

Intelligent Bidding Strategies in Local Energy Markets using the Advantage Actor-Critic Algorithm

Sania Khaskheli and Amjad Anvari-Moghaddam

Department of Energy Technology, Aalborg University, Aalborg 9220, Denmark

Abstract- The growth of renewable energy and local trading has led to Local Energy Markets (LEM), enabling producers to trade surplus energy. Traditional bidding strategies struggle with market fluctuations, causing inefficiencies. This paper presents an agent-based framework using the Advantage Actor-Critic (A2C) algorithm to develop intelligent bidding strategies in a simulated market. The intelligent agent mimics real-world dynamics, considering fluctuating capacity and variable marginal costs. It determines optimal bid prices and quantities, improving decision-making under dynamic conditions. A2C integrates policy-based and value-based learning: the actor network optimizes trading actions, while the critic network evaluates and enhances expected returns. Simulation results confirm that our proposed intelligent agent (I1) effectively adapts bids to dynamic conditions and compared to ordinary agents (P1, P2 and P3) improves the Average Revenue (32.35 to 44.64%), Average Profit (37.97 to 45.46%), Profit Margin (0.6 to 6.71%), and Bid Success Rate (57.65 to 94.21).

Index Terms: Advantage Actor-Critic, Energy Markets, Energy Trading, Intelligent Bidding Strategies, Reinforcement Learning

I. INTRODUCTION

Energy trading involves power exchange between generators and suppliers, who distribute it to consumers. Optimized trading strategies enhance operational efficiency and incentivize participant engagement [1]. However, as market participants increase, maintaining stability and operational efficiency becomes challenging [2]. Local Energy Markets (LEMs) enable prosumers and communities to trade electricity locally, optimizing resource management and ensuring transactions occur near production points [3]. Existing literature lacks insights into LEM bidding mechanisms and optimal strategies. LEMs operate as time-series markets, requiring continuous bid submissions, making the process inefficient and time-intensive, necessitating an intelligent bidding mechanism [4]. Formulating optimal bidding strategies is challenging due to incomplete knowledge of others' strategies, price and quantity uncertainties, and market dynamics. Traditional rule-based methods often fail to adapt to these complexities. In contrast, Reinforcement Learning (RL) allows agents to refine strategies through trial-and-error interactions, converging on profitable decisions. However, conventional value-based RL struggles in continuous action spaces, requiring alternative approaches for effective bidding in LEMs [5]. To address this challenge, an Actor-Critic (AC)-based bidding strategy is proposed for energy traders. The actor learns optimal bidding decisions, while the critic evaluates long-term rewards. As a model-free RL method, it continuously interacts with the market, refining actions and value estimation. This work introduces Deep Reinforcement

Learning (DRL)-based intelligent bidding strategies using the Advantage Actor-Critic (A2C) algorithm in a simulated LEM. The designed market environment replicates real-world dynamics, allowing agents to observe fluctuating capacity factors and variable marginal costs.

A. Related Works

The energy trading bidding strategies can be traditional model-based optimization techniques [6], [7] and intelligent learning-based methods [8], [9]. The former uses mathematical optimization to determine optimal bidding strategies, considering market dynamics, operational constraints, and participant objectives. According to [10], intelligent bidding strategies include zero-intelligence and intelligent agent approaches. Zero-intelligence agents generate bids/offers randomly within predefined limits, while intelligent agents use optimization, algorithms, game theory, and learning to refine trading decisions based on experience or data. [11] proposed a linear bidding/offering strategy for LEMs allows agents to adjust prices linearly within a time slot—lowering bid prices or raising offer prices. This enables prosumers to submit multiple bids/offers, increasing market matching opportunities. [12] developed metrics to assess bidding strategy performance in LEMs, modeling risk-averse agent behaviors. By integrating expected profit and risk criteria, they formulated optimal multi-step energy quantity-price bidding strategies based on agent risk preferences. In [13], an optimal bidding /offering strategy is proposed for prosumers for improving cost savings and community welfare by optimizing energy trading for individual and collective benefits. [14] enhanced deep Q-learning for local energy trading by modifying the Deep Q-Network, optimizing prosumers' decision-making for adaptive trading strategies. [15] proposed a Q-learning-based bidding strategy for prosumers in a two-sided pay-as-bid LEM, enabling optimal bidding through market interactions to maximize individual benefits. [16] proposed an actor-critic bidding approach for LEMs, selecting a trained RL agent from a pool. It integrates SAC-based bidding with fixed acceptance, training multiple agents against diverse opponents to handle varied negotiation behaviors. [17] proposed a LEM framework where agents share cost functions with a market operator, enabling cooperative, decentralized network optimization for efficient resource allocation and reduced centralization. [18] uses a noncooperative Markov game framework with discrete multi-agent Q-learning, enabling generators to learn optimal bidding strategies, handle uncertainties, and achieve near-optimal outcomes in an information-limited market.

B. Contributions and Organization of work

The growing interest in RL for optimal bidding in energy trading comes from its ability to learn strategies through trial

and error without a precise environment model. RL suits LEMs due to their complex, non-deterministic nature. However, high-dimensional action spaces and 24-hour-ahead bidding increase learning challenges, demanding extensive training and large datasets for convergence. This work adopts A2C, a state-of-the-art DRL algorithm using an off-policy actor-critic framework based on maximum entropy. This approach enhances exploration by optimizing policies for both profit rewards and entropy maximization. The key contributions are as follows.

- I. Development of a DRL-based A2C framework for optimal day-ahead bidding in LEMs.
- II. Proposed A2C Bidding Algorithm and Pseudocode for A2C Algorithm
- III. Modeling of intelligent and ordinary agents (producers) employing random trading strategies.

The article is structured as follows: Section II defines the MDP problem, including state, action dynamics, reward function, and the proposed A2C framework. Section III details the A2C-based bidding strategy. Section IV analyzes simulation results for intelligent and random-bidding agents. Section V concludes the work and word on potential bidding strategy to improve market competitiveness.

II. SYSTEM MODEL

A. Problem Formulation

In RL, the environment's design defines state space, transition dynamics, and the reward function, shaping learned policies. Since these elements are usually unknown, the agent learns solely from interaction feedback, as shown in Figure 1. This work models the single agent bidding problem as an MDP to define optimal policies, value functions, and expected returns. RL then learns an optimal policy to maximize expected rewards for each state.

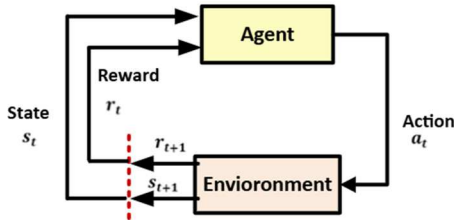


Fig 1: agent-environment interaction.

A finite MDP can be represented as $(S, A, \mathcal{R}, \mathcal{T}, \mu)$. Each element is defined as follows: S , is a finite set of states, A is a finite set of actions, and $\mathcal{R}: S \times A \times \mathbb{R}$ is the reward function. The state transition probability function, represented by $\mathcal{T}: S \times A \times \mathbb{R}$, where,

$$\mathcal{T}(s'|s, a) = P_r(s^{t+1} = s' | s^t = s, A = a), \text{ where}$$

$$s, s' \in S, a \in A, \forall s \in S, a \in A: \sum_{s' \in S} \mathcal{T}(s'|s, a) = 1 \quad (1)$$

Initial state distribution is represented by $\mu: S \rightarrow [0,1]$, where, $\sum_{s \in S} \mu(s) = 1$.

An MDP starts in an initial state $s_0 \in S$ sampled under the distribution function μ . At each time t the agent observes the current state $s^t \in S$ and chooses an action a^t based on some probability distribution $\pi(a^t | s^t)$ known as ‘policy’. The MDP transitions to the new state s^{t+1} given the current state

s^t and the action of agent a^t . This transition is ruled by $\mathcal{T}(s^{t+1} | s^t, a^t)$. Moreover, the agent receives a reward of r^{t+1} that is the result of the reward function $\mathcal{R}(s^t, a^t, s^{t+1})$. This cycle of observing the state, choosing an action, receiving the next state and reward continues until the MDP reaches the terminal state $s \in \bar{S}$ or the time t reaches the time limit T if the MDP is episodic. $\bar{S} \subset S$ is the set of terminal states where the probability of transitioning from the any $s \in S$ to all the other states in $s \in S$ is zero. The result of this interaction loop is a trajectory of state, action, and reward, which can be represented as:

$$s^0, a^0, r^1, s^1, a^1, r^2, \dots, s^{T-1}, a^{T-1}, r^T, r^T.$$

B. A2C Model Framework

Optimal bidding in LEM is challenging due to its high-dimensional, dynamic nature. To address this, we learn a capacity factor instead of directly optimizing bids, enabling dynamic bid price adjustments per state. In trading, agents lack explicit knowledge of others' responses, making behavior uncertain. Temporal difference (TD) methods address this by learning without prior environment dynamics and handling continuous action spaces. Using observed transitions and rewards, TD methods iteratively refine policies for precise decision-making. Considering these factors, we adopt the A2C algorithm within the maximum entropy RL framework. As shown in Figure 2, AC algorithms train two components simultaneously: (i) Actor – a parameterized policy selecting actions per state, and (ii) Critic – a value function estimating action quality based on expected returns.

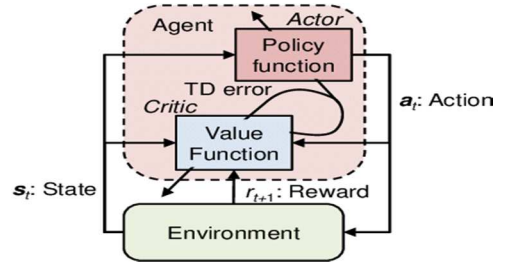


Figure 2: The A2C model interacting with environment.

The A2C algorithm develops an intelligent map trading states to actions, ensuring mutual trading needs effectively compared to ordinary bidders. While the policy does not explicitly guarantee optimality, the maximum entropy framework enhances exploration and robustness, enabling efficient navigation of complex state-action spaces. Over time, the algorithm converges toward an optimal policy, making it ideal for dynamic and uncertain trading environments. Here, the advantage function is calculated as,

$$adv^\pi(s, \alpha) = Q^\pi(s, \alpha) - V^\pi(s) \quad (2)$$

where $adv^\pi(s, \alpha)$ shows how effectively action α in state s is chosen and then follow the policy π , compared to the state-value of state s where the policy already chooses the action for the state s . The advantage formula can be reformulated to depend solely on the state-value function, eliminating the need for training separate networks to compute advantage. Initially, the action-value $Q^\pi(s, \alpha)$ function can be written as state-value function $V^\pi(s)$.

$$Q^\pi(s, \alpha) = \mathbb{E}_\pi[G^t | s^t = s, A^t = \alpha]$$

$$\begin{aligned}
&= \mathbb{E}_\pi[G^t | s^t = s, A^t = \alpha] \\
&= \mathbb{E}_\pi[r^{t+1} + \gamma G^{t+1} | s^t = s, A^t = \alpha] \\
&= r^{t+1} + \gamma \mathbb{E}_\pi[G^{t+1} | s^t = s, A^t = \alpha] \\
Q^\pi(s, \alpha) &= r^{t+1} + \gamma V^\pi(s^{t+1} = s, A^t = \alpha)
\end{aligned} \quad (3)$$

The new formulation of advantage $adv^\pi(s, \alpha)$ will be:

$$adv^\pi(s^t, \alpha^t) = r^{t+1} + \gamma V^\pi(s^{t+1}) - V^\pi(s^t) \quad (4)$$

III. PROPOSED A2C BIDDING ALGORITHM

Figure 3 illustrates the A2C agent's learning structure, where iterative self-exploration optimizes bids in a finite continuous state and action space. The ordinary agent generates bid B_t^o , processed by the state creator function to generate state s_t . The agent utilizes an experience replay buffer and demonstration buffer to store past experiences as (S, A, R, D, S') tuples, which represents a transition experience in RL. Each element is defined as follows:

- S (State): Current state of the environment at time step t ,
- A (Action): Action taken by the agent at state s_t ,
- R (Reward): Reward received after taking action A_t ,
- D (Done): Boolean flag indicating whether the episode has ended (1 if the episode is done, 0 otherwise).
- S' (Next State): State reached after executing action A_t .

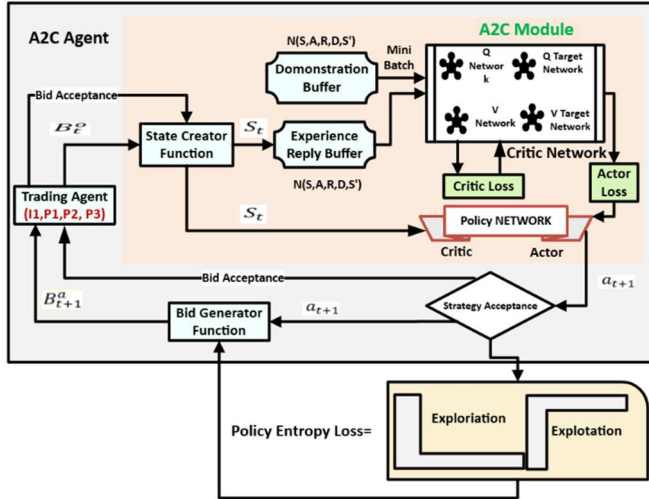


Fig 3. RL based A2C agent framework for energy market bidding.

A mini batch of experiences is sampled for training. The actor selects actions a_t , while the critic evaluates them using Q and V networks. Critic loss and actor loss are computed using respective target networks to update parameters. The bid generator formulates a_{t+1} , evaluated by an acceptance strategy before submission. This cycle iteratively optimizes bidding strategies. Entropy measures action selection uncertainty, encouraging exploration (new strategies) by balancing new and exploitation (learned strategies). Higher entropy prevents convergence to suboptimal policies, enhancing robustness. The trade-off between expected return and entropy stabilizes learning, reducing the risk of premature exploitation or overfitting. This makes A2C effective in dynamic, uncertain environments. As defined in [19], Equation 6 shows the policy objective with entropy maximization where \mathcal{H} denotes to entropy function and α determines relative importance of entropy. We use the same behavior to identify the loss function (actor and critic).

$$\pi^* = \arg \max_{\pi} \sum_t E_{S_t, a_t \sim \pi} [r(S_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | S_t))] \quad (5)$$

Equation 6 shows the entropy loss computed only on the demonstration examples for training actor, introduced in [20]. This loss aims to improve the learned policy without going beyond the demonstration data. Considering the gradient applied to the parameters θ_π described as $\lambda_1 \nabla \theta_\pi R - \lambda_2 \nabla \theta_\pi \text{Loss}_{function}$, the aim is to maximize expected return R and minimize loss function.

$$\text{Loss}_{function} = \sum_{i=1}^{N_D} \|\pi(S_i | \theta_\pi) - a_i\|^2 \quad (6)$$

The actor loss follows the REINFORCE formula, while the critic loss is the mean squared error between the critic's output and the target value from environment interactions. Pseudocode 1 presents the A2C algorithm and loss function.

PSEUDOCODE 1: Advantage Actor-Critic

- 1: Initialize parameters ϕ of a actor network π at random
- 2: Initialize parameters θ of a critic network V at random
- 3: For each episode:
- 4: For $t = 0, 1, 2, 3, \dots, T - 1$ do:
- 5: Observe current state s^t
- 6: Sample action $a^t \sim \pi(\cdot | s^t, \phi)$
- 7: Apply action a^t , observe next state s^{t+1} and reward r^{t+1}
- 8: If s^{k+1} is terminal
- 9: Advantage $adv^\pi(s^t, a^t) = r^{t+1} - V^\pi(s^t | \theta)$
- 10: Critic target $y^t \leftarrow r^{t+1}$
- 11: Else:
- 12: Advantage $adv^\pi(s^t, a^t) = r^{t+1} + \gamma V(s^{t+1} | \theta) - V(s^t | \theta)$
- 13: Critic target $y^t \leftarrow r^{t+1} + \gamma V(s^{t+1} | \theta)$
- 14: Calculate loss $\mathcal{L}(\phi) = -adv^\pi(s^t, a^t) \log_\pi(a^t | s^t, \phi)$
- 15: Calculate loss $\mathcal{L}(\theta) = (y^t - V(s^t | \theta))^2$
- 16: Update parameters ϕ to minimize the loss $\mathcal{L}(\phi)$
- 17: Update parameters θ to minimize the loss $\mathcal{L}(\theta)$

IV. SIMULATION RESULTS AND ANALYSIS

This work considers a 24-hour day-ahead market with hourly price quotes based on buy and offer bids using a double auction approach. Pay-as-clear mechanism is used for market price clearing mechanism. The simulation analyzes datasets on capacity factors and marginal costs of power producers. It tests RL-trained intelligent agent (II) and three random bidders (P1, P2, P3) based on varying price and quantity in LEM. Training includes up to 70,000 episodes with simulation setup executed at different episodes in Python language using PyTorch library. To balance data collection and simulation time, a 50-day trade-off is chosen. Fewer days risk bias results from random fluctuations, while more can be time-consuming. The metrics are computed for different training episodes were:

Training rewards (€/time step): For RL agent, maximizing profit.

Actor & critic loss: The actor loss function is computed based on policy gradients thus no direct physical unit.

Policy Entropy Loss: It measures uncertainty in the agent’s policy, balancing exploration and exploitation. As a probability function, it is dimensionless.

Average Revenue (€/day): Total daily earnings, indicating market engagement and bid effectiveness.

Average Profit (€/day): Net earnings after costs, reflecting efficiency in converting revenue into gains.

Profit Margin (%): Percentage of revenue retained as profit, showing cost efficiency and pricing strategy.

Bid Success Rate (%): The percentage of successful bids placed, indicating market competitiveness and the effectiveness of an agent’s bidding strategy.

The RL training rewards, actor and critic loss function and policy entropy for 10,000 to 70,000 training episodes are shown in Figure 4.

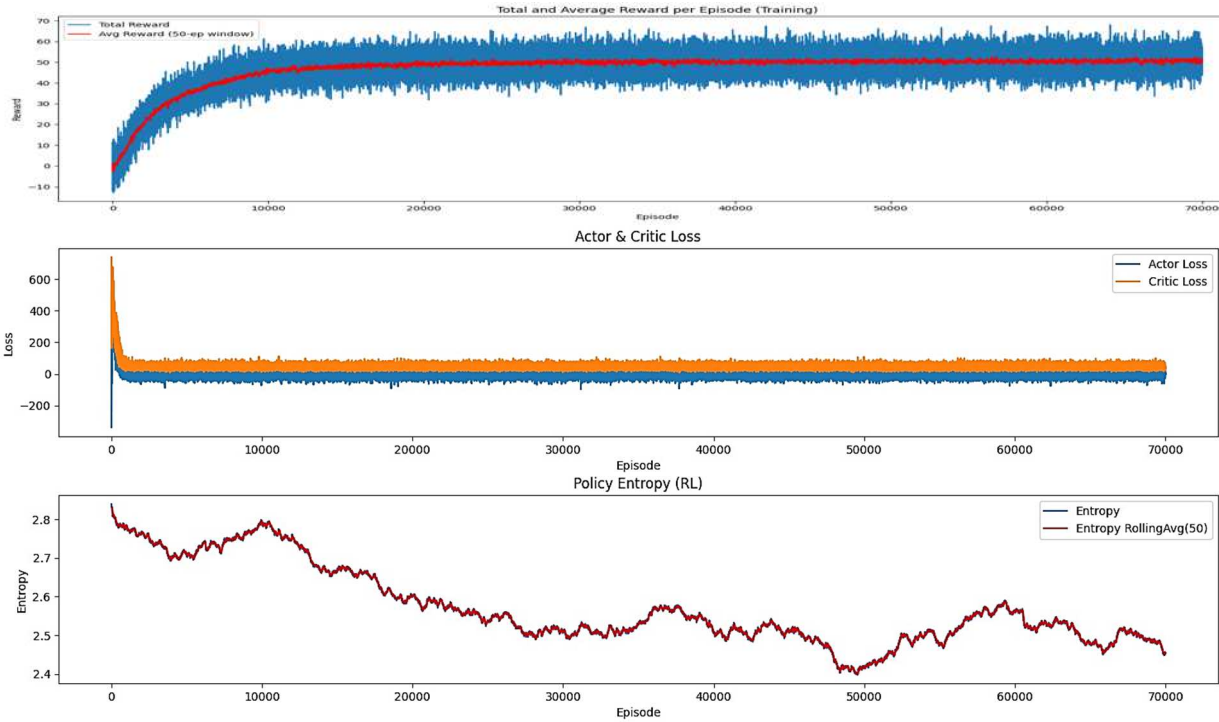


Fig 4: RL training rewards, Actor and Critic Loss Function and policy entropy for 70,000 episodes.

Figure 4 presents total reward per training episode, actor-critic loss, and policy entropy across episodes. The results indicate clear learning progression and convergence for the energy trading RL model. At 10,000 episodes, total reward is low (~30 €/MWh) as the model explores, with critic loss dropping from 600 to below 200, actor loss stabilizing around -100 to -200, and policy entropy decreasing from 2.8 to ~2.7, signaling a transition to structured decisions. By 20,000 episodes, reward increases sharply to (~40€/MWh), critic loss remains between 100–200, actor loss around -150, and policy entropy drops to ~2.65, indicating a shift from exploration to exploitation. At 30,000 episodes, reward stabilizes (45-50 €/MWh), actor-critic losses fluctuate minimally, confirming policy convergence, while policy entropy declines to 2.5, showing reduced randomness. At 40,000 episodes, the reward stabilizes at (50-55 €/MWh), with steady actor-critic loss and entropy reducing to ~2.45, indicating a largely deterministic policy. By 50,000 episodes, reward reaches (55-58 €/MWh), with entropy at ~2.4, balancing exploration and exploitation. At 60,000 episodes, reward rises to (58-62 €/MWh), actor-critic loss remains stable, and entropy increases to ~2.58, ensuring minimal randomness while retaining adaptability. At 70,000 episodes, reward peaks at (62 €/MWh), confirming optimized, profit-maximizing energy trading decisions. Actor-critic loss

remains steady, confirming learning saturation, while policy entropy stabilizes at ~2.5, marking final policy refinement.

Overall, the training reward stabilizes after 20,000 episodes, confirming successful learning. Actor-critic losses stabilize, preventing overfitting, while entropy declines from 2.8 to 2.4, indicating refined decision-making with retained adaptability. The results confirm full convergence, ensuring a robust and optimized policy. Based on results presented in Figure 4, we further trained trading agents (I1, P1, P2, and P3) and analyzed their learning and convergence behavior for Average Revenue, Average Profit, Profit Margin, and Bid Success Rate across training episodes (10,000 to 70,000). The results of this evaluation are presented in Figures 5 to 8.

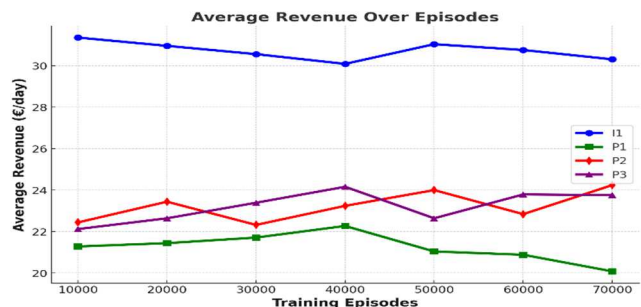


Figure 5: Average revenue for different training episodes

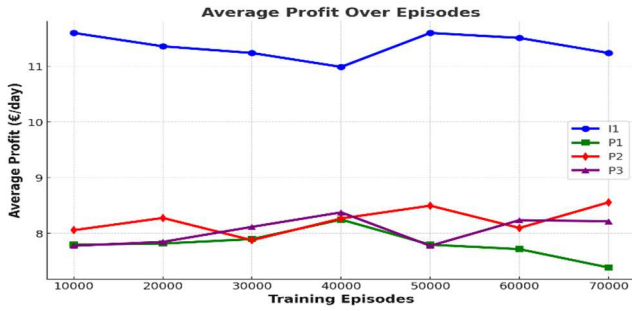


Figure 6: Average profit for different training episodes

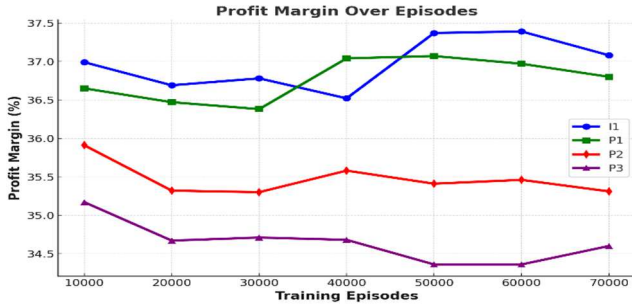


Figure 7: Average profit margin for different training episodes



Figure 8: Bid success rate for different training episodes

Figures 5 to 8 shows the simulation results of I1, P1, P2, and P3 learning performance analyzed across Average Revenue, Average Profit, Profit Margin, and Bid Success Rate over increasing training episodes. I1 consistently outperformed all other agents showing the best performance across all parameters, with stable revenue (~30-31€/day), high profitability (~11€/day), and the highest bid success rate (~98-99% across training episodes. This indicates I1's superior bidding strategy, ensuring consistent market participation and maximizing earnings. P1 is the weakest performer, with low revenue (~20-22€/day), poor profitability (~7.3-8.2€/day), and the worst bid success rate (~48-53%), making it inefficient for energy trading. P1 has shown a steady but slower learning curve, meaning it takes a longer time to optimize its bidding approach. P2 has strong profitability (~8.5€/day), an improvement in profit margin and an aggressive bidding strategy but lacks stability, with bid success rates fluctuating between 56-66%, indicating inconsistency, implying an aggressive but unstable bidding strategy. P3 maintains a moderate position, with revenue (~22-24€/day), profit (~7.8-8.3€/day), and bid success (~55-64%), but does not excel in any specific category, indicating a balanced but less competitive strategy. Table 1 summarizes a parameteric comparison of trading agents, how our proposed agent has outperformed the other agents. Table 2 summarizes the trading agents (I1, P1, P2, P3) based on their

performance across key metrics and identifying the best and worst agents for each parameter

Table 1: Parameteric comparison of trading agents.

Proposed Agent	Ordinary Agents	Average Revenue (€/day)	Average Profit (€/day)	Profit Margin (%)	Bid Success Rate (%)
I1	P1	44.64	45.46	0.60	94.21
	P2	32.34	37.97	4.24	57.65
	P3	32.35	41.10	6.71	59.27

Table 2: Performance evaluation tradeoff between trading agents

Metric	Best Agent	Worst Agent	Performance Evaluation
Average Revenue	I1	P1	P1 should optimize its bidding strategy to improve revenue generation.
Average Profit	I1	P1	P1 should focus on better cost management and competitive bidding.
Profit Margin	I1	P3	P3 should refine its bidding strategy for higher profitability.
Bid Success Rate	I1	P1	P1 should adopt a more aggressive bidding to win more trades.
Stability & Convergence	I1	P2	P2 should balance risk-taking with stability to improve performance.

The results indicate that I1's bidding strategy is the most effective, balancing high revenue, profitability, and bid success rate. P2 shows strong profitability but lacks stability in bid success, requiring fine-tuning for consistency. P1 converges slowly, necessitating strategy optimization for faster adaptation, potentially through adaptive learning techniques. P3, as an average performer, could either adopt a more aggressive strategy like P2 or enhance bid stability like I1 for better overall performance.

V. CONCLUSION

This study validates an A2C agent-based framework for intelligent bidding in LEM. The A2C algorithm balances exploration and exploitation which ensures robust learning, convergence and refining bidding strategies over extended training. Episodes. By combining policy- and value-based learning, the A2C optimally adjusts bid prices and quantities to dynamic market conditions while considering fluctuating capacity, marginal costs, and strategic interactions. The RL-based energy trading model successfully converges, stabilizing after 20,000 episodes with increasing rewards and steady actor-critic losses, ensuring policy robustness. Policy entropy decreases from 2.8 to ~2.4, indicating a shift to exploitation while retaining adaptability. With 70,000 episodes, the model optimizes profit-maximizing decisions with no loss divergence, confirming learning saturation. These results validate its stability and efficiency in market participation. Simulation results show that the RL-trained I1 has converted to a most effective and stable strategy, achieving superior revenue generation, high profitability, and consistent bid success rates (~98-99%). RL techniques like A2C with dynamic pricing adjustments could further refine bidding decisions, leading and adaptive bidding to maximize efficiency, profitability and competitive energy trading strategy.

REFERENCES

- [1] S. Khaskheli and A. Anvari-Moghaddam, "Energy Trading in Local Energy Markets: A Comprehensive Review of Models, Solution Strategies, and Machine Learning Approaches," *Applied Sciences*, vol. 14, no. 24, p. 11510, 2024.
- [2] S. Khaskheli, I. A. Halepoto, and A. Khalid, "Residential Community Micro Grid Load Scheduling and Management System Using Cooperative Game Theory," *3c Tecnología: glosas de innovación aplicadas a la pyme*, vol. 8, no. 1, pp. 534–551, 2019.
- [3] R. Faia, F. Lezama, J. Soares, T. Pinto, and Z. Vale, "Local electricity markets: A review on benefits, barriers, current trends and future perspectives," *Renewable and Sustainable Energy Reviews*, vol. 190, p. 114006, 2024.
- [4] J. Wang *et al.*, "Two-stage distributionally robust offering and pricing strategy for a price-maker virtual power plant," *Appl Energy*, vol. 363, p. 123005, 2024.
- [5] L. Cheng *et al.*, "Integrating Evolutionary Game-Theoretical Methods and Deep Reinforcement Learning for Adaptive Strategy Optimization in User-Side Electricity Markets: A Comprehensive Review.," *Mathematics (2227-7390)*, vol. 12, no. 20, 2024.
- [6] N. Uthayansuthi and P. Vateekul, "Optimization of Peer-to-Peer Energy Trading with a Model-Based Deep Reinforcement Learning in a Non-Sharing Information Scenario," *IEEE Access*, 2024.
- [7] O. M. Sedeh and B. Ostadi, "Optimization of bidding strategy in the day-ahead market by consideration of seasonality trend of the market spot price," *Energy Policy*, vol. 145, p. 111740, 2020.
- [8] J. Wu, J. Wang, and X. Kong, "Strategic bidding in a competitive electricity market: An intelligent method using Multi-Agent Transfer Learning based on reinforcement learning," *Energy*, vol. 256, p. 124657, 2022.
- [9] A. Taghizadeh, M. Montazeri, and H. Kebriaei, "Deep reinforcement learning-aided bidding strategies for transactive energy market," *IEEE Syst J*, vol. 16, no. 3, pp. 4445–4453, 2022.
- [10] S. N. Islam, "A Review of Peer-to-Peer Energy Trading Markets: Enabling Models and Technologies," *Energies (Basel)*, vol. 17, no. 7, p. 1702, 2024.
- [11] G. C. Okwuibe, "Evaluation of hierarchical, multi-agent, community-based, local energy markets based on key performance indicators. *Energies* 15 (10), 3575 (2022)."
- [12] J. D. Schölzel, S. Henn, M. Tings, R. Streblov, and D. Müller, "Comparative analysis of bidding strategies for auction-based local energy markets," *Energy*, vol. 291, p. 130211, 2024.
- [13] G. C. Okwuibe, J. Bhalodia, A. S. Gazafroudi, T. Brenner, P. Tzscheutschler, and T. Hamacher, "Intelligent bidding strategies for prosumers in local energy markets based on reinforcement learning," *IEEE Access*, vol. 10, pp. 113275–113293, 2022.
- [14] L. Cheng *et al.*, "Integrating Evolutionary Game-Theoretical Methods and Deep Reinforcement Learning for Adaptive Strategy Optimization in User-Side Electricity Markets: A Comprehensive Review.," *Mathematics (2227-7390)*, vol. 12, no. 20, 2024.
- [15] G. C. Okwuibe, M. Wadhwa, T. Brenner, P. Tzscheutschler, and T. Hamacher, "Intelligent bidding strategies in local electricity markets: A simulation-based analysis," in *2020 IEEE Electric Power and Energy Conference (EPEC)*, IEEE, 2020, pp. 1–7.
- [16] A. Sengupta, Y. Mohammad, and S. Nakadai, "An autonomous negotiating agent framework with reinforcement learning based strategies and adaptive strategy switching mechanism," *arXiv preprint arXiv:2102.03588*, 2021.
- [17] M. I. Azim *et al.*, "Coalition game theoretic P2P trading in a distribution network integrity-ensured local energy market," *Sustainable Energy, Grids and Networks*, vol. 36, p. 101186, 2023.
- [18] C. Wu, W. Gu, Z. Yi, C. Lin, and H. Long, "Non-cooperative differential game and feedback Nash equilibrium analysis for real-time electricity markets," *International Journal of Electrical Power & Energy Systems*, vol. 144, p. 108561, 2023.
- [19] H. Xu, Q. Wu, J. Wen, and Z. Yang, "Joint bidding and pricing for electricity retailers based on multi-task deep reinforcement learning," *International journal of electrical power & energy systems*, vol. 138, p. 107897, 2022.
- [20] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2018, pp. 6292–6299.