Delft University of Technology

Non-stationarity in multiagent reinforcement learning in electricity market simulation

Renshaw-Whitman, Charles; Zobernig, Viktor; Cremer, Jochen L.; de Vries, Laurens

DOI
10.1016/j.epsr.2024.110712
Publication date
2024
Document Version
Final published version
Published in
Electric Power Systems Research

Citation (APA)
Renshaw-Whitman, C., Zobernig, V., Cremer, J. L., & de Vries, L. (2024). Non-stationarity in multiagent reinforcement learning in electricity market simulation. *Electric Power Systems Research*, *235*, Article 110712. https://doi.org/10.1016/j.epsr.2024.110712

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Electric Power Systems Research 235 (2024) 110712

Contents lists available at ScienceDirect

Electric Power Systems Research

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

Non-stationarity in multiagent reinforcement learning in electricity market simulation

Charles Renshaw-Whitman [a,*], Viktor Zobernig [b,a], Jochen L. Cremer [b,a], Laurens de Vries [a]

[a] Delft University of Technology, Delft, Netherlands
[b] AIT Austrian Institute of Technology, Vienna, Austria

ARTICLE INFO

Keywords:
Deep learning
Game theory
Market-design
Market simulation
Reinforcement learning

ABSTRACT

The design of electricity markets may be facilitated by simulating actors' behaviors. Recent studies model human decision-makers within markets as agents which learn strategies that maximize expected profits. This work investigates the problem of 'non-stationarity' in the context of market simulations, a problem with the learning-algorithms used by such studies which results in agents behaving irrationally, thus limiting the studies' applicability to real-world strategic behavior. Isolating the source of the problem for a day-ahead electricity market, this paper proposes methods which meliorate this problem in simple test-cases, and proves requirements under which 'centralized-training, decentralized-execution' value-learning methods will converge to correct behavior in general. Subsequently, this paper proposes a framework for 'adversarial market design' that includes the market-designer as an agent. This allows the optimization of market-designs subject to possibly strategic behavior of participating firms — in turn enabling the automated selection of the optimal market from any set of markets.

1. Introduction

The simulation of participants' behavior in electricity markets is an important tool for the effective design of electricity markets. The increasing prevalence of renewables will require increasingly complex, interlocking markets in order to ensure the effective procurement of the concomitant auxiliary services which prevent violation of the grid's technical operating constraints (see, e.g., [1,2]). However, as this complexity increases, so too does the number and complexity of bids available to market participants, making it progressively more difficult to ensure that such market designs align participants' incentives to be conducive to grid security and consumer satisfaction. Simulating participants' actions provides an alternative to designing solely from first principles, but more complex and interlocking markets entail an exponentially increasing number of bidding-strategies. Market designers are thus confronted with the dual task of determining actors' strategically optimal actions and simultaneously designing to mitigate the associated possibility for abuse of market power.

Conventional methods for analyzing strategic behavior of diverse market players have utilized optimization-algorithms subject to game-theoretic equilibria constraints [3]. Such algorithms have the advantage that they can exactly determine actors' optimal behavior using well-developed convex optimization techniques. However, they face challenges in modeling sequential problems (e.g., those involving the submission of several consecutive bids) and non-convex constraints such as those found in the unit commitment problem [3]. Thus, this constrains multi-agent analyses driven by such methods, thereby complicating the realistic evaluation of market designs.

Notably, reinforcement learning (RL) methods, well-suited to sequential decision-problems, are increasingly used for electricity market modeling in place of conventional optimization algorithms (e.g., [4–6]) - in part spurred on by the achievements of 'Deep RL' such as [7–9]. RL methods have the distinct advantage of being unconstrained by limitations such as convexity or other conventional optimization constraints, making their applicability more versatile.

While RL can model the decision-making of a single rational actor, multiple such actors are needed to constitute realistic market dynamics. Our work centers on the exploration of multi-agent RL (MARL) methods, to advance market model analyses.

Regarding computational tractability, MARL methods, on account of being *heuristic* optimization algorithms, may also be expected to scale to larger number of agents or higher dimensional search spaces more readily than their analytical compeers — compare the seemingly intractable superexponential complexity bounds given in [10] to the remarkable performance shown in, e.g., [11]. While MARL methods may show inferior performance on low-dimensional problems, they can be used to find approximate solutions even when the search-space becomes intractably large or complex for any analytical algorithm,

* Corresponding author.
E-mail address: CharlesRW@protonmail.com (C. Renshaw-Whitman).

https://doi.org/10.1016/j.epsr.2024.110712
Received 1 October 2023; Received in revised form 14 March 2024; Accepted 17 June 2024
Available online 2 July 2024
0378-7796/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

as occurs in, e.g., complex sequential decision-making problems like developing a bidding-strategy with intricate conditionals. In the present work, we consider cases with between one and five agents, focusing principally on the case with only two competing firms. The required computation per update is linear in the number of agents, while the number of samples required to reach convergence is difficult to predict in advance.

In this study, we highlight a critical flaw in commonly-used MARL methods, which can result in incorrect equilibria determinations. This flaw is analogous to the concept of 'relative overgeneralization' discussed in the literature in cooperative RL [12], extended to general-sum games. To analyze this problem, we show how the convergence theorems for traditional Q-learning become invalid in the presence of multiple-agents. This is concretized in a case-study in which convergence-failure is identified in a simple day-ahead market setting using both tabular and deep RL methodologies. Our results indicate that agents converge to strategies inconsistent with known-optimal oligopoly behavior.

For the purposes of this paper, we focus our attention on the popular deep deterministic policy-gradient algorithm (DDPG, [13]) and its multi-agent adaptation, multi-agent DDPG (MADDPG, [8]). We select these two algorithms because they differ only in the form of the learned value-function - the value-function being the function which learns to predict the long-term value of taking a particular action, conditional on having certain information. Our results show that MADDPG, a centralized learning algorithm, converges where the decentralized DDPG does not — and prove a theorem indicating centralization of the value-function is necessary and sufficient to achieve convergence.

Moreover, we broaden the scope of these methodologies to include market *design* through our novel approach, dubbed "adversarial market-design", treating the market-designer as an agent. Framing market-design as an RL problem assures convergence to Nash equilibria, based on the theorems above. Adversarial market-design allows streamlining of the market-design process by incorporating it within the RL training procedure. In this framework, the market-design agent is incentivized to design a market for which firm-agents' optimal behavior is aligned with social welfare. In the present work we consider only power-producing agents for simplicity, though power-purchasers or other types of agents may be incorporated straightforwardly.

Thus, the core research contributions of this work are (i) an explication of the conditions under which MARL methods will converge in electricity markets (along with accompanying theorems) and (ii) the introduction of the adversarial market-design framework. We use the temporal difference error (TD-error) as a measure of convergence, and further define a measure of distance from equilibrium. This metric, the optimality-deficit, is used to assess when agents have converged to a correct Nash equilibrium. Moreover, through the application of adversarial market design, we aim to demonstrate the potential of MARL methods to enable the efficient design and evaluation of novel electricity market structures.

The structure of the work is as follows: Section 2 introduces the relevant background in game theory and RL, along with a brief overview of the DDPG and MADDPG algorithms and the adversarial market-design framework. Section 3 discusses the mathematical form of the non-stationarity problem, and introduces some relevant tools for investigation. Section 4 introduces the simulations to be carried out and illustrates their results. Section 5 discusses the significance of these results, followed by a concluding recapitulation and suggestions for future work.

2. Reinforcement learning for electricity markets

2.1. Reinforcement learning

RL is a paradigm of machine learning in which an agent attempts to learn a sequential decision-rule that maximizes a given reward function. It operates within the Markov Decision Process (MDP) framework, where actions taken in each state drive the agent's transition to the next state based on the chosen action to find the optimal policy. A popular textbook on the subject is [14], while a review in the context of electricity markets is presented in [15] — unless otherwise noted, the notation used here is as in the former.

We will be mostly concerned with the family of reinforcement learning methods, called 'value-learning' methods, which seek to learn a "Q-function". For our purposes, the Q-function predicts the profit a firm would obtain by taking a particular action, a (an example of an action might be the submission a particular bid). The firm's choice of bidding-strategy is known as a policy, denoted π; this strategy may be contingent on an observed state, s, such as a firm's marginal cost of generation, or the current demand-curve. RL algorithms learn a policy π which maximizes this Q-value. A number of such algorithms are detailed in [14]. While RL readily handles sequential decision-making problems, we concern ourselves only with single-timestep markets (i.e., those which do not require making a sequence of decisions, but one only) for simplicity.

2.2. MARL algorithms: DDPG and MADDPG

In this work we consider first 'tabular' and then 'continuous' environments. In a tabular environment, agents choose from a finite set of actions (e.g., a choice from a list of possible offer-prices). In a continuous environment, they submit some continuously varying quantities (e.g., the free choice of a price, possibly restricted to some range). We focus on environments in which there is only a single-timestep — a single round of bid submission and subsequent market-clearing. We will briefly examine a case with hidden state-variables, but restrict the mathematical discussion below to the stateless case for simplicity.

The tabular case is included as it is the most straightforward. Further, convergence theorems developed for single-agent reinforcement-learning typically only apply to tabular environments. This work uses a tabular-learning algorithm known as 'independent Q-learning', in which agents learn precisely as in the single-agent setting (and thus, taking no account of other agents' actions). By failing to account for others' actions, independent agents may be said to consider other agents as 'part of the environment'. For electricity markets, this amounts to an inability to distinguish between other firms' bids effecting the market and a general change in market conditions (e.g., in demand).

For independent learning algorithms, if multiple agents train simultaneously, each agent faces, in effect, an environment which changes as other agents' policies change. This violates the assumptions of a Markov Decision Process (MDP, the mathematical structure underlying RL problems), and thus causes difficulty in arriving at equilibria. This problem is known as 'non-stationarity' from the apparent non-stationarity (in the probabilistic sense) of the environment [16].

In addition to the tabular case, this work also considers the case of a continuous environment. In our case, the continuous environment is distinguished from the tabular one by the use of bids in which quantity and price are allowed to be any positive number, rather than a choice from some pre-determined set of bid pairs. This is, of course, more faithful to the real-world bidding process. When the number of possible inputs are no longer finite, it is necessary to learn a parameterized function which outputs an action — for this we use neural networks. In our investigations here, we use 'actor–critic' methods, in which an agent is assigned two neural networks, the first representing the Q-function, and the second the agent's policy π. The Q-function is trained based on information from the environment, while the policy π is trained to maximize the Q-function.

Simplified for the stateless, single-timestep, full-information case we consider, the loss-functions used are:

$$\begin{aligned} \mathcal{L}^{Q}_{\text{DDPG}}\left[\phi^i\right] &= \mathbb{E}\left[\left(r^i(A^{-i},A^i) - Q_{\phi^i}(A^i)\right)^2\right] \\ \mathcal{L}^{Q}_{\text{MADDPG}}\left[\phi^i\right] &= \mathbb{E}\left[\left(r^i(A^{-i},A^i) - Q_{\phi^i}(A^{-i},A^i)\right)^2\right] \\ \mathcal{L}^{\pi}\left[\theta^i\right] &= \mathbb{E}\left[Q(\pi_{\theta^i})\right] \end{aligned} \tag{1}$$

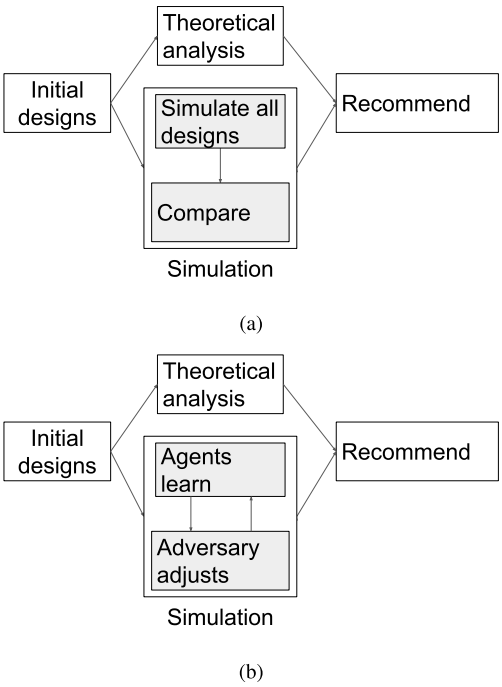


Fig. 1. Comparison of the use of simulation in traditional and in adversarial market-design. (a) In conventional market-design a designer uses simulations to select amongst a family of designs by simulating each and manually comparing the results. (b) In adversarial market-design, a market-designer-agent ('adversary') is trained to select the best design — agents learn to behave strategically in each new environment as the adversary adjusts.

Parameters not relevant to the particular loss function are suppressed — θ and ϕ represent the policy and value-function networks, respectively.

2.3. Nash-equilibria and electricity markets

Market environments, such as the electricity markets, may be modeled by assuming different firms act as economically-rational expected-profit-maximizers (or more general expected-utility-maximizers, if, e.g., firms are risk-averse). RL has the additional utility that, by specifying a different reward function, one can model other incentive structures, like risk-aversion or short-termism (each of which may be modeled using time-discounted rewards). If each agent has a family of actions $\mathcal{A}^i$ (and calling the Cartesian product $\mathcal{A} = \bigtimes_i \mathcal{A}^i$), a "game" between N agents is any function $R : \mathcal{A} \to \mathbb{R}^N$ — it assigns for each possible joint action a 'reward' to each agent. Agent i's reward is R^i.

In game-theory, a Nash-equilibrium is a joint action a^* such that no agent may unilaterally profitably change their action:

$$\forall i \ \forall a^i \in \mathcal{A}^i, R^i(a^{-i,*}, a^i) \leqslant R^i(a^{-i,*}, a^{i*}) \tag{2}$$

Deterministic games may have zero, one, or no Nash equilibria; when agents play probabilistically (i.e., they select probability-distributions over each $\mathcal{A}^i$, called 'mixed-strategies'), it is a famous theorem that all games then have at least one Nash-equilibrium [16]. We will refer briefly to cases where agents have access to hidden-information; both the case of hidden-information and mixed-strategies have analogous definitions for Nash-equilibria as in Eq. (2), with expectation-values taken over appropriate probability-distributions.

2.4. Adversarial market-design

"Adversarial market-design" aims to determine the best market-design from a set of possible designs. It models the market-designer as an RL agent whose 'actions' are the selection of a market design, and whose 'reward' is some social-welfare function. For example, suppose the market-designer-agent selects as its action some parameter-vector θ. These parameters describe the form of the market, such as a price-cap, or the shape of some administratively-set demand-curve — the aim is to find the optimal value of these parameters. Then the firm-agents learn a policy of the form $\pi^i(a^i|\theta)$, and receive a reward $R^i_\theta(a^i, a^{-i})$ — the profit obtained when the market, specified by θ, is cleared. The reward of the designer-agent is some social-welfare function $\mathrm{SW}(a^i, a^{-i}, \theta)$ (e.g., the sum of the consumer and producer surpluses). The market-design and firm-agents may act either simultaneously or in succession; our experiments below employ the former convention.

The significance of this paradigm is that, where a Nash equilibrium played by firm-agents informs us about possible strategic or malign behavior, a Nash equilibrium played by the firm-agents *and* the designer-agent represents a market-design selected to maximize social welfare *in spite of* strategic behavior. A schematic comparison of conventional and adversarial market-design is given in Fig. 1.

3. The non-stationarity problem

3.1. TD-error and optimality deficit

In seeking to examine when multi-agent reinforcement-learning algorithms will converge to a correct Nash-equilibrium, we must introduce a metric which describes training-convergence, and a metric which describes distance from Nash-equilibrium. We consider here only the full-information, single-clearing electricity market, though the concepts generalize readily to the full MARL problem with hidden information and multiple timesteps.

The TD-error is the difference between an agent's expected and received reward; the TD-error δ is defined for the Q-function's predicted profit, Q, and received reward, r, simply as

$$\delta = Q - r \tag{3}$$

If the TD-error goes to zero, agents obtain exactly the reward they predict, and if they are playing a simple argmax policy, their policy ceases to change as well (modulo exploration).

We define agent i's optimality-deficit λ^i as the difference between the expected reward an agent achieves with its current policy and the expected reward it would achieve by optimal play, holding fixed other agents' policies. That is, it quantifies how much profit a firm is leaving 'on the table' by bidding as it does now, fixing other firms' bidding-strategies. Mathematically, this is

$$\begin{aligned}\lambda^i = \max_a \ & \mathbb{E}\left[R^i(A^{-i}, a)|A^{-i} \sim \pi^{-i}\right] - \\ & \mathbb{E}\left[R^i(A^{-i}, A^i)|A^{-i} \sim \pi^{-i}, A^i \sim \pi^i\right]\end{aligned} \tag{4}$$

By definition, a Nash-equilibrium has all agents playing with a zero optimality deficit.

If a multi-agent simulation should exhibit a near zero TD-error for all agents, while not all agents show a similarly near zero optimality-deficit, it is clear that the simulation will have converged to a non-Nash-equilibrium, rendering the results not useful as models of strategic firm behavior.

3.2. Convergence theorems

Here we present theorems showing that, under certain conditions, MARL algorithms which use centralized training are guaranteed to converge. We denote the entries of a matrix with subscripts m, n, as in $Q^i_{t,mn}$; e_{mn} denotes the matrix which is 1 in the m, n-th entry, and 0 in all other entries. Proofs are provided in the Appendix.

Theorem 1. *Let Q_t^i denote agent i's estimated Q-function, after t iterations of learning, as a matrix whose entries correspond to possible action-pairs of all agents. Similarly, let R^i be a matrix denoting the actual reward obtained, and $f(Q^i, R_{mn}^k e_{mn})$ be an update rule which alters the Q-estimate upon observing a reward R_{mn}^i corresponding to the action pair (m, n). Then, if there exists a matrix-norm $\|\cdot\|$ such that always $\|Q_{t+1}^i - R^i\| \leqslant \|Q_t^i - R^i\|$, then the error $\|Q_t^i - R^i\|$ converges to a constant c as $t \to \infty$.*

Theorem 2. *Suppose the following three conditions hold:*

- *Seeing the (m, n)-action pair, the update rule alters only Q's m, n-th entry*
- *For each m, n there exists a ρ, $0 \leqslant \rho < 1$, $|Q_{t+1,mn}^i - R_{mn}^i| \leqslant \rho|Q_{t,mn}^i - R_{mn}^i|$*
- *All action-pairs (m, n) for which $Q_{mn}^i \neq R_{mn}^i$ occur with non-zero probability*

Then Q_t^i converges element-wise and almost-surely to R^i as $t \to \infty$ (i.e., each entry of Q converges to the corresponding entry in R with probability 1).

Corollary. *The convergence theorem stated above applies to stateful MDPs with multiple timesteps and hidden information if a Q-function over trajectories is used.*

In the independent-learning case (analogous to DDPG, as opposed to MADDPG), the Q-"table" is actually just a vector indexed by the agent's own action. In this case, the contraction property cannot be satisfied in general, so that the theorems above do not hold.

4. Case study

Here we outline the three classes of experiment we shall run: those with a discrete environment, those with a continuous environment (first in Bertrand competition, and then in Q, P-competition), and those regarding adversarial market design.

4.1. Experiment setup

For what follows, unless otherwise noted, all agents' marginal costs are $c = 40$USD/MWh, while demand (in USD) is given by $P_D(Q) = 500 - 2Q$. Clearing occurs in as-bid merit-order until the price at the quantity contracted equals the demand at that price. Exploration occurs for the first half of a run only, while learning occurs throughout. Agents are trained on one hundred thousand market-clearings.

All neural networks used in this work have two hidden layers of 300 nodes, with Layer-Norm and ReLU activations. The Adam optimizer was used. For exploration, the policy-generated act was added to zero-mean Gaussian noise with standard-deviation 0.3. All acts were scaled such that 1.0 corresponded to $Q_{\max}$ or MC. Rewards used for learning were scaled down by a factor of 1000. A replay buffer large enough to store one quarter of the total episodes for learning was used (storing in FIFO order); batches of size 128 were drawn at random from this buffer; learning updates occurred every 128 timesteps. These parameters were selected after a brief trial-and-error process, and found to be adequate to demonstrate the relevant phenomena.

The experiments were run on a personal laptop with a hybrid graphics-card setup, utilizing an Intel P630 and an Nvidia Quadro M2200 Mobile. The tabular- and continuous-environment experiments ran in under an hour (for the whole experiment including all runs and all configurations). The adversarial market-design experiment completed in under four hours (primarily because of the much larger number of configurations compared to the tabular- and continuous-environment experiments).

4.2. Tabular day-ahead market

In light of the theory presented previously, we investigate the circumstances under which independent learners will converge to an incorrect bidding strategy. To this end, we simulate a simple day-ahead market, described in detail below, with agents trained implementing simple tabular Q-learning. The optimality-deficit is computed, and examples are shown in which convergence occurs while agents maintain a non-zero optimality deficit.

4.2.1. Description of the environment

For the sake of clarity, we model a simple day-ahead electricity market. A single market-clearing is simulated — technical operation constraints are not considered. In cases with hidden information/states, each agent observes its own maximum generation capacity (either 50 or 100 MW) and marginal cost (either 20 or 50 USD/MWh), but not its opponents'. In cases without hidden information, the capacity and marginal cost of all agents are 50 MW and 20 USD/MWh, respectively

In each round, agents select a pair of actions (a, b) corresponding to the fraction of their total generation capacity to bid, and the fraction of their marginal cost to bid — i.e., the action selection (a, b) corresponds to a bid to sell aQ_{max}^i units of electricity at a price bP_{MC}^i. Allowed values of a and b are the same for all agents at all times, and are each drawn from a distinct discrete set. $Q_{\max}$ fractions were permitted to be $0.5,\ 0.75$, and 1.00. P_{MC} fractions were permitted to be $0.5, 1.56, 2.61, 3.67, 4.72, 5.78, 6.83, 7.89, 8.94$, and 10.00 (i.e., 10 linearly spaced points between 0.5 and 10).

The market clears in as-bid merit-order, with a uniform clearing price set to be the minimum price at which sufficient bids are accepted to satisfy demand at that same price level. In case of shortage, the clearing price is the willingness-to-pay of demand for the total offered quantity. Ties are resolved randomly for simplicity, and bids may be partially accepted (e.g., the market may procure, say, 15 MW of a 40 MW bid).

4.2.2. Training procedure

Agents were initialized with an empty Q-table. After every round, an agent learns from an observation. This learning takes the form of a 'soft update' with a learning rate of 0.99; the exploration technique used was epsilon-exploration [14].

4.2.3. Tabular day-ahead market results

Fig. 2 illustrates the TD-error and optimality-deficit associated with the tabular case, for each of the different environmental configurations. The two cases with a single learning agent ("1 Agent, 1 State" and "1 Agent, 4 States") show nearly zero TD-error and optimality-deficit, compared to both cases with multiple learning agents ("3 Agents, 1 State" and "3 Agents, 4 States"), which each show a positive optimality-deficit; only the case with hidden-information, "3 Agents, 4 States" shows a substantial TD-error. The "final" values shown are, in each case, the average over the most recent thousand timesteps, or the coterminous batches; within each run, the values are averaged over all learning agents. The tabular experiments indicate that independent learners are capable of learning to bid the correct monopoly price (the optimality-deficit is near zero in the monopoly case). However, they fail to behave strategically as soon as multiple firms are involved (the optimality-deficit is non-zero when there are multiple agents). This pathology is clearly related to non-stationarity, as the TD-error in all cases (except that with hidden-information, as expected) approaches zero while the optimality-deficit does not whenever there are multiple agents.

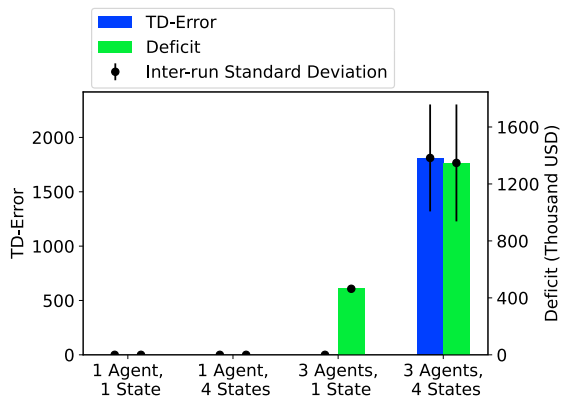


Fig. 2. Summary of the tabular experiment results. For each case, the final TD-error and optimality-deficit are shown; uncertainty bars represent the standard deviation over all three runs. Left vertical axis corresponds to the scale of the TD-error, and the right to the deficit.

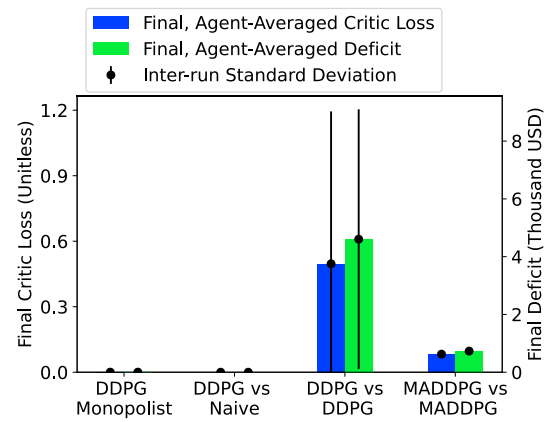


Fig. 3. Summary of the continuous environment experiment results for Case 1 (hypercompetitive, Bertrand competition). For each case, the final critic-loss and optimality-deficit are shown; uncertainty bars represent the standard deviation over all three runs. Left vertical axis corresponds to the scale of the TD-error, and the right to the deficit.

4.3. Extension to the continuous day-ahead market

4.3.1. Price-only (Bertrand), hypercompetitive

The first set of continuous experiments consists of testing these DeepRL algorithms in a continuous environment under Bertrand competition (i.e., in which agents bid a single positive real price, with the lowest-bidding agent fulfilling all demand) [17]. This experiment studies the connection between the tabular experiments — restricting to Bertrand competition allows the calculation of the best-response function, and thus the optimality-deficit; this is not readily done when agents are allowed to submit both quantity- and price-bids.

We are interested in verifying the adequate implementation of the DeepRL algorithms, and in illustrating the continued failure of independent learning due to non-stationarity. As in the tabular case, we examine a simple day-ahead market; we restrict the experiment to price-only bidding in the hypercompetitive regime (in which each agent can unilaterally meet all demand). This allows comparison of agent performance to the expected analytical solution of the Bertrand market discussed above. Negative quantities and negative prices are forbidden, though agents are expected to learn to bid above their marginal costs rather than being hard-coded to do so. Shortage is impossible as the Bertrand model assumes non-bindingness of capacity constraints.

We examine four Bertrand-competition scenarios, the results of which are summarized in Fig. 3: in the first, the DDPG agent is a monopolist (and so is expected to converge to the monopoly price); this provides evidence as to whether the learners have been implemented correctly. In the second, a DDPG agent competes with a naive marginal-cost bidder (the maximum profit is thus zero); this should illustrate that the non-stationarity problem does not occur when the other agents are fixed. In the third, two DDPG agents are pitted against one another in hopes of illustrating the failure to converge to a Nash equilibrium, analogous to the discrete case. Finally these two DDPG learners are replaced with MADDPG learners, which it is hypothesized will suffice to ensure that a correct equilibrium is learned.

4.3.2. Q,P bidding, competitive

The purpose of the second set of continuous-environment experiments is to illustrate that the same problems of non-stationarity occur as previously when agents participate in a more complex market. In this case, the action-space of the agents is extended to permit an agent to choose both the price and quantity of its bid. Further, the capacity of each agent is lowered to $Q_{\max} = 200$ MW so that no individual agent can meet all demand alone. Because both quantity and price are set independently, the analytical solution is fairly complicated compared to the Bertrand case; instead we report profits as a proxy for relative performance. This increase in complexity is meant to illustrate that DDPG learners are capable of operating in (slightly) more complex environments, while still failing when pitted against one another. In such a case, it is hypothesized that the centralized MADDPG will be capable of converging where independent-learning algorithms like DDPG fail to do so.

4.3.3. Continuous day-ahead market results

Fig. 3 summarizes the results for the continuous Bertrand environment. The two cases with a single learning agent ("DDPG Monopolist" and "DDPG vs. Naive") show near-zero critic-loss (which is functionally analogous to the TD-error) and optimality deficit, compared to the cases with multiple learning-agents ("DDPG vs. DDPG" and "MADDPG vs. MADDPG"). The MADDPG case shows much lower critic-loss and optimality-deficit than the DDPG case.

Fig. 4 shows the results for the continuous, Q,P-bidding environment. The "DDPG vs. DDPG" case shows a larger TD-error compared to the "DDPG Monopolist" and "MADDPG vs. MADDPG" cases. The agents in the MADDPG case obtain less profit on average than in the DDPG case. The uncertainty bars represent the standard deviation over all three runs.

The continuous Q, P-bidding cases show that the use of different algorithms can result in different profits being obtained — in particular, a market-designer wishing to use RL to examine this market would, seeing the greater profits of the DDPG case, incorrectly infer a much greater potential for strategic behavior. This illustrates the importance of correctly handling non-stationarity if one wishes to understand the market-design in question.

4.4. Adversarial market-design

4.4.1. Price-cap environment formulation

We consider a market-designer aiming to optimize the social-welfare function

$$\mathrm{SW} = 2\int_0^{Q^*} dq(\alpha P_D(q) - (1-\alpha)P_S(q)) + (1-2\alpha)P^*(Q^*) \tag{5}$$

that (dis-)privileges consumer welfare relative to producer welfare according to a factor α ($0 \leqslant \alpha \leqslant 1$). $\alpha = 1$ corresponds to completely preferring consumer-welfare, and $\alpha = 0$ to completely preferring producer-welfare.

For ease of reference, we call α the "adversariality", as it controls how strongly "adversarial" the designer-agent is to the firm-agents. Once the designer-agent sets the price-cap, the market is cleared as though the demand-curve were $P_{\mathrm{eff}}(Q) = \min(P_{\mathrm{cap}}, P_{\mathrm{D}}(Q))$, with P_{D} the uncapped demand-curve and P_{cap} the price cap.

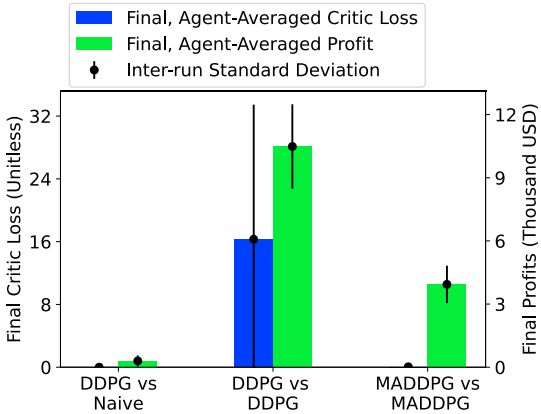


Fig. 4. Summary of the continuous environment experiment results for case 2 (competitive, Q,P-competition). For each case, the final critic-loss and profit are shown.

4.4.2. Description of the environment

This third set of experiments is conducted according to Cournot competition [17] — firm's submit quantity-bids only, with the priced determined according to the total quantity offered. The key difference is the introduction of an adversary-agent who sets a market price-cap. The reason for selecting Cournot competition is that it allows for ready computation of the best-response function, and requires fewer modifications to accommodate the adversarial price-cap than would the Bertrand solution. The goal of this experiment is to examine the performance of an adversary tasked with optimizing the market to maximize a certain social welfare-function, while learning agents *simultaneously* learn to behave strategically as the adversary adjusts the market-rules.

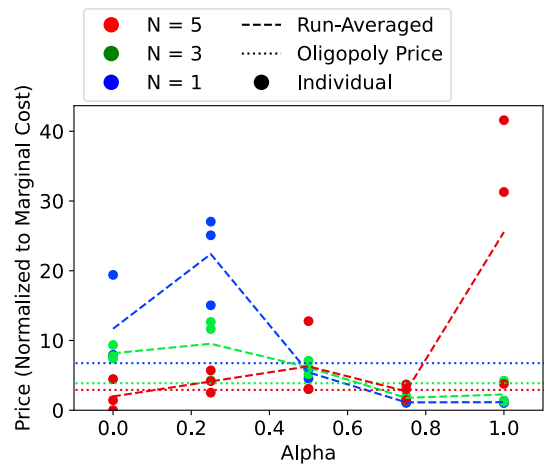
The adversary's action-space is a single continuous scalar, the price-cap; firms' actions are quantity-only bids of the same form as before. As previously, no agents observe any state-information — instead the channel for information to flow is the agents' critic functions. All agents use MADDPG to learn.

The purpose of this experiment is not, of course, to present policy-advice based on such a simplistic model; instead, it is hoped that the simplicity of the model will render the application and its limitations more legibly than a fully realistic environment. A real application of this technique would entail careful selection of market-designs such that they can be usefully represented as the output of a neural network, as well as a more thoroughgoing consideration of the appropriate state- and action-spaces.

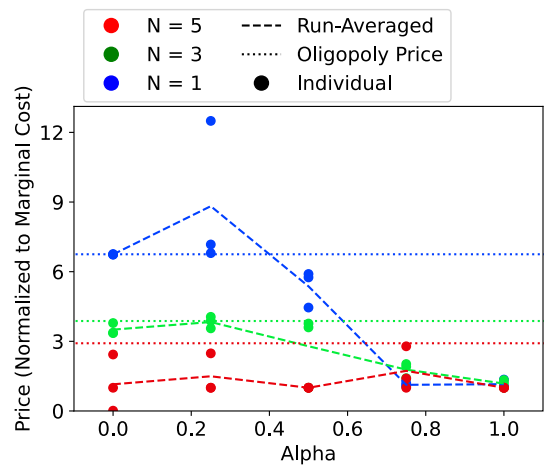
4.4.3. Adversarial market-design results

Rather than presenting the complicated best-response functions, we here offer a few remarks about the optimal behavior of the price-cap environment. First, in general, to the extent that they are free to choose their prices, the strategically acting firms should cause a deadweight loss. Clearly, when $\alpha = 1$, the optimal price cap is the marginal cost (plus an arbitrarily small real number) – this maximizes not only the consumer welfare, but the total social welfare for any $\alpha > 1/2$ – since lowering the cap in this regime converts producer welfare into a strictly greater amount of consumer welfare (which is also valued more). When $\alpha < 1/2$, the price cap can force the oligopolists to produce a quantity greater than the oligopoly quantity, but not less — thus the optimal cap in this region involves trading off (relatively more valuable) producer welfare to consumer welfare in an attempt to recoup by decreasing the deadweight loss.

Fig. 5 displays the results of the adversarial market-design experiments — for each of five values of the adversariality $\alpha = 0.00, 0.25, 0.50, 0.75, 1.00$, and each of $N = 1, 3, 5$ competing firms are displayed the final price-cap, and the final clearing-price.



(a)



(b)

Fig. 5. Adversarial market-design: parameter-sweep summary. Plots illustrating the behavior of the final clearing price and price cap. (a) final learned price-cap. (b) clearing-price.

These adversarial market-design experiments show that the simultaneous optimization of the market-design (the price cap) and the agents' strategic behavior yields outcomes showing the main patterns expected from the analytic solution — in particular, a high clearing price when the market-designer prefers producer welfare, decreasing to approach the producers' marginal-costs as this preference shifts to consumer-welfare. On the other hand, it is clear that there was substantial variation amongst the different scenarios, indicating that the simple setup used here may not be optimal for practical market-design.

5. Conclusion

This work has examined the role of non-stationarity on MARL simulations of rational agent behavior for market design. We demonstrate that independent-learning algorithms like DDPG are incapable, in general, of arriving at correct equilibria. We employed the TD-error as a convergence metric and a measure of the distance from equilibria to determine whether the agents converge to the correct Nash-equilibrium; conditions were proved under which CTDE algorithms are guaranteed to converge to Nash-equilibria. Finally, taking advantage of these convergence guarantees, we introduced the concept of adversarial market-design, and illustrated its use in a simple example.

Guaranteed convergence to true Nash-equilibria means that MARL market simulations may be used as a reliable source of evidence about a market-design's ability to mitigate abuses of market power. The framework of adversarial market-design provides a new, MARL-based method for automatic market-design optimization.

As this work focused primarily on methodology, little attention was paid to 'realism' of the simulations — one interesting direction of future work would be the examination of non-stationarity in more realistic electricity-markets. In a more practical context, provided that stability can be enhanced (e.g., by having the market-designer's action visible to the learners), the application of adversarial market design holds the potential to offer significant benefits in shaping new market rules, particularly in scenarios involving multiple interconnected markets. This approach may prove invaluable in assisting policymakers in making informed decisions concerning market regulations during periods of market crisis or transition.

CRediT authorship contribution statement

Charles Renshaw-Whitman: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Viktor Zobernig:** Supervision, Writing – review & editing. **Jochen L. Cremer:** Supervision, Writing – review & editing. **Laurens de Vries:** Supervision, Writing – review & editing.

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.

Data availability

No data was used for the research described in the article.

Appendix. Proof of convergence theorems

In what follows, we seek to prove sufficient conditions for convergence to a correct equilibrium; to make this simpler, we consider the discrete case with only two agents; the generalization to many agents is straightforward, while convergence theorems are impractical to prove for the continuous case in which Q-function updates rely heavily on the complicated neural network parameterization and specifics of the optimizer used for gradient-descent.

Proof of Theorem 1. First, we consider the centralized policy: Let Q^i_t be a matrix whose entry m, n represents the agent's Q-value estimate of the reward obtained for itself and its opponent taking actions a^m and a^n, respectively. Upon the action-pair (a^m, a^n) occurring, an update rule dictates that agent k updates as

$$Q^i_{t+1} = f(Q^i_t, R^i_{mn} e_{mn}) \tag{6}$$

(where e_{mn} is the matrix which is 1 in entry m, n and 0 in all others) For instance, the soft-update rule has the form $f(A, B) = A(1 - e_{mn}) + e_{mn}(\rho A + (1 - \rho))B$. This is simply to say that the update rule updates only the entry for the action-pair which is observed, which updates to a weighted average of the estimated and observed value.

The following condition suffices to show that an update rule for Q is non-increasing: if we have for some matrix-norm $\|\cdot\|$ that for any observation (a^i, a^j)

$$\|Q^i_{t+1} - R^i\| \leqslant \|Q^i_t - R^i\| \tag{7}$$

then the update-rule is non-increasing and bounded above by $\|Q_0 - R\|$, and bounded below by 0 - thus the sequence converges (not necessarily to 0).

Proof of Theorem 2. In addition to the above, requiring that, for each possible observation-pair, (a^m, a^n), $Q^i_{t+1,mn} = R^i_{mn}$ or that the observation-pair occur with non-zero probability, and upon occurring, cause $|Q^i_{t+1,mn} - R^i_{mn}| \leqslant \rho |Q^i_{t,mn} - R^i_{mn}|$ for some $0 \leqslant \rho < 1$. That this criterion is sufficient for convergence follows from considering any observation-pair (a^m, a^n) where $|Q^i_{t,mn} - R^i_{mn}| = c_t > 0$ for some value c_t. Let the stopping-time $\tau_1 = t$ denote the event that t is the first time such that $c_t < c_{t-1}$. By hypothesis, τ_1 is finite with probability one. Then let τ_2 be the stopping-time associated with the second such decrease — conditioning on $t \geqslant \tau_1$ (a set which is nonempty as τ_1 is finite), τ_2 is likewise finite, etc. Thus, for any integer N, there exists, with probability one, a time t_N such that (a^m, a^n) has been drawn N times; in each case, the mismatch $|Q^i_{t,mn} - R^i_{mn}|$ decreases by at least a factor of $\rho < 1$; as this is true for each pair of indices and for all N, Brouwer's fixed-point theorem [18] implies that the Q-estimate converges element-wise and almost-surely to the true reward.

References

[1] L. Hirth, The market value of variable renewables, Energy Econ. 38 (2013) 218–236, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0140988313000285.

[2] D. Jay, K.S. Swarup, A comprehensive survey on reactive power ancillary service markets, Renew. Sustain. Energy Rev. 144 (2021) 110967, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032121002598.

[3] S.A. Gabriel, A.J. Conejo, J.D. Fuller, B.F. Hobbs, C. Ruiz, Complementarity Modeling in Energy Markets, Springer Science & Business Media, 2012, Google-Books-ID: Lu1L5wUea8IC.

[4] Y. Du, F. Li, H. Zandi, Y. Xue, Approximating Nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning, J. Mod. Power Syst. Clean Energy 9 (3) (2021) 534–544, [Online]. Available: https://ieeexplore.ieee.org/document/9406572.

[5] K. Poplavskaya, J. Lago, L. de Vries, Effect of market design on strategic bidding behavior: Model-based analysis of European electricity balancing markets, Appl. Energy 270 (2020) 115130, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261920306425.

[6] C. Graf, V. Zobernig, J. Schmidt, C. Klöckl, Computational performance of deep reinforcement learning to find Nash equilibria, Comput. Econ. (2023) [Online]. Available: https://doi.org/10.1007/s10614-022-10351-6.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning, 2013, [Online]. Available: http://arxiv.org/abs/1312.5602. arXiv:1312.5602 [cs].

[8] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, 2020, [Online]. Available: http://arxiv.org/abs/1706.02275. arXiv:1706.02275 [cs].

[9] J. Perolat, B. de Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J.T. Connor, N. Burch, T. Anthony, S. McAleer, R. Elie, S.H. Cen, Z. Wang, A. Gruslys, A. Malysheva, M. Khan, S. Ozair, F. Timbers, T. Pohlen, T. Eccles, M. Rowland, M. Lanctot, J.-B. Lespiau, B. Piot, S. Omidshafiei, E. Lockhart, L. Sifre, N. Beauguerlange, R. Munos, D. Silver, S. Singh, D. Hassabis, K. Tuyls, Mastering the game of stratego with model-free multiagent reinforcement learning, 2022, [Online]. Available: http://arxiv.org/abs/2206.15378. arXiv:2206.15378 [cs].

[10] F.A. Oliehoek, C. Amato, A Concise Introduction to Decentralized POMDPs, in: SpringerBriefs in Intelligent Systems, Springer International Publishing, Cham, 2016, [Online]. Available: http://link.springer.com/10.1007/978-3-319-28929-8.

[11] O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J.P. Agapiou, M. Jaderberg, A.S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T.L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354, [Online]. Available: https://www.nature.com/articles/s41586-019-1724-z. Number: 7782 Publisher: Nature Publishing Group.

[12] L. Wei, A.I. Sarwat, W. Saad, S. Biswas, Stochastic games for power grid protection against coordinated cyber-physical attacks, IEEE Trans. Smart Grid 9 (2) (2018) 684–694, Conference Name: IEEE Transactions on Smart Grid.

[13] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, 2019, [Online]. Available: http://arxiv.org/abs/1509.02971. arXiv:1509.02971 [cs, stat].

[14] R.S. Sutton, A.G. Barto, Reinforcement Learning: an Introduction, second ed., in: Adaptive Computation and Machine Learning Series, The MIT Press, Cambridge, Massachusetts, 2018.

[15] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, G. Strbac, Deep reinforcement learning for strategic bidding in electricity markets, IEEE Trans. Smart Grid 11 (2) (2020) 1343–1355, [Online]. Available: https://ieeexplore.ieee.org/document/8805177/.

[16] Y. Shoham, Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations.

[17] C. Shapiro, Chapter 6 Theories of oligopoly behavior, in: Handbook of Industrial Organization, Vol. 1, Elsevier, 1989, pp. 329–414, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1573448X89010095.

[18] J.R. Munkres, Topology, second ed., Prentice Hall, Upper Saddle River, NJ, 2000.