear [30]. Also, and as described in Ref. [31] similar approaches such as described above exist. However, as described in Ref. [31] and expanded on in Ref. [13,32], again, neural networks show the best performance of approximation due to their highly non-linear characteristics.

Thus, and as shown in Fig. 2, the here proposed approach combines a single framework to apply the strength of deep learning on both the multi-period aspect as well as the complexity of the approximated supply functions in order to model competition amongst virtual power plants.

3. The Market Model.

In commodity markets, Cournot competition provides a popular model for interactions between generators [33]. The optimization problem for a single market participant j, below also referred to as 'players', in such a model can be described as:

$$\Pi_j^*(\xi) = \max_{q^S, q^D, s, l} \sum_{t\in T} \left[\sum_{i\in I_j^S} \left[p_t^* q_{i,t}^S - c_i\left(q_{i,t}^S\right) \right] - \sum_{i\in I_j^D} p_t^* q_{i,t}^D \right] \tag{1a}$$

s.t.

$$\underline{q}_{i,t}^S(\xi) \le q_{i,t}^S \le \overline{q}_{i,t}^S(\xi) \quad \forall i\in I^S, t\in T \tag{1b}$$

$$q_{i,t}^D + s_{i,t} - \eta_i^s s_{i,t-1} - l_{i,t} + \eta_i^l l_{i,t-1} = d_{i,t}(\xi) \quad \begin{matrix} \forall i\in I_j^D, \\ t\in T \end{matrix} \tag{1c}$$

$$0 \le s_{i,t} \le \overline{s}_{i,t}(\xi) \quad \forall i\in I_j^D, t\in T \tag{1d}$$

$$0 \le l_{i,t} \le \overline{l}_{i,t}(\xi) \quad \forall i\in I_j^D, t\in T \tag{1e}$$

The objective function (1a) maximizes the profits of the generation units whilst minimizing the payments for the purchased

[2] layers meaning approximations that are stacked on each other.

energy.[3] Capacity constraint (1b) describes the ranges in which the generation units operate. The state constraint (1c) describes that a player can utilize storage and demand shift in order to fulfill the uncertain demand given. Capacity constraints (1d) and (1e) describe the operational ranges for storage units and demand shift.

Here it can be observed that the VPP formulation provides a generalization of traditional power plants, aggregators and wholesale consumers (i.e. 'non-flexible' demand aggregators). For example could a traditional thermal power plant be modeled as the single plant of a single player without any (flexible) demand or storage capacities. This allows modeling the direct competition of traditional plants to virtual power and (flexible) demand aggregators.

In the here provided example, a market clearing for energy prices in every period t is assumed:

$$p_t^* = p_t\left(\xi, \sum_{i \in I^S} q_{i,t}^S\right) \quad \forall t \in T \tag{2}$$

The uncertainties presented in this model are:

- Generation provided by e.g. wind power plants subject to weather effects.
- Flexible demand provided by e.g. automatic heating subject to outside temperatures.
- Available storage capacity provided by e.g. electric vehicles subject to usage and thus disconnection from the grid.
- Energy prices provided by e.g. spot markets subject to price fluctuations.

As the subsequent sections will illustrate, having such a large uncertainty space presents the main hurdle in finding the equilibrium solutions.

Further, the competition is Cournot, i.e. a competition in quantity, which means the players will exercise their market power via strategically reducing or increasing the quantities bid and deviate from the social welfare optimal quantities. In traditional Cournot games this decision space is limited by the generation capacities (1b) of the generators. In the here presented model, however, there are also decisions on stored and shifted loads, i.e. decisions over several time periods. As the later presented case study shows, the agents also utilize those 'temporal' decisions strategically.

4. Solving the market model

Despite the problems' notational simplicity and disregarding issues with efficient scaling of the problem size via additional players,[4] the core problem itself has two issues: on one hand are time periods connected via state constraint (1c) and will for longer periods thus lead to combinatorial explosion (also referred to as the "curse of dimensionality"), on the other hand will uncertainties negatively affect the convergence to an equilibrium state.

Dealing with such uncertainty efficiently is already an issue in single player (and thus price-taker) optimization problems, resulting in stochastic [35,36] or robust optimization problems [37]. In games under uncertainty, solutions are thus often obtained via linearizations [38]. This paper will instead circumvent the requirement for dependence on (manual) linearizations and instead use (automatic) non-linear function approximators in form of neural networks.

Considering uncertainty in a game representing the previous market would lead to the following Nash equilibrium that would have to hold for all players j:

$$\Pi_j(\xi | q_i^* \, \forall i \in I_j) \geq \Pi_j(\xi | q_i \, \forall i \in I_j) \quad \begin{matrix} \forall q_i \in \mathbb{O}_i(\xi), \\ i \in I_j \end{matrix} \tag{3}$$

However, neither stability nor existence of such an equilibrium would be provided, as both the profit as well as the feasible space defined by the constraints of problem (1) are subject to uncertainty.

In order to address this, policy function approximations have been utilized in literature [39]. This means that it is assumed that instead of an optimal decision, an optimal policy $\pi_i^*(\xi)$ is decided upon. This leads to the following reformulation of the Nash equilibrium (3):

$$\Pi_j\left(\xi \middle| \pi_j^*(\xi)\right) \geq \Pi_j(\xi | \pi_j(\xi)) \tag{4}$$

This policy Nash equilibrium can be considered converged, if no change in policy $\pi_j(\xi)$ leads to a change in profits for any player j. This is a generalization of the previous Nash equilibrium (3), which considers fixed policies, i.e. players not reacting to uncertainty. A proof of convergence for a stochastic game is provided in Refs. [39].

In the real world, such Nash equilibrium approximations have successfully been used in real-time decision making in games with large decision spaces [40].

The advantage of using such a policy approximation, in literature also often referred to as an *actor*, is not only limited to that it allows to deal with finding the equilibrium under uncertainty for a static system. In addition to that, it also allows to solve the dynamic problem nested in the equilibrium model. Extending the problem by a value function approximation, also referred to as a *critic*, of the future states allows for an actor-critic model. Doing this allows the dynamic equilibrium problem to be solved similar to other approximate dynamic problems [13].

Establishing this value function approximation requires defining a reformulation of the objective function (1a) of the player problem as a Markov Decision Process.[5] This is done by assuming uncertainty becomes known in a given period t:

$$\Pi_{j,t}^*(\xi_t, s_{i,t-1}, l_{i,t-1} \, \forall i \in I_j) = \max_{q^S, q^D, s, l} \left[\sum_{i \in I_j^S} \left[p_t^* q_{i,t}^S - c_i\left(q_{i,t}^S\right) \right] - \sum_{i \in I_j^D} p_t^* q_{i,t}^D \right] + \varphi_{j,t+1}(s_{i,t}, l_{i,t} \, \forall i \in I_j) \tag{5a}$$

s.t.

$$\underline{q}_{i,t}^S(\xi_t) \leq q_{i,t}^S \leq \overline{q}_{i,t}^S(\xi_t) \quad \forall i \in I^S \tag{5b}$$

$$q_{i,t}^D + s_{i,t} - \eta_i^s s_{i,t-1} - l_{i,t} + \eta_i^l l_{i,t-1} = d_{i,t}(\xi_t) \quad \forall i \in I_j^D \tag{5c}$$

[3] For sake of simplicity, this model considers the decision maker to be unaffected by additional technical details such as degradation of storage or cost of load shifting and instead make price decisions purely based on operational profits. Nonetheless implementing this would be possible by adding an adequate cost component as a function of storage to this objective.

[4] An issue common in equilibrium problems [34].

[5] It has to be noted that in the here presented case, the uncertainties ξ_t and $x_t + 1$ are considered stage-wise independent. In case the uncertainties are dependent (e.g. in decision trees), the value function approximation for the next stage would have to consider the uncertainty of the current stage: $\varphi_{j,t+1}(\xi_t | s_{i,t}, l_{i,t} \, \forall i \in I_j)$. For the sake of simplification, this is however omitted here.

$$0 \leq s_{i,t} \leq \bar{s}_{i,t}(\xi_t) \quad \forall i \in I_j^D \tag{5d}$$

$$0 \leq l_{i,t} \leq \bar{l}_{i,t}(\xi_t) \quad \forall i \in I_j^D \tag{5e}$$

$$p_t^* = p_t\left(\xi, \sum_{i \in I_j^S} q_{i,t}^S + \sum_{j2 \in J \setminus j} \pi_{j2,t}(\xi_t)\right) \quad \forall t \in T \tag{5f}$$

This formulation shows a difference to traditional actor-critic models [32] and to techniques intended to map the model environment [19,41]: the characteristic of the optimization model being known allows for traditional solution techniques instead of utilizing the policy gradient to solve the problem for a specific player *j*. Instead the actor function is used to model the reactions to uncertainty of the other players *j*2, reducing the algorithms search space. This does not only allow a faster convergence for fitting the actor (as there are techniques to efficiently solve such deterministic single-stage problems), but also decreases the size of the actor, as only the variables interacting with other players have to be considered. In the here presented example, only generation influences the prices and thus demand, storage and load shedding do not have to be included in the actor but instead are only used in the critic.

The model also demonstrates another difference to dynamic competition models from literature: in the given formulation, the problem incorporates assumptions on time periods $t+ \quad 1 \notin T$. Compared to traditional problems under storage [42], the here provided formulation thus not require assumptions on parameters such as end-inventory levels or monetary value of such storage[6] and instead choses these values based on the assumptions for the future values provided by the function approximators, i.e. the neural networks. This removes another, traditionally manual, approximation task.

Due to the dynamic problem being influenced by the (policy) decisions on storage and load shift and the future uncertainties being unknown, the approximated value (i.e. the approximated profit) of period *t* thus can only consider the approximated values for the future outcome. As those values themselves consider the subsequent periods, the approximation is the classical Bellman function for dynamic programming. Convergence of such an approximation problem of the value function, i.e. to find $\varphi_{j,t}(s_{i,t-1}, l_{i,t-1} \forall i \in I_j) \approx \Pi_{j,t}(\xi_t, s_{i,t-1}, l_{i,t-1} \forall i \in I_j)$, can be achieved via bootstrapping of (enough) samples for the uncertainty.

initialize ϕ, π, R, l_0, s_0
for each episode **do**:
  sample ξ
  for each $t \in T$ **do**:
    solve problem (5) $\forall j \in J$
    update l_t, s_t
    calculate $\Pi'_{j,t}(\xi_t, q^*_{i,t} \forall i \in I) \forall j \in J$
    store $[t, l_t, s_t, \xi_t, \Pi'_t, q^*_t]$ in R
    sample batch $r \in R$ and **train** $\phi_{j,t}, \pi_{j,t} \forall j \in J$

(6)

After introducing a replay memory *R* and considering the profits within a single specific period *t* as Π_t, the algorithm achieving this convergence can be formulated as shown in Eq. (6). Information on convergence can be found in A.

Solving problem (5) in a specific period *t* for all players yields a minibatch $[l_t, s_t, \xi_t, \Pi_t^t, q_t^*]$ that is stored in this replay memory. In the same manner, a batch consisting of a number of *b* minibatches can be sampled via random draw from this replay memory *R*. It takes the form of:

$$r = \begin{bmatrix} \left[t^1, l_t^1, s_t^1, \xi_t^1, \Pi\prime_t^1, q_t^{*1}\right] \\ [\quad \cdots \quad] \\ \left[t^b, l_t^b, s_t^b, \xi_t^b, q_t^{*b}\right] \end{bmatrix} \tag{7}$$

Compared to traditional power market clearing models such as Ref. [43], here, similar to Ref. [4], the player decisions can be solved in parallel of each other. This is due to a decoupling as a player does not utilize other players value functions or its own policy to find an optimal solution in a single period.

The schematic presented in Fig. 4 outlines the dynamic process of a single such episode from the algorithm.

initialize ϕ, π, R, l_0, s_0
for each episode **do**:
  sample ξ
  solve problem (5) $\forall j \in J, t \in T$
  update l, s
  calculate $\Pi'_{j,t}(\xi_t, q^*_{i,t} \forall i \in I) \forall j \in J, t \in T$
  store $[t, l_t, s_t, \xi_t, \Pi'_t, q^*_t]$ in R $\forall t \in T$
  sample batch $r \in R$ and **train** $\phi_{j,t}, \pi_{j,t} \forall j \in J, t \in T$

(8)

Alternatively, and in case the uncertainties in a time period are stage-wise independent, all player and periodical decisions can be solved in parallel. This is shown in Eq. (8).

The schematic presented in Fig. 5 outlines a single episode of this parallel algorithm.

Compared to previous problems in literature [12] this problem converges based on drawn samples and thus not require approximations of the uncertainty space such as scenario trees or lattices but instead can receive real data or samples taken from distributions as input.

An analysis of the given problem is presented in A, which shows that for the given problem approximation, the stage-wise game is in fact deterministic, albeit non-convex, and the quality of the policy function approximation is not affected by the quality of the future value function approximation.

Actor-critic methods have been utilized in literature in agent

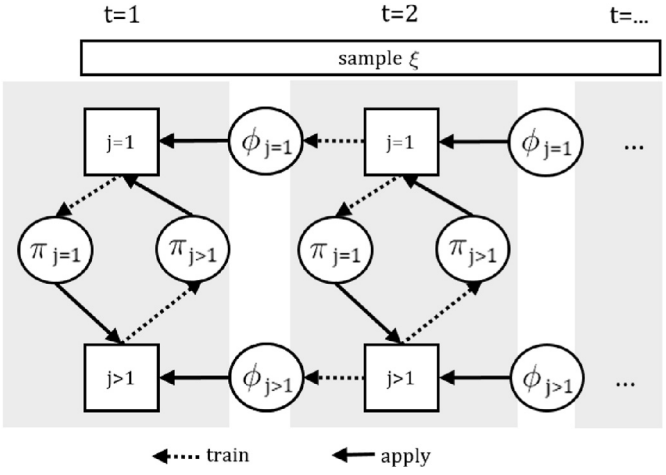

Fig. 4. Serial algorithm.

[6] In hydropower optimization this value is also known as 'water value'.

based models and shown to converge towards a Nash equilibrium [19,41]. However, similar to traditional machine learning techniques, those techniques have mostly been utilized with either discrete or discretized decision spaces. By using non-linear optimization to find the equilibrium from a perspective from each individual player j (i.e. solving problem (5)) instead of relying on the actor-critic model itself to find the optimal point, the here presented algorithm is thus able to use continuous decisions (as it does not solve for them but instead only approximates them).

inputs: batch r, critic network ϕ

get $t^r = \begin{bmatrix} t^1 \\ \dots \\ t^b \end{bmatrix}$, $l^r = \begin{bmatrix} l^1_{t^1} \\ \dots \\ l^b_{t^b} \end{bmatrix}$, $s^r = \begin{bmatrix} s^1_{t^1} \\ \dots \\ s^b_{t^b} \end{bmatrix}$, $\Pi^r = \begin{bmatrix} \Pi'^1_{t^1} \\ \dots \\ \Pi'^b_{t^b} \end{bmatrix}$

calculate critic value $\phi^r = \begin{bmatrix} \phi_{t^1}(s^1, l^1) \\ \dots \\ \phi_{t^b}(s^b, l^b) \end{bmatrix}$

calculate critic loss function $\text{loss}^r = (\Pi^r - \phi^r)^2$

calculate gradients ∇loss^r

update ϕ via optimizer using gradients ∇loss^r

(9)

Applying Mean Squared Error as a loss function, the critic function can be trained as in Eq. (9). The utilized optimizer and structure of the neural network applied in this paper can be found in Appendix B.

inputs: batch r, actor network π

get $t^r = \begin{bmatrix} t^1 \\ \dots \\ t^b \end{bmatrix}$, $\xi^r = \begin{bmatrix} \xi^1_{t^1} \\ \dots \\ \xi^b_{t^b} \end{bmatrix}$, $q^{*r} = \begin{bmatrix} q^{*1}_{t^1} \\ \dots \\ q^{*b}_{t^b} \end{bmatrix}$

calculate actor value $\pi^r = \begin{bmatrix} \pi_{t^1}(\xi^1) \\ \dots \\ \pi_{t^b}(\xi^b) \end{bmatrix}$

calculate actor loss function $\text{loss}^r = (q^{*r} - \pi^r)^2$

calculate gradients ∇loss^r

update π via optimizer using gradients ∇loss^r

(10)

In similar manner, the actor network can be trained as shown in Eq. (10).

The training process of the critic and actor approximation networks is also displayed in Figs. 6 and 7 respectively. Fig. 7 displays both games under perfect and imperfect information. For games under perfect information, knowledge on the outcome of parameters of all players is used, whereas in incomplete information knowledge on uncertain parameters can be withheld. In the here presented example, this could e.g. mean that an actor is trained only on the outcomes of available storage and load shift capacities $\bar{s}_{i,t}(\xi_t), \bar{l}_{i,t}(\xi_t)$ for the specific player, i.e. for $I^D_{j,t}$, instead of also including other players' flexible capacities, i.e. training on I^D_t. In the case study presented within this paper, a game under complete information was assumed.

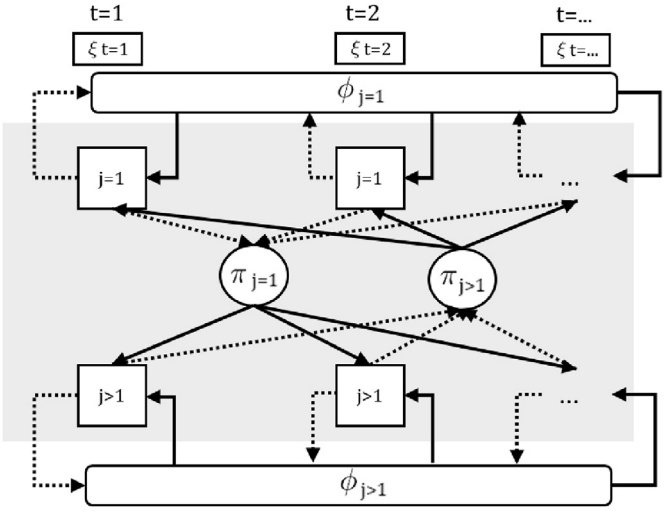

Fig. 5. Parallel algorithm.

5. The transmission system problem

As area pricing does not adequately represent the physical reality of the grid, the transmission system operator has to make adjustments to the market clearing results in order to ensure transmission line limits are accounted for. The corrective actions considered in the here applied model are those of load shedding and generation rescheduling as e.g. presented in Ref. [44].

To formulate this, a single-period DC-Optimal Power Flow (DC-OPF) is chosen as a representation of the adjustments the Transmission System Operator (TSO) has to conduct every period. The model utilizes the assumption presented in Ref. [45]: adjustments are made after realization of the uncertain variables. However, and as an accurate representation of the reality of many electricity markets such as found in the Scandinavian power system, the transmission system operator is considered to hold a reactive role disconnected from the market decisions, having the mission of ensuring adequate transmission under consideration of the clearing result presented by the power market. This leads to the following deterministic DC-OPF for every period t:

$$\min_{\Delta q^D, \Delta q^S, \delta} \text{VCC} \sum_{i \in I^S} \left| \Delta q^S_{i,t} \right| + \text{VLL} \sum_{i \in I^D} \Delta q^D_{i,t} \tag{11a}$$

s.t.

$$\forall i \in I^D_j :$$

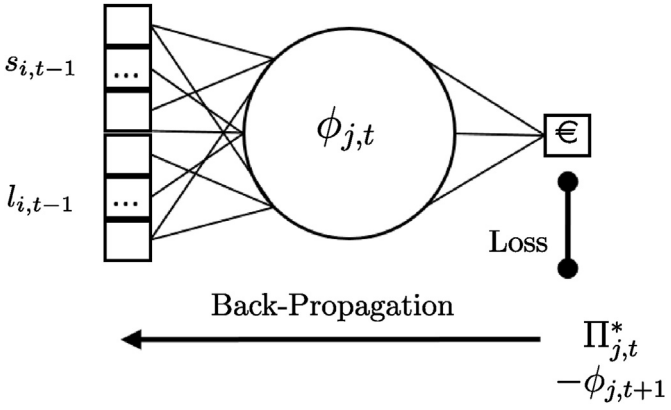

Fig. 6. Critic network.

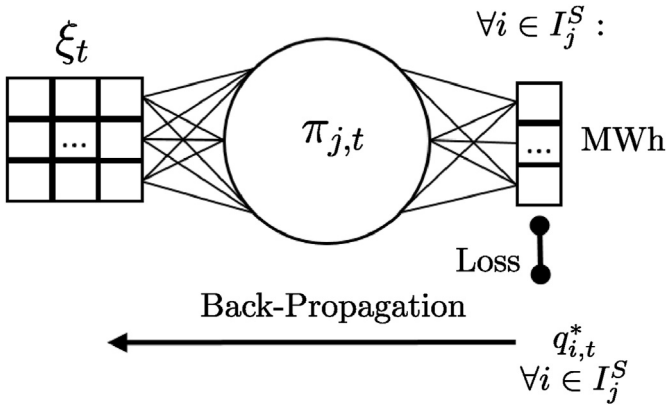

Fig. 7. Actor network.

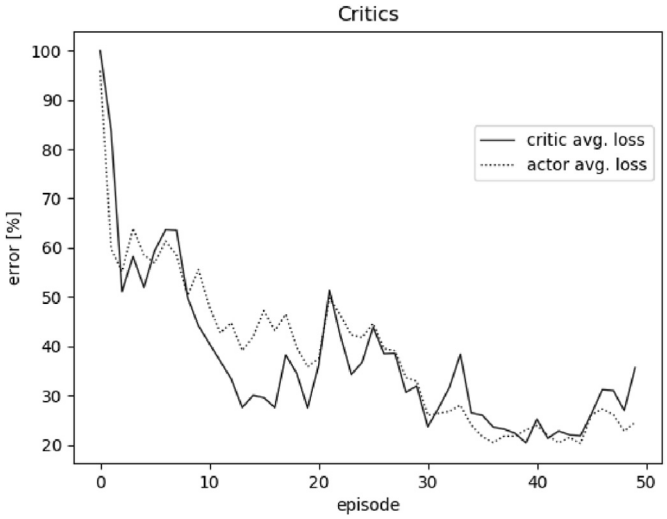

Fig. 8. Case I: 30 bus with 4 players.

$$\underline{q}^S_{i,t} \le q^S_{i,t} + \Delta q^S_{i,t} \le \overline{q}^S_{i,t} \quad \forall i \in I^S \tag{11b}$$

$$0 \le \Delta q^D_{i,t} \le q^D_{i,t} \quad \forall i \in I^D \tag{11c}$$

$$\sum_{i \in I^S_n} \left[q^S_{i,t} + \Delta q^S_{i,t} \right] - \sum_{i \in I^D_n} \left[q^D_{i,t} - \Delta q^D_{i,t} + s_{i,t} - \eta^s_i s_{i,t-1} - l_{i,t} + \eta^l_i l_{i,t-1} \right] = \sum_{n_2 \in N} B_{n,n_2} (\delta_{n,t} - \delta_{n_2,t}) \quad \forall n \in N \tag{11d}$$

$$-\overline{F}_{n,n_2} \le B_{n,n_2} (\delta_{n,t} - \delta_{n_2,t}) \le \overline{F}_{n,n_2} \quad \forall n \in N, n_2 \in N \tag{11e}$$

$$\delta_{n^{\text{slack}},t} = 0 \tag{11f}$$

The objective (11a) is the minimization of cost accrued by load shedding and corrective generation rescheduling. As strategic bidding on those would be considered a break of laws and against the interest of the power system operators, these costs are not considered in the market optimization problem of the players as presented in Eq. (1). Constraints (11b) and (11c) formulate the limits of the adjustment variables. The physical constraints given by Kirchhoff's laws are introduced via the nodal balance condition (11d). The physical line capacities are enforced via constraint (11e). The slack bus is defined via condition (11f). By replacing the adjustment of supply with two bidirectional variables, one of up- and one for downwards regulation, allows replacement of $\left|\Delta q^S_{i,t}\right| = \Delta q^{S,\text{up}}_{i,t} + \Delta q^{S,\text{down}}_{i,t}$. This transforms the transmission system problem into a Linear Problem.

6. Case study I - IEEE 30 bus & 3 days

The first chosen case study is that of an IEEE 30 bus test system extended to three periods. As mentioned in the model description, the units were considered to show uncertainty in price slope and elasticity, minimum and maximum generation limits, consumption and available storage and demand shift capacities.

Regarding performance, the convergence speed per episode was 100 s for the Nash game on an Intel i7-8850H CPU (all players in parallel) and training the networks took between 5 s per episode on a Nvidia Quadro P2000 GPU (all networks trained in parallel). Further information on the topology of the chosen networks used for both case studies can be found in Appendix B.

Fig. 8 shows the convergences of the actors and critics of the model, whereas 100% is the initial starting error.

Table 1 describes the setup of the four considered players. Player 1 is a pure generator that also holds the largest units (with capacities of around 200, 90 and 40 MW) and the other three players hold units with around 30, 25 and 40 MW respectively. The demand nodes are partioned equally amongst player 2, 3 and 4, with the players owning total storage capacities of 65, 18 and 5 MW respectively. Flexible demands where 72, 47 and 19 MW respectively. As the demand side players aim to fulfill this demand, their results can be expected to be negative (as the transaction between the end consumers to these aggregators is not considered in this model and instead their goal is cost minimization of the demand fulfillment). Nonetheless, one of the virtual power plants, namely the aggregator that is player 4, manages to succeed in making a profit over the duration of the three episodes.

Considering the demand results in Fig. 10 and comparing them to the supply results in Fig. 11 shows that the total demand of this player is outweighted by their available generation capacity. Solving the problem single-period would result in a negative result of −26.8€ for player 4 in period 1. Instead they not only manage to make a positive result in period 1 but also stay positive over the total time frame of three periods.

The reason for this can be identified in the potential of multi-period shifts via flexible demand and storage. As shown in Fig. 12 player 4 utilizes the storage and flexible demand capacities more than other players. Even though the other players had higher capacities in storage and flexibility, player 4 had the highest available

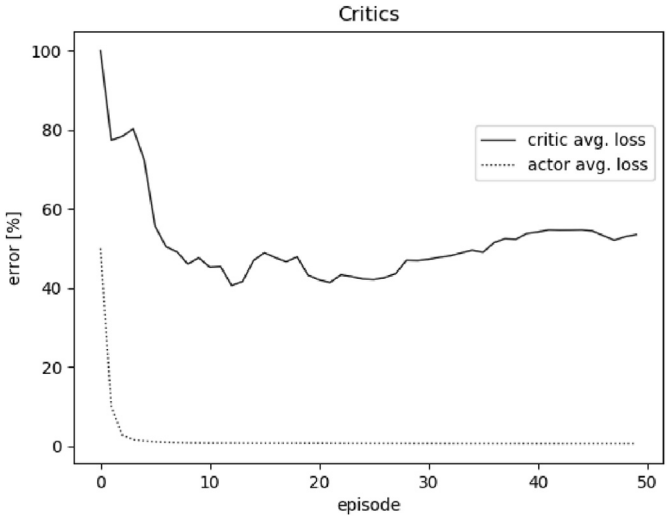

Fig. 9. Case II.1: 118 bus with 10 players.

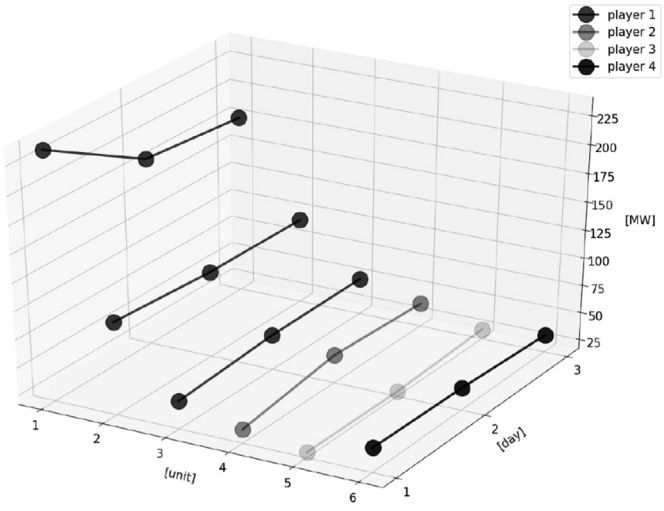

Fig. 11. Case I - supply results.

Table 1
Case I - player description and profit results.

Player	Generation Busses	Demand Busses	Profit period 1	Profit period 2	Profit period 3
$j = 1$	$I_1^S = \{1, 2, 5\}$	$I_1^D = \{0/\}$	13738 €	11853 €	11995 €
$j = 2$	$I_2^S = \{8\}$	$I_2^D = \{1, 2, \ldots, 10\}$	−5264 €	−5954 €	−5258 €
$j = 3$	$I_3^S = \{11\}$	$I_3^D = \{11, 12, \ldots, 20\}$	−368 €	−490 €	−315 €
$j = 4$	$I_4^S = \{13\}$	$I_4^D = \{21, 22, \ldots, 30\}$	428 €	−6 €	−329 €

generation capacity. This balance of generation and consumption allows the player to influence the prices via capacitating generation and benefit on the resulting price changes optimally via demand and supply shift. Due to the single-period result being negative, this effect would not have shown. This highlights the necessity for multi-period models as the one proposed in this paper.

However, this profit-maximizing utilization of storage by the player not only affects the profit results of the supply and demand side. As Fig. 13 illustrates, allowing the players to operate over multiple periods changes the transmission line utilization. This can also be expected, as due to multi-period operation, the players will now aim to purchase less energy in periods of high prices and buy more in periods of low prices. In the given example, this increases the line utilization from an average of 23% in the single-period case to 28% in the three-period case. Albeit higher utilization of the lines (and thus potentially having an effect on required long-term investments into line capacities), the result of the grid operator decreases by 13% from the single-to the three-period case. This means that such coordination on the demand and supply side can also have a positive impact on the underlying grid.

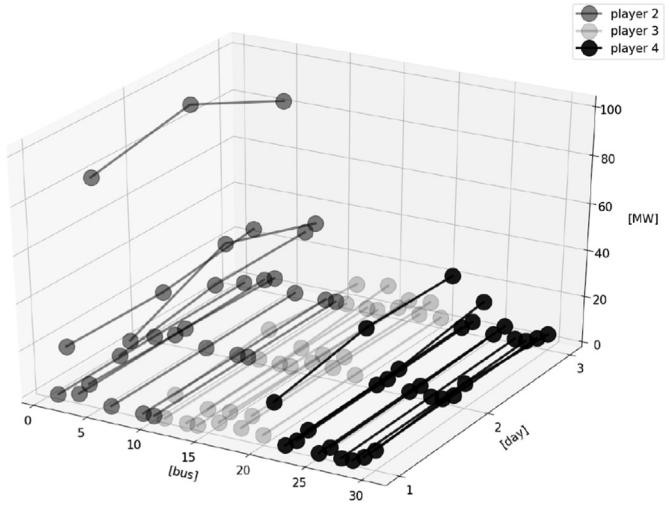

Fig. 10. Case I - demand results.

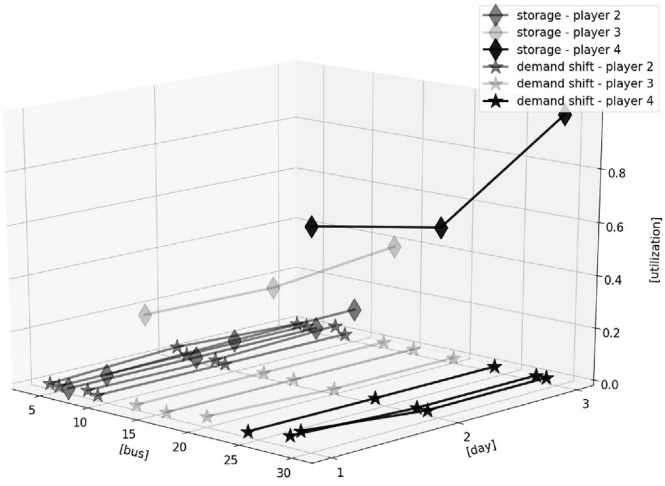

Fig. 12. Case I: Store and demand shift results.

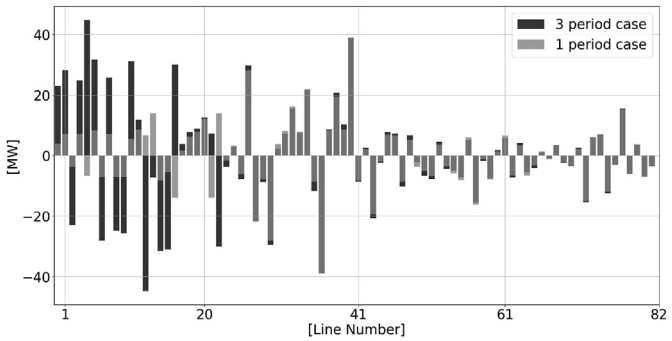

Fig. 13. Case I: Transmission results 3- and 1-period case.

7. Case study II - IEEE 118 bus & 7 days

The second analyzed case is an adjusted version of the IEEE 118 bus test system solved over a period of 7 days. As Fig. 14 illustrates, connecting these problems via storage and demand shift makes the problem equivalent to a $118 \times 7 = 826$ bus problem. The uncertainty space for the problem is a number greater than 10^{91} different possible outcomes an average standard deviation of 50% of the mean. Two cases were analyzed, the first case consisting of 5 generation companies owning the 19 generation units and 5 additional players managing the flexible demand that covers the entire demand side. Comparison case two merges the units of these 5 demand side players in a single entity. The convergence for the case is shown in Fig. 9. The graphs show a convergence to an equilibrium on the decisions (i.e. the actors) with continuing uncertainty on the stochastic future outcomes (i.e. persistent errors on the critic). In other words, even though the agents' assumptions on the future outcome is not correct, the actors still manage to converge and thus, by definition, find a equilibrium for this case with a high number of busses and large uncertainty The definition of this single equilibrium due to the decoupling of future and current states is shown in Appendix A.

Solution speed for a single episode of all decision problems was 180 s and training of the networks took between 10 and 15 s per episode.

To compare the market clearing results, a Monte Carlo analysis of 1000 different scenarios was conducted, whereas the average results are shown here.

These market clearing results for period 1 are displayed in Fig. 15 for case 1 ('separated') and in Fig. 16 for case 2 ('single'). The demand shown is adjusted for storage and shift of flexible loads to subsequent periods and the generation decisions are the (deterministic) choices made by the agents for this period, considering the future uncertainties.

The difference between the demand provided in case 1 to the demand provided in case 2 is −7.53%, where the supplies are nearly identical with a difference of −0.09%. As the clearing results show, the main difference in demand is clustered in the first 20 buses of the 118 bus network, as shown by the bottom left side of Figs. 15 and 16.

At first glance, this result seems paradox, considering aggregating the five separate demand side players to a single demand side player causes demand in the first period to decrease disproportionately over supply.

However, this difference in changes on demand and supply side be explained by storage and load shift changes. The case study illustrates how an aggregated player utilizes storage and demand shift differently over the separated players. This is shown in Figs. 17 and 18 for separated and single demand side players respectively. Even though 35 of the busses were supplied with storage capacity and 45 busses with flexible load capacity, the results show only a few selected of the busses having these capacities being utilized to a greater extend. Compared to the market clearing results, these changes in storage and load shift are less subtle between case 1 and case 2. It can be seen clearly that the single player chooses to utilize less of the available storage capacity in period 1, leading to more available loads in period 1 that, due to Kirchhoff's law have to be consumed then.

In single-period Cournot markets it can be shown that players with market power tend to withhold available supply in order to increase their personal profits [34]. In the here provided multi-period example, storage and demand shifts are used by the demand aggregator of case II.2, i.e. the entity that holds all the flexible demand and storage, to increase their demand surplus over the more competitive case II.1.

This is shown by the shrinking supply side surpluses in Fig. 19. Especially period 2 yields higher returns to the agents in the competitive case II.1, with most of the total results still lying above 20000 Euro per period. In the case of the demand aggregator (player 1 of Case II.2), this particular agent reduces the returns of the other agents that are now not able to exercise demand shifts and storage in order to minimize the cost for the flexible demand it is responsible for catering to. This shows that an agent that holds a monopoly on demand-side inter-temporal decisions uses market power to minimize the cost of catering to this demand.

The larger extend of the exercise of these demand shifts and storage is also shown in the flattening effect that temporal arbitrage causes (as e.g. discussed in Ref. [12]) - the surpluses are more even between the periods where a single entity controls all inter-temporal decisions.

Fig. 20 shows that an effect of aggregating the demand side into a single player seems to level the competition, leading to lower welfares of the involved players. The negative welfare outcomes are a result of not taking into account that the demand aggregators will be reimbursed for their fulfilled demand, thus, adjusting for this the real welfare result will be positive. Nonetheless, it shows that aggregating of the demand side resources resulted in lower supply side profits, as this is the side that gains the benefits of lower prices. In addition, and as discussed, the results are in line with outcomes of Cournot models from literature with monopolies on demand. The difference here is that the given model also considers uncertainty and inter-temporal decisions, with similar results as traditional Cournot models.

The comparison of the transmission line results for period 1 is displayed in Fig. 21, which illustrates case II.1 as black dots and case II.2 as gray stars. Merging the supply side players into a single entity increased the transmission system utilization by 0.4% but also decreased the system operators cost of adjusting demand and supply by 0.66%. This shows that a different constellation of players can have significant effects on the utilization of the transmission network. Even though in the here presented case studies the highest utilized transmission lines (left quadrant of Fig. 20) were left nearly identical in their utilization, the average utilization was

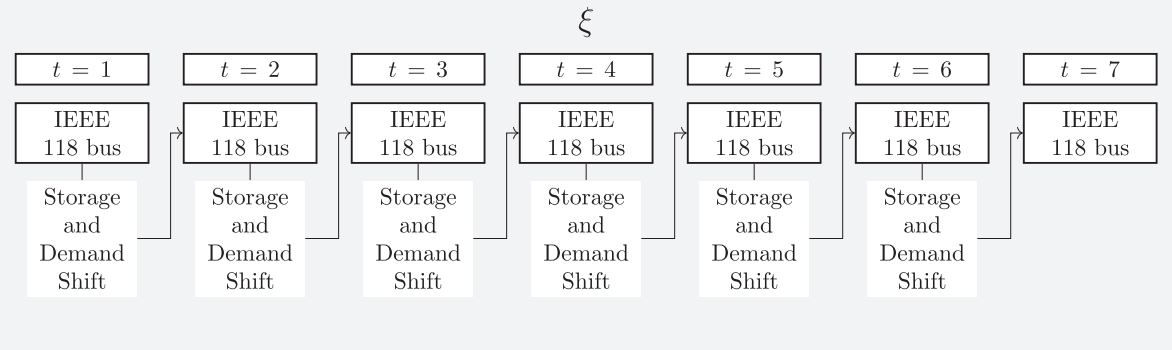

Fig. 14. Case study II problem.

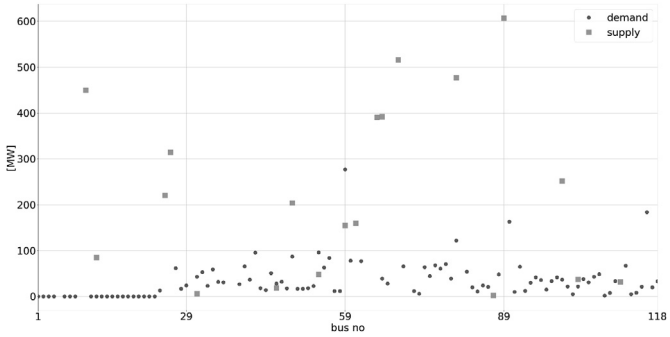

Fig. 15. Case II.1: demand and supply.

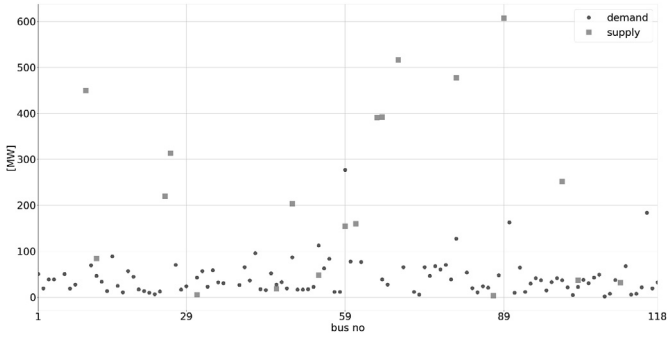

Fig. 16. Case II.2: demand and supply.

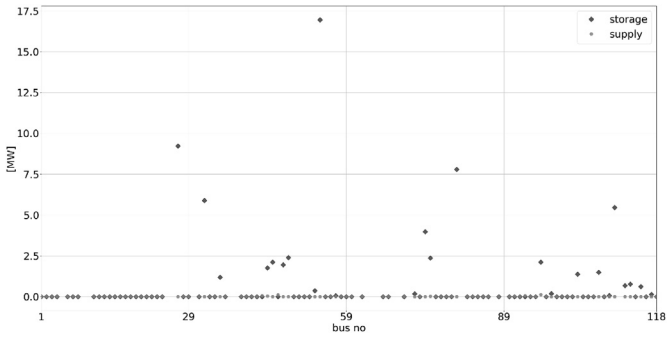

Fig. 17. Case II.1: storage and load shift.

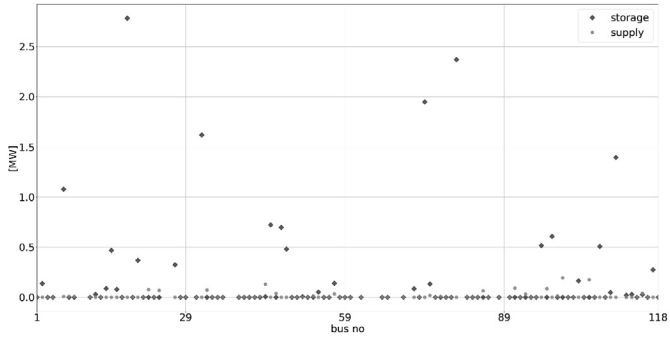

Fig. 18. Case II.2: storage and load shift.

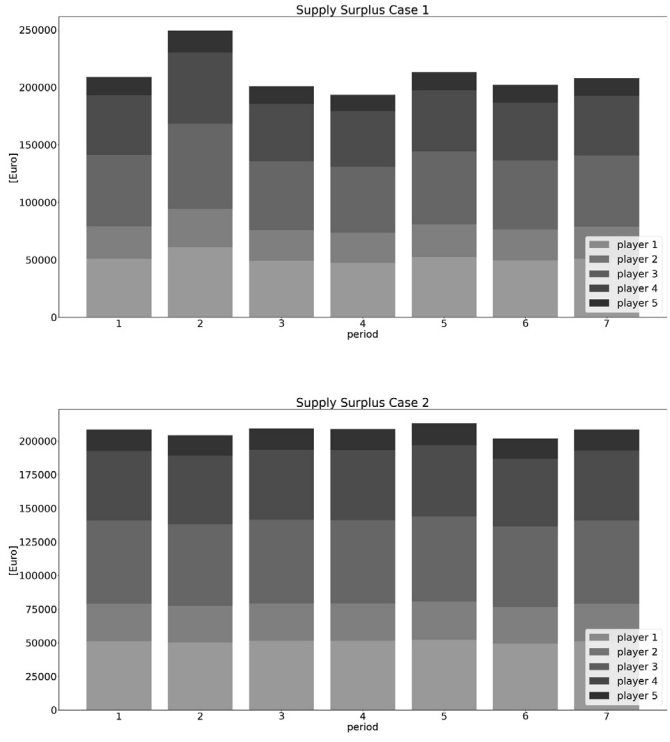

Fig. 19. Case II - supply surplus.

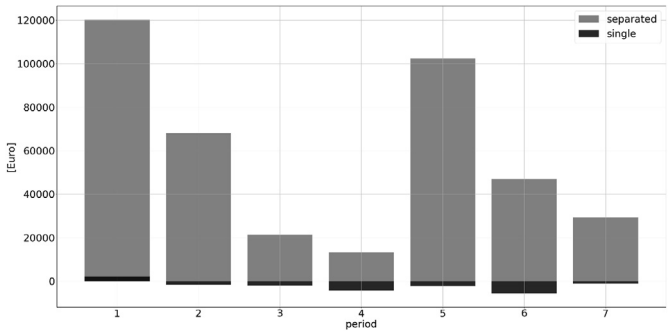

Fig. 20. Case II - welfare results.

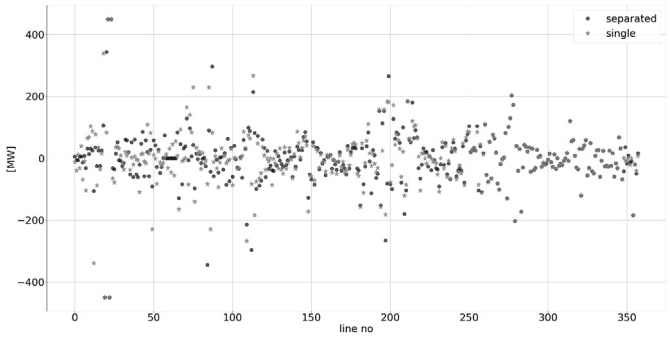

Fig. 21. Case II - System clearing: transmitted capacity in period 1.

higher for case 2. The reason for this can be attested to the aggregated player spanning a larger area of (i.e. the entire) transmission network, whereas the five separated players were localized in specific parts of the IEEE 118 bus network. Due to this distribution of units over the network, the actions of the single player would thus also affect the network to a greater extend.

The resulting case study thus displays intuitive results that are aligned with previous literature on the topic, whilst being capable to cover a medium-sized case study with a large-sized uncertainty set. In summary, the case study suggests that monopolization on the demand side in a system under distributed storage and flexible loads can lead to lower market clearing prices in Cournot competition. In this case, the winners of this monopolization would be the agents holding this inter-temporal capacity and subsequently the consumers profiting from a more efficient utilization of such. The losers in such an monopolization would however be the grid operators as such a monopolization could affect required balancing capacities.

Nonetheless, it has to be noted that the given case study was intended to provide an intuitive example of application of the proposed framework and should not be utilized for supporting real-world policy changes.

8. Conclusion

This paper presents a novel, multi-period framework to solve dynamic competition problems amongst Virtual Power Plants consisting of aggregate flexible demand, distributed generation and/or generation units. Neural networks are utilized as approximations for players on the demand and supply side. An iterative game is played between those approximations in order to converge towards an equilibrium under a vast uncertainty space. In addition, a DC-Optimal Power Flow transformation is proposed in order to yield network results from an area/zonal market clearing. The presented framework is aimed to represent liberalized electricity markets found in Europe that traditionally show a separation between market clearing and system operations planning.

Further, a proof of the model finding a Nash equilibrium even under high uncertainty (and thus unreliable approximations of future periods) is provided. The presented case studies on an extended version of the IEEE 30 and 118 bus test systems with an uncertainty space of over 10^{91} different possible scenarios demonstrates the capabilities of the model converging to an equilibrium solution.

The case studies also show the need for similar multi-period market clearing models. An example is given by a Virtual Power Plant consisting of generation capacity, demand shift capacity and storage capacities. In the single-period 30 bus case, this aggregator does not manage to yield a profit, as the demand of the units it manages exceed its generation capacity. Utilizing its storage and demand shift, however, this Virtual Power Plant manages to yield a profit over the periods.

A similar mechanism is suggested by the larger case-study, which shows that Virtual Power Plants holding monopolies on demand capacities utilize this market power in order to decrease the system prices, thus decreasing their cost of fulfilling the demand.

The proposed model framework is shown to be versatile, as an example it is pointed out that games on symmetric as well as asymmetric information are similarly implementable. In addition, the computational efficiency is highlighted: respective players' decision problems can be solved for in parallel, with the approximations being able to be trained in parallel as well.

In case of stage-wise independence, the periodical problems can be parallelized as well. This is also shown in the second case study, which shows the problems converging for 10 players competing over 7 time periods, i.e. solving 70 problems in parallel.

In addition to this computational aspect, the provided example also reduces the requirements of user assumptions such as end inventory levels of storage or approximations of the distributions, as the framework is trained iteratively on samples and thus can also cope with distributions directly, where traditional models instead required approximations of uncertainty in form of e.g. scenario trees.

One downside of the model, however, is the inaccessibility of results for visual presentation due to high dimensionality. In the provided paper the results for period 1 were presented as Monte Carlo analysis of the networks, which however show higher

dimensions than the 1000 scenarios bootstrapped this way. This problem is an inherent issue of utilizing neural networks and could increase the difficulty of trouble-shooting models and also have an effect on acceptance of advice to decision makers that is based on such a framework. This issue of data visualization, however, is an active topic of research and future developments in this field would thus positively affect the importance of the here proposed model framework.

Declaration of competing interest

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.

Appendix A. Policy Nash Equilibrium

Theorem 1 (Policy Nash Equilibrium). *The deterministic policy Nash equilibrium (4) is equal to the stochastic Nash Equilibrium (3) if every policy yields decisions* $\pi_{i,t} \in \pi_j^*(\xi_t)$ *representing the optimal player decisions, i.e.* $q_{i,t}^{S*} = \pi_{i,t}$

- *without requiring a perfect fit of the policy approximation, i.e.* $q_{i,t}^{S} \approx \pi_{i,t}$.
- *regardless of the future profit expectation, i.e. the quality of the critic function approximation* φ.

Proof.For the single period case, state equation (5c) can be reformulated as a function $f^{\text{state}}(s,l) = 0$ and market clearing condition (5f) as a function $f^{\text{market}}(q^S) = 0$. Denoting the profits in a single period as Π' allows for a reformulation of the periodic decision problem (5):

$$\begin{aligned} &\Pi'_{j,t}\left(q_{i,t}^* \forall i \in I\right) + \varphi_{j,t+1}\left(s_{i,t}^*, l_{i,t}^* \forall i \in I_j\right) \\ &\quad f^{\text{market}}\left(q^S\right) = 0, \qquad f^{\text{state}}\left(s,l,q^D\right) = 0 \\ &\quad \forall q_i \in \mathbb{O}_i^q(\xi_t) \qquad \forall i \in I_j \\ &\quad \forall s_i \in \mathbb{O}_i^s(\xi_t) \qquad \forall i \in I_j \\ &\quad \forall l_i \in \mathbb{O}_i^s(\xi_t) \qquad \forall i \in I_j \end{aligned} \tag{A.1}$$

As the feasible space of market clearing decisions and state decisions is only influenced by uncertainties and not any other decisions, the state problem can be decoupled from the market clearing problem.

This means that the following reformulation of the objective function (5a) holds for any state decisions:

$$\Pi'^*_{j,t}(\xi_t) = \Pi^t_{j,t}\left(\xi_t, q_{i,t}^* \forall i \in I\right) \tag{A.2}$$

Thus follows, regardless of the outcome of the uncertainty in the following periods and the quality of the future profit approximation $\varphi_{j,t+1}$, the optimal market clearing decisions are dependent only on the uncertainty in period t. As per definition, this uncertainty ξ_t becomes known in the current period. This makes the market clearing problem deterministic.

Assuming the critic is trained sufficiently to allow yielding policy decisions $\pi_{i,t} \in \pi_j^*(\xi_t)$ that adequately represent $q_{i,t}^{S*} = \pi_{i,t}$ results in the formulation of:

$$\Pi'_{j,t}\left(\xi_t, q_{i,t}^* \forall i \in I\right) = \Pi^t_{j,t}(\xi_t, \pi_{i,t} \forall i \in I) \tag{A.3}$$

This is valid even in cases where the approximation only fits the optimal decisions, i.e.:

$$q_{i,t} \neq \pi_{i,t} \text{ where } q_{i,t} \neq q_{i,t}^* \forall i \in I, t \in T \tag{A.4}$$

Thus it can be stated that finding the Policy Nash equilibrium (4) not only is a deterministic problem but also an accurate representation even for an imperfect approximation of the actor.

Compared to traditional implementations of actor-critic methods such as Ref. [46], the here presented problem fully knows the model(/environment). This disconnection and the disconnection of state and market clearing variables allows training the actor and critic individually without requiring reliance on the policy gradient theorem in Refs. [47] and thus increasing computation speed by being able to train critic and actor in parallel instead of in succession.

In addition and similar to Ref. [39] this proof shows that the actor can be trained from results obtained by the market clearing problem (5f) and will converge towards a Nash Equilibrium in case the fit of the decisions yielded by the actor network converge towards the market clearing decisions that they are being trained on, a problem that designs such as 'deep Q networks' have proven themselves capable of [48]. This is supported by the decision problem for a single period and player being known, replacing exploration in space with unknown boundaries (such as epsilon-greedy approaches) or policy gradient techniques (based on the so-called 'reinforcement trick') through a non-linear optimization problem over the space defined by Eq. (A.1) [32].

Appendix B. Neural Network Topology

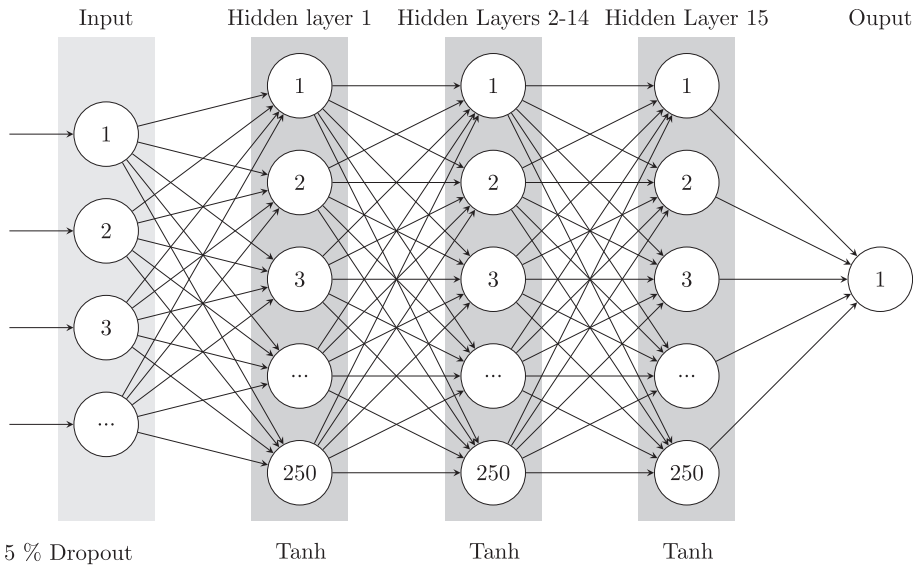

Fig. B.22. Neural Network Architecture

The chosen topology for the function approximations used in the presented case study is shown in Fig.B21. Each critic and actor was a neural network consisting of 15 layers consisting of linear layers with a size of 250 wrapped in hyperbolic tangent activation functions, a dense layer to merge the networks and a 5% dropout layer before the first layer. The utilized optimizer was 'Adam' [49].

Appendix C. Acknowledgement

The author would like to thank the project consortium of the SINTEF research project "Modeling Flexible Resources in Smart Distribution Grid - ModFlex" (255209/E20) and the Norwegian Research Council for supporting this work.

References

[1] Commission européenne and Direction générale de la mobilité et des transports, *EU energy, transport and GHG emmissions: trends to 2050 : reference scenario 2016*. Luxembourg: Office for official publications of the european communities, 2016, oCLC: 960351415.

[2] M. P. Moghaddam, A. Abdollahi, and M. Rashidinejad, "Flexible demand response programs modeling in competitive electricity markets," Applied Energy, vol. 88, no. 9, pp. 3257–3269, Sep. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261911001516.

[3] Gong X, De Paola A, Angeli D, Strbac G. A game-theoretic approach for price-based coordination of flexible devices operating in integrated energy-reserve markets. Energy Dec. 2019;189:116153 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544219318481.

[4] J. L. Mathieu, M. Kamgarpour, J. Lygeros, G. Andersson, and D. S. Callaway, "Arbitraging Intraday Wholesale Energy Market Prices With Aggregations of Thermostatic Loads," IEEE Transactions on Power Systems, vol. 30, no. 2, pp. 763–772, Mar. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/6866265/.

[5] Bruninx K, Pandzic H, Le Cadre H, Delarue E. On the Interaction between Aggregators, Electricity Markets and Residential Demand Response Providers. 1–1 IEEE Transactions on Power Systems 2019 [Online]. Available: https://ieeexplore.ieee.org/document/8848616/.

[6] Nekouei E, Alpcan T, Chattopadhyay D. Game-Theoretic Frameworks for Demand Response in Electricity Markets. IEEE Transactions on Smart Grid 2015;6(2):748–58 [Online]. Available: http://ieeexplore.ieee.org/document/6965656/.

[7] S. M. Moghaddas Tafreshi and A. Saliminia Lahiji, "Long-Term Market Equilibrium in Smart Grid Paradigm With Introducing Demand Response Provider in Competition," IEEE Transactions on Smart Grid, vol. 6, no. 6, pp. 2794–2806, Nov. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7078923/.

[8] Saliminia Lahiji A, Moghaddas Tafreshi SM. Merchant transmission planning in smart paradigm with introducing demand response aggregator in wholesale market: Merchant Transmission Planning in Smart Paradigm. International Transactions on Electrical Energy Systems Oct. 2016;26(10):2148–71. https://doi.org/10.1002/etep.2196 [Online]. Available:.

[9] S. Hadayeghparast, A. SoltaniNejad Farsangi, and H. Shayanfar, "Day-ahead stochastic multi-objective economic/emission operational scheduling of a large scale virtual power plant," Energy, vol. 172, pp. 630–646, Apr. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544219301598.

[10] Tajeddini MA, Rahimi-Kian A, Soroudi A. Risk averse optimal operation of a virtual power plant using two stage stochastic programming. Energy Aug. 2014;73:958–67 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544214008093.

[11] Z. Tan, G. Wang, L. Ju, Q. Tan, and W. Yang, "Application of CVaR risk aversion approach in the dynamical scheduling optimization model for virtual power plant connected with wind-photovoltaic-energy storage system with uncertainties and demand response," Energy, vol. 124, pp. 198–213, Apr. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544217302372.

[12] Löschenbrand M, Wei W, Liu F. Hydro-thermal power market equilibrium with price-making hydropower producers. Energy Dec. 2018;164:377–89 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544218316980.

[13] D. Bertsekas, Reinforcement Learning and Optimal Control. Massachusetts: Athena Scientific, 2019.

[14] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," Applied Energy, vol. 235, pp. 1072–1089, Feb. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261918317082.

[15] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, "Game Theory and Multi-agent Reinforcement Learning," in *Reinforcement Learning*, M. Wiering and M. van Otterlo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 12, pp. 441–470, series Title: Adaptation, Learning, and Optimization.

[16] E. Moiseeva and M. R. Hesamzadeh, "Bayesian and Robust Nash Equilibria in Hydrodominated Systems Under Uncertainty," IEEE Transactions on Sustainable Energy, vol. 9, no. 2, pp. 818-830, Apr. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8064663/.

[17] Z. Ni and S. Paul, "A Multistage Game in Smart Grid Security: A Reinforcement Learning Solution," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 9, pp. 2684–2695, Sep. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8603817/.

[18] M. Löschenbrand, "Finding multiple Nash equilibria via machine learning-supported Gröbner bases," *European Journal of Operational Research*, p. S0377221720300783, Jan. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0377221720300783.

[19] J. Heinrich and D. Silver, "Deep Reinforcement Learning from Self-Play in Imperfect-Information Games," arXiv:1603.01121 [cs], Mar. 2016, arXiv: 1603.01121. [Online]. Available: http://arxiv.org/abs/1603.01121.

[20] Ye Y, Qiu D, Li J, Strbac G. Multi-Period and Multi-Spatial Equilibrium Analysis in Imperfect Electricity Markets: A Novel Multi-Agent Deep Reinforcement Learning Approach. 130 515–130 529 IEEE Access 2019;7 [Online]. Available:

https://ieeexplore.ieee.org/document/8826539/.
[21] Zhang Z, Zhang D, Qiu RC. Deep reinforcement learning for power system: An overview. CSEE Journal of Power and Energy Systems 2019.
[22] Liang Y, Guo C, Ding Z, Hua H. Agent-Based Modeling in Electricity Market Using Deep Deterministic Policy Gradient Algorithm. IEEE Transactions on Power Systems 2020:1–13 [Online]. Available: https://ieeexplore.ieee.org/document/9106862/.
[23] Ye Y, Qiu D, Sun M, Papadaskalopoulos D, Strbac G. Deep Reinforcement Learning for Strategic Bidding in Electricity Markets. IEEE Transactions on Smart Grid Mar. 2020;11(2):1343–55 [Online]. Available: https://ieeexplore.ieee.org/document/8805177/.
[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017, p. 4.
[25] Powell MJD. Direct search algorithms for optimization calculations. Acta Numerica 1998;7:287–336.
[26] F. Biscani, D. Izzo, and C. H. Yam, "A Global Optimisation Toolbox for Massively Parallel Engineering Optimisation," arXiv:1004.3824 [cs, math], Apr. 2010, arXiv: 1004.3824. [Online]. Available: http://arxiv.org/abs/1004.3824.
[27] W. E. Hart, C. Watson, D. L. Woodruff, G. A. Hackebeil, B. L. Nicholson, and J. D. Siirola, Pyomo-optimization modeling in python, Berlin, 2017, vol. 67.
[28] Baldick R, Grant R, Kahn E. Theory and Application of Linear Supply Function Equilibrium in Electricity Markets. Journal of Regulatory Economics Mar. 2004;25(2):143–67 [Online]. Available: http://link.springer.com/10.1023/B:REGE.0000012287.80449.97.
[29] Day C, Hobbs B, Pang Jong-Shi. Oligopolistic competition in power networks: a conjectured supply function approach. IEEE Transactions on Power Systems Aug. 2002;17(3):597–607 [Online]. Available: http://ieeexplore.ieee.org/document/1033699/.
[30] Shapiro A. Analysis of stochastic dual dynamic programming method. European Journal of Operational Research Feb. 2011;209(1):63–72 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0377221710005448.
[31] D. Bertsekas, Dynamic Programming and Optimal Control, 4th ed. Massachusetts: Athena Scientific, 2012, vol. 2, no. 2.
[32] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. MIT Press, 2018.
[33] Ventosa M, Baíello A, Ramos A, Rivier M. Electricity market modeling trends. Energy Policy May 2005;33(7):897–913 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0301421503003161.
[34] S. A. Gabriel, A. J. Conejo, J. D. Hobbs, and C. Ruiz, Complementarity modeling in energy markets. Springer Science & Business Media, 2012, vol. 180.
[35] PandÅ¼iÄ‡ H, Morales JM, Conejo AJ, Kuzle I. Offering model for a virtual power plant based on stochastic programming. Applied Energy May 2013;105:282–92 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261913000044.
[36] Kardakos EG, Simoglou CK, Bakirtzis AG. Optimal Offering Strategy of a Virtual Power Plant: A Stochastic Bi-Level Approach. 1–1 IEEE Transactions on Smart Grid 2015 [Online]. Available: http://ieeexplore.ieee.org/document/7095586/.
[37] Wei W, Liu F, Mei S. Charging Strategies of EV Aggregator Under Renewable Generation and Congestion: A Normalized Nash Equilibrium Approach. IEEE Transactions on Smart Grid May 2016;7(3):1630–41 [Online]. Available: http://ieeexplore.ieee.org/document/7283652/.
[38] Moiseeva E, Hesamzadeh MR, Biggar DR. Exercise of Market Power on Ramp Rate in Wind-Integrated Power Systems. IEEE Transactions on Power Systems May 2015;30(3):1614–23 [Online]. Available: http://ieeexplore.ieee.org/document/6914628/.
[39] Hu J, Wellman MP. Nash Q-Learning for General-Sum Stochastic Games. Journal of Machine Learning 2003;4:1039–69.
[40] K. Arulkumaran, A. Cully, and J. Togelius, "AlphaStar: An Evolutionary Computation Perspective," Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO '19, pp. 314–315, 2019, arXiv: 1902.01724. [Online]. Available: http://arxiv.org/abs/1902.01724.
[41] M. Lanctot, V. Zambaldi, and A. Lazaridou, "A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning," Advances in Neural Information Processing Systems, pp. 4190 – 4203, 2017.
[42] M. Löschenbrand and M. Korpås, "Multiple Nash Equilibria in Electricity Markets With Price-Making Hydrothermal Producers," IEEE Transactions on Power Systems, vol. 34, no. 1, pp. 422–431, Jan. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8417933/.
[43] Contreras J, Candiles O. A Cobweb Bidding Model for Competitive Electricity Markets. IEEE TRANSACTIONS ON POWER SYSTEMS 2002;17(1):6.
[44] I. B. Sperstad, S. H. Jakobsen, and O. Gjerde, "Modelling of corrective actions in power system reliability analysis," in *2015 IEEE Eindhoven PowerTech*. Eindhoven, Netherlands: IEEE, Jun. 2015, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7232453/.
[45] L. Roald, S. Misra, T. Krause, and G. Andersson, "Corrective Control to Handle Forecast Uncertainty: A Chance Constrained Optimal Power Flow," arXiv: 1609.02194 [math], Sep. 2016, arXiv: 1609.02194. [Online]. Available: http://arxiv.org/abs/1609.02194.
[46] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms," 2014.
[47] Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. Advances in neural information processing systems 2000:1057–63.
[48] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529.
[49] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv: 1412.6980 [cs], Dec. 2014, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980.