P2P Energy Trading through Prospect Theory, Differential Evolution, and Reinforcement Learning

ASHUTOSH TIMILSINA and SIMONE SILVESTRI, Department of Computer Science, University of Kentucky, USA

Peer-to-peer (P2P) energy trading is a decentralized energy market where local energy *prosumers* act as peers, trading energy among each other. Existing works in this area largely overlook the importance of user behavioral modeling, assume users' sustained active participation, and full compliance in the decision-making process. To overcome these unrealistic assumptions, and their deleterious consequences, in this paper we propose an *automated* P2P energy trading framework that specifically considers the users' perception by exploiting *prospect theory*. We formalize an optimization problem that maximizes the buyers' perceived utility while matching energy production and demand. We prove that the problem is NP-hard and we propose a Differential Evolution-based Algorithm for Trading Energy (*DEbATE*) heuristic. Additionally, we propose two automated pricing solutions to improve the sellers' profit based on reinforcement learning. The first solution, named Pricing mechanism with Q-learning and Risk-sensitivity (*PQR*), is based on Q-learning. Additionally, the given scalability issues of *PQR*, we propose a Deep Q-Network-based algorithm called *ProDQN* that exploits deep learning and a novel loss function rooted in prospect theory. Results based on real traces of energy consumption and production, as well as realistic prospect theory functions, show that our approaches achieve 26% higher perceived value for buyers and generate 7% more reward for sellers, compared to recent state-of-the-art approaches.

Additional Key Words and Phrases: Peer-to-peer energy trading, differential evolution, dynamic pricing, non-linear optimization, prospect theory, Q-learning, Deep Q-Network.

1 INTRODUCTION

1.1 Background and Motivation

The detrimental effects of the energy sector on the environment have raised the urgency to move towards an environment-friendly, efficient, and sustainable energy landscape [9, 32]. As a result, Renewable Energy Technologies (RETs), and more specifically Distributed Energy Resources (DER) such as rooftop solar and wind turbine, have been driving a transformation of modern power systems [32, 42]. DER have seen widespread proliferation among consumers in recent years [2], fueled by the Internet of Things (IoT)-enabled Smart Grid (SG) [11, 12, 15, 27], that embraces the use of cutting edge technologies to make the grid smarter and an active ecosystem for energy exchanges among all the stakeholders [52]. The advent of SG technologies, such as the Advanced Metering Infrastructures (AMI) [8] and home energy management systems [28], have resulted in additional flexibility for consumers to generate and consume energy. This, in turn, has allowed traditionally passive consumers to become actively involved in energy trading by sharing the excess energy generated at their premise to either the grid or other buyers [1, 35, 44, 45]. These active consumers with energy production capabilities have been referred to as *prosumers* [35], as a portmanteau of "producers" and "consumers".

Authors' address: Ashutosh Timilsina, ashutosh.timilsina@uky.edu; Simone Silvestri, simone.silvestri@uky.edu, Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA.

2688-3007/2023/6-ART $15.00
https://doi.org/10.1145/3603148

However, the potential of prosumers in energy markets has been only minimally exploited with the adoption of incentive schemes like *Feed-in-Tariff* (FiT) mechanisms [47, 50]. Through these schemes, prosumers can sell the generated excess energy to the grid, and buy from the grid in case of deficiency [24, 47]. FiT are generally adopted in the form of *net-metering*, i.e., the grid only offers minimal, or even none, monetary incentives for the acquired excess energy [3, 35, 47, 50]. Additionally, several grids throughout the world that employ net-metering have placed capping in terms of the amount of energy that prosumers can sell to the grid. As a result, FiT strategies are deemed economically inconvenient for the prosumers and have been discontinued altogether in several locations [47, 50]. Nevertheless, the availability of IoT-enabled SG technologies have the potential to enable more efficient and convenient energy trading mechanisms, as discussed below.

1.2 P2P Energy Trading

Peer-to-peer (P2P) energy trading is a recently proposed decentralized alternative to the traditional energy trading modality. It provides flexibility for end-users to be involved in energy trading and has been gaining popularity in recent years [35, 51]. Specifically, P2P energy trading provides a prosumer-centric platform that allows prosumers to trade energy with each other at a negotiated price. The trading may or may not involve the grid [40, 50]. Typically, the operation range of trading price would be higher than FiT price offered by grid and lower than the electricity tariff charged by utilities. Further, utility companies offering rates that change with the time-of-day make the P2P paradigm even more convenient, particularly when the price of electricity charged by the grid is at its peak [40]. Monetary incentives resulting from P2P energy trading, for both selling prosumers (producers) and buying prosumers (consumers), are therefore far better compared to existing mechanisms. As a result, prosumers are more incentivized to keep engaging with the trading for the long-term [40, 50].

P2P energy trading also aims at minimizing the dependency of prosumers from grid for energy [47], resulting in an increased reliability of the overall system. Additionally, a higher amount of local energy generation and consumption resulting from P2P trading leads to the minimization of the overall system energy loss, as well as an effective way to achieve demand side management [55]. Benefits extend also to the grid operator, by providing savings in investments that would have been otherwise required to develop/maintain transmission infrastructure in a centralized power distribution architecture [35, 50]. Therefore, P2P energy trading offers a prosumer-centric approach that has potential to benefit all stakeholders involved, as highlighted extensively in recent studies [4, 33, 36, 44, 47, 49, 54].

Existing literature on P2P energy trading has predominantly focused on the technical and physical aspects of P2P energy trading, such as voltage regulation and loss minimization [33, 36], while few others have examined pricing mechanisms for P2P trading among prosumers [4, 47, 49]. Most of these works assume a constant active participation of users in the trading scheme. There are very few existing literature that take into account the user behavioral aspects in designing energy trading systems so far, like the works in [16, 44, 46–48, 53]. However, these works too demand users' active participation which, over the long run, may be overwhelming and could result in irrational decisions, and even the abandonment of trading system [17, 21]. As studies [18] have shown, users' subjective perceptions of gain and loss can significantly influence their decision-making processes, and as a result, their willingness to engage in P2P energy trading. To address these challenges, as established in [35, 44, 47], new approaches that incorporate user perceptions and irrational decision-making into the design of P2P energy trading systems are necessary for ensuring user's long-term engagement and sustainability.

1.3 Paper Contributions

To overcome the limitations of existing works, in this paper we specifically take into account and model the prosumers' decision-making behavior and their perceptions of loss and gain. We capture the perceived utility through an optimization framework aimed at matching energy between sellers and buyers in an effective and

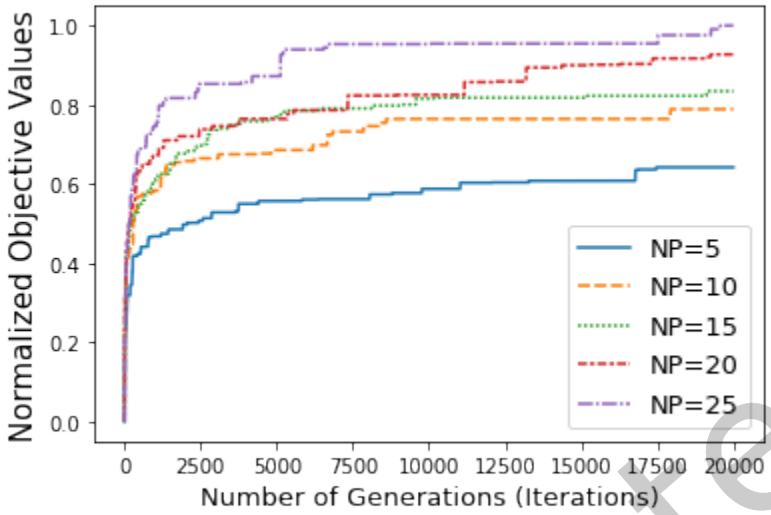

Fig. 1. P2P Energy Trading System Overview.

efficient manner. This is a complex task, as it involves numerous individual prosumers with varying levels of energy demand and production, as well as their differing perceptions of loss and gain. To this purpose, we utilize the widely accepted notion in behavioral decision-making called *Prospect Theory* (PT) [25]. PT captures the non-rational decision making of humans in the face of uncertainty, and it provides a mathematical tool to quantify the subjective perception loss and gain. Specifically, we propose a PT-based optimization framework for prosumer-centric P2P energy trading that incorporates perceived utility into the trading and automates the price updating for sellers using reinforcement learning. This is more clearly depicted in Fig. 1. The proposed framework aims at matching energy demand and production between buyers and sellers (step 1 in Fig. 1). The objective is to maximize the perceived utility of individual buyers, by taking into account their intrinsic perception and heterogeneity. We formalize this as a non-linear and non-convex optimization problem, and prove that it is NP-hard. Given the non-linear and non-convex nature of the problem on top of being NP-hard, we further propose a Differential Evolution-based Algorithm for Trading Energy (*DEbATE*) to find a solution to the problem in polynomial time (*energy allocation*, step 2).

In order to require minimal participation of prosumers, we employ a Reinforcement Learning (RL) framework, called Pricing mechanism with Q-learning and Risk-sensitivity (*PQR*), which is executed in tandem with *DEbATE*, to automate the pricing mechanism for sellers (*pricing mechanism*, step 3). Sellers are not aware of their competitiveness in the market. Therefore, *PQR* adjusts the price dynamically based on the market demand as well as seller's competitiveness and perceived utility. *PQR* learns the optimal selling price for each seller using a PT-based risk-sensitive RL approach [39]. However, *PQR* inherits the typical scalability and stability limitations of a standard tabular approach for the Q-learning function. To avoid such limitations, we further improve *PQR* by proposing a Deep Reinforcement Learning based alternative heuristic, called *ProDQN*, that uses a PT-based loss function to include the sellers' perceived utility. The output of the RL algorithms are then published to the prosumers for the next matching of demand and production. Finally, the output of the matching is then implemented by the P2P energy system to execute the physical energy transactions (step 4). The proposed techniques address the limitations of previous works by modeling individual prosumers' behavior, incorporating perceived utility, and automating the price updating process for sellers through a unique and novel optimization

problem in conjunction with a reinforcement learning framework. Employing a Differential Evolution-based heuristic, paired with reinforcement learning based pricing mechanisms, allows to efficiently find a solution to the non-linear and non-convex problem, which is especially critical for large systems with many prosumers. Additionally, by incorporating PT-based approaches, the individual subjective perception loss and gain can be quantified, which is an essential aspect of prosumer-centric P2P energy trading. Finally, through the use of reinforcement learning, the system can learn from the prosumers' behavior and adapt to changes in market conditions, leading to a more efficient and effective P2P energy trading system. In summary, this paper makes novel contributions in the domain of evolutionary computing by defining a novel demand/production matching problem, and a differential evolution algorithm to solve it, that take into account the users' perceptions through prospect theory. Additionally, the paper makes contributions in the domain of reinforcement learning by defining Q-learning based algorithms that learn competitive prices while taking into account the sellers' perceived utility.

We validate the proposed approaches through extensive simulations using real datasets for energy production and consumption, paired with recent survey data for PT perception modeling. Experimental results show that *DEbATE* performed 26% better in terms of buyer's perceived utility than a state-of-the-art approach. Additionally, *PQR* is able to generate 7% higher sellers' reward. The deep reinforcement learning solution *ProDQN* is able to further improve the sellers' reward by 8%, at the expense of a small reduction in buyers' utility.

The rest of the paper is organized as follows. The system model and problem statement are described in Section 2. Then, the proposed automated P2P energy trading framework is explained in detail in Sections 3. Furthermore, the experimental results are elaborated in Section 4. Section 5 reviews the modeling of optimization problems adopted for P2P energy trading in the literature. Finally, Section 6 concludes the paper.

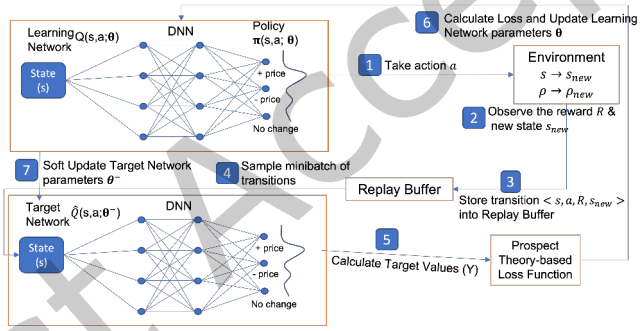

Fig. 2. Proposed Framework of P2P Energy Trading.

2 SYSTEM MODEL AND PROBLEM FORMULATION

The components of the proposed framework for P2P energy trading system are shown in Fig. 2. The systems consists of three distinct components, namely (i) Prosumers, (ii) Energy allocation, and (iii) Pricing mechanism. We describe the modeling of the three components in detail in following subsections.

2.1 Modeling Prosumers

The prosumers are users of the power system, equipped with an energy management system that allows to buy and sell energy in the energy market through an existing distribution network. Some prosumers are equipped with energy generation capabilities. The grid serves as backup for prosumers if the generated and traded energy is insufficient to satisfy the demand. Let P be the set of all prosumers participating in the P2P energy market. We refer to $B_t \subset P$ as the set of *Buyers*, i.e., the set of prosumers that have a higher self-consumption than generation

at a timeslot t. B_t also include prosumers without energy generation capabilities. Similarly, $S_t \subset P$ is the set of *Sellers*, i.e., prosumers that have excess generation at a timeslot t. For simplicity of notation, we drop the subscript t in the following.

2.2 Modeling Energy Allocation

In this subsection, we present the energy allocation mechanism which determines how to match the buyers' demand with the sellers' production, while maximizing their individual perception. We model the perceived loss and gain of prosumers using the *prospect theory* (PT) value function to capture user perception of gains and losses as shown in Fig. 2. Specifically, consider the excess energy generation of seller $i \in S$ be e_i and demand of buyer $j \in B$ be d_j. Then, let $x_{ij} \in [0, 1]$ be a variable representing the fraction of d_j that buyer j buys from seller i. Also, let ρ_{gs}, ρ_{gb} be the energy selling and purchasing prices from the grid, respectively, ρ_i be the selling price of seller i, and ρ_j is the *reference price* of buyer j. Prices are expressed as cost per kWh.

We adopt a modified PT value function to model realistic user perception in an energy market [25]. The function quantifies the human perceived utility towards gain and loss based on the degree of deviation from a reference point. In our problem, it captures the difference between the total actual buying cost for the buyer j i.e., y_j and their desired reference cost $\rho_j d_j$ for buying d_j amount of energy at their reference price. The utility function is then formulated as

$$v(y_j) = \begin{cases} k_{+,j}(\rho_j d_j - y_j)^{\zeta_{+,j}}, & y_j < \rho_j d_j \\ -k_{-,j}(y_j - \rho_j d_j)^{\zeta_{-,j}}, & y_j \geq \rho_j d_j \end{cases} \tag{1}$$

where $k_{+,.}, k_{-,.}, \zeta_{+,.}, \zeta_{-,.}$ are the parameters that control the degree of loss-aversion and risk-sensitivity. Similarly to [6, 20, 38], we assume that these parameters can be obtained by surveys completed by the prosumers when the energy trading system is installed in their home, and potentially updated later on with sporadic feedback to the energy management system. Recent studies have shown that these parameters are highly heterogeneous and vary from person to person based on factors like gender and age group [6, 38]. In the equation above, y_j is the total actual cost of buying energy for buyer j that incorporates the total cost of buying energy from P2P setting as well as the grid – in case the demands are not met locally. Therefore the term y_j is given by

$$y_j = \sum_{i \in S} \rho_i x_{ij} d_j + \rho_{gs}(1 - \sum_{i \in S} x_{ij}) d_j$$

Note that, similar to the PT value function in [25], the utility function in Eq. (1) is concave in the gain domain (i.e., $y_j < \rho_j d_j$) while convex in loss domain (i.e., $y_j \geq \rho_j d_j$).

The problem of matching demand and production of heterogeneous prosumers is formalized as follows.

$$\max \quad f(y) : \sum_{j \in B} v(y_j) \tag{2}$$

$$\text{s.t.} \quad \sum_{j \in B} (1 + \ell_{ij}) x_{ij} d_j \leq e_i, \quad \forall i \tag{2a}$$

$$\sum_{i \in S} x_{ij} \leq 1, \quad \forall j \tag{2b}$$

$$x_{ij} = 0, \text{ if } \ell_{ij} \geq \ell_{max}, \quad \forall i, j \tag{2c}$$

$$\rho_{gb} \leq \rho_i, \rho_j \leq \rho_{gs}, \quad \forall i \tag{2d}$$

$$\mu_j z_{ij} \leq x_{ij} d_j \leq d_j z_{ij}, \quad \forall i, j \tag{2e}$$

$$z_{ij} \geq x_{ij}, \quad \forall i, j \tag{2f}$$

$$x_{ij} \in [0, 1], \; z_{ij} \in \{0, 1\}, \quad \forall i, j \tag{2g}$$

The problem maximizes the sum of perceived utility for buyers in Eq. (2). There is an *energy loss* during the physical energy transfer through wires [55], which depends on the wire-length between i and j and it is directly proportional to the amount of energy exchanged. We model such loss as a fraction $\ell_{ij} \in [0, 1]$ of the energy exchanged. As a result, the constraint in Eq. (2a) prevents the problem from exceeding the amount of energy being sold by each sellers while incorporating the losses in electric lines. The constraint in Eq. (2b) ensures that the energy demand for each buyer is not exceeded, while constraint (2c) limits the loss between sellers and buyers to be within the loss threshold l_{max}. On the other hand, constraint (2d) limits the upper and lower bound for energy price to the selling and buying price of the grid. Constraint (2e) sets the minimum amount μ_j of an energy transaction for buyer j, using a binary decision variable z_{ij}, that is equal to 1 if $x_{ij} > 0$, and to 0 otherwise (constraint (2f)). The value of μ_j is generally a system parameter to prevent impractical solutions containing infinitesimal amounts [44]. Finally, the range of operation of the decision variables are defined in (2g).

Theorem 2.1. *The optimization problem in Eq.* (2) *is NP-hard.*

Proof. Proof of NP-hardness is presented in the Appendix A □

It is to be noted that, in addition to the NP-Hardness, the problem in Eq. (2) is also non-linear and non-convex. There is not any general procedure to solve such optimization problems dealing with continuous solution sets [7]. Hence, in order to solve this combinatorial problem of matching demand and supply of energy, we propose a heuristic based on Differential Evolution [41] which finds a feasible solution through iterative recombination and improvement of the candidate solutions along with constraint handling. Adopting this heuristic approach is particularly beneficial in large systems, where the complexity of the problem would make it impossible to find the optimal solution in reasonable time. In the next section, we motivate the need for a dynamic pricing mechanism mechanism, before presenting the Differential Evolution heuristic in Section 3.

2.3 Modeling Pricing Mechanism

In the proposed P2P energy trading model, the selling price is considered to be a fixed amount for a given trading period, and it is used as the trading price for a transaction. However, the reference price ρ_i of seller i is a personal value which may under- or over-estimate the competitiveness of market. In order to improve the sellers' competitiveness, we implement a dynamic pricing model for sellers as exhibited in the Fig. 2. Note that, to expect sellers to manually adjust the price based on the performance of the energy trading system (e.g., their revenue) is impractical. Demanding such active participation could easily be overwhelming, and severely affect their performance and level of engagement with the system. To avoid such active participation, we model the

adjustment of the price as a Markov Decision Process, and exploit reinforcement learning to update the selling price at each trading period.

To maximize the sellers' perceived objectives through prospect theory, we resort to a risk-sensitive reinforcement learning approaches that forms the basis of the automated pricing mechanism within our P2P energy trading framework. In the following Section 3, we present two algorithms based on reinforcement learning that incorporate the seller's perceptions on loss and gains to update the prices while automating the process in order to ensure sustained prosumer participation. First, in Subsection 3.2, we employ risk-sensitive Q-learning algorithm [39] and then given its efficiency limitation due to the tabular representation of the Q-function, we also present a Deep Q-network (DQN) [31] based algorithm in Subsection 3.3 that proposes a novel loss function founded on prospect theory value function.

Algorithm 1: DEbATE

Input : set of buyers B, sellers S, fitness function $f(.)$, max iterations G_{max}, population size NP, crossover probability CR, differential weight F

Output : best identified feasible solution $\mathbf{x}^*$

Update set of buyers B and sellers S, $count = 0$;

Generate initial population $\mathcal{X} = \{\mathbf{x_k} |\ k = 1, \ldots, NP\}$;

while $count < G_{max}$ **do**

  for *each* $\mathbf{x_k} \in \mathcal{X}$ **do**

    Choose 3 different vectors $\{\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c}\} \in \mathcal{X}$ at random and $R \sim U(1, |S| \times |B|)$;

    Create mutated solution $\bar{\mathbf{x}}_\mathbf{k} = \mathbf{x_k}$;

    /* **Mutation and Crossover** */

    for *each* $i \in |S|$, $j \in |B|$ **do**

      Select $u \sim U(0, 1)$;

      if $u < CR || (i \times j) == R$ **then**

        $\bar{x}_{ij}^{(k)} = x_{ij}^{(a)} + F \times (x_{ij}^{(b)} - x_{ij}^{(c)})$;

        $\bar{x}_{ij}^{(k)} = \min(1, \max(0, \bar{x}_{ij}^{(k)}))$;

    end

    /* **Check Constraints** */

    $\forall i, j$, **if** $\ell_{ij} \geq \ell_{max}$ **then** $\bar{x}_{ij} = 0$;

    $\forall i$, **if** $\sum_j (1 + \ell_{ij}) \bar{x}_{ij} d_j > e_i$ **then** $\bar{x}_{ij} = \frac{\bar{x}_{ij} r_i}{\sum_j (1+l_{ij}) \bar{x}_{ij} w_j}$;

    $\forall j$, **if** $\sum_i \bar{x}_{ij} > 1$ **then** $\bar{x}_{ij} = \frac{\bar{x}_{ij}}{\sum_i \bar{x}_{ij}}$;

    /* **Compare fitness** */

    if $f(\bar{\mathbf{x}}_\mathbf{k}) > f(\mathbf{x_k})$ **then** $\mathcal{X} = (\mathcal{X} \setminus \{\mathbf{x_k}\}) \cup \{\bar{\mathbf{x}}_\mathbf{k}\}$;

  end

  count = count++;

end

/* **Find the best solution and execute trading** */

Let $\mathbf{x}^* = \arg\max_{\mathbf{x_k} \in \mathcal{X}} f(\mathbf{x_k})$;

Execute transactions for each prosumer to $\mathbf{x}^*$;

3 SOLUTION APPROACHES AND HEURISTICS

In this section, we describe the *Differential Evolution-based Algorithm for Trading Energy (DEbATE)* heuristic (Alg. 1), designed for matching demand and production according to the problem presented in Section 2, the *Pricing mechanism with Q-learning and Risk-sensitivity (PQR)*, and the *Prospect theory-based Deep Q-Network (ProDQN)*, designed to dynamically adjust the sellers' prices.

3.1 DEbATE

DEbATE is executed at each trading period (e.g., 12 hours) to solve the non-linear optimization problem presented in Section 2. It uses Differential Evolution to determine an optimal amount of energy to be traded between prosumers that maximizes the perceived utility of buyers. *DEbATE* initially updates the list of buyers (B) and sellers (S) based on the expected production and consumption for the current trading period. These can be predicted accurately with recent approaches [10, 29]. The Differential Evolution-based optimization begins on line 2 where an *initial population* $\mathcal{X}$ is generated with population size of NP. An element $\mathbf{x_k} \in \mathcal{X}$, with $k = 1, 2, \ldots, NP$ is a *candidate solution* vector of variables x_{ij} representing the amount of energy to be traded between the i^{th} seller and j^{th} buyer. These variables correspond to the decision variables of our optimization problem.

The *while*−loop (line $3-19$) is the differential evolution loop that aims at finding a solution to the non-linear optimization problem with Eq. (2) as the fitness function. The loop is executed for G_{max} iterations. At each iteration, for each candidate solution $\mathbf{x}_k \in \mathcal{X}$, the algorithm creates a *mutated solution* $\bar{\mathbf{x}}_\mathbf{k}$. Initially, $\bar{\mathbf{x}}_\mathbf{k} = \mathbf{x_k}$. The mutated solution is subsequently updated through mutation and crossover with 3 random candidates $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c \in \mathcal{X}$ (line 5). A value $R \in [1, |S| \times |B|]$ is selected at random. R will be used in the following for−loop to ensure a minimum mutation. The for loop in line 7 iterates over the components (dimensions in evolutionary terms) of $\bar{\mathbf{x}}_\mathbf{k}$. During each iteration, a value $u \in [0, 1]$ is sampled at random as mutation probability (line 8). Subsequently, a mutation occurs for the component ij of $\bar{\mathbf{x}}_\mathbf{k}$ with crossover probability CR (line 9). The mutation occurs irrespective of the probability if $(i \times j) = R$ (to ensure at least one minimum mutation). A mutation is executed by combining the corresponding component of $\mathbf{x_a}$, $\mathbf{x_b}$, and $\mathbf{x_c}$ with the differential weight parameter $F \in [0, 2]$ as in line 10. The mutated component $\bar{\mathbf{x}}_\mathbf{ij}^{(\mathbf{k})}$ is clipped to ensure that it falls within $[0, 1]$ as minimum and maximum threshold to satisfy constraint Eq. (2g) in line 11 of the algorithm.

After the mutated solution is finalized, it is checked, and adjusted if needed, to meet the constraints in Eqs. (2a)-(2c) of the optimization problem. Specifically, line 13 ensures that no exchange occurs (i.e., $\bar{\mathbf{x}}_\mathbf{ij}^{(\mathbf{k})} = 0$) between users having a loss higher than l_{max}. Lines $14-15$ ensure that the production of a seller and the demand of each buyer are not exceeded, respectively. Finally, in line 16, the fitness function $f(.)$ of the mutated solution $\bar{\mathbf{x}}_\mathbf{k}$ is compared against the original candidate solution $\mathbf{x_k}$. If $f(\bar{\mathbf{x}}_\mathbf{k}) > f(\mathbf{x_k})$, then $\bar{\mathbf{x}}_\mathbf{k}$ replaces $\mathbf{x_k}$ in the set of candidate solutions $\mathcal{X}$. At the end of the while loop, *DEbATE* selects the best solution $\mathbf{x}^*$ in $\mathcal{X}$ (line 20) and executes the transactions accordingly (line 21). In the following Lemma A.2, we show that *DEbATE* has polynomial time complexity, and hence it is computationally efficient. The theorem focuses on the asymptotic complexity, a typical mathematical formulation to characterize the upper-bound of the running-time for sufficiently large inputs [14].

Lemma 3.1. *The time complexity of the DEbATE algorithm is* $O(G_{max} \times NP \times |S||B|)$.

Proof. Proof of the time complexity of DEbATE algorithm can be found in the Appendix A. □

3.2 PQR

After determining the solution to the energy allocation problem in *DEbATE*, the selling price for sellers is then updated through the *P*ricing mechanism with *Q*-learning and *R*isk-sensitivity (*PQR*) algorithm as presented in Alg. 2. In order to learn the optimal selling price dynamically over time, we model the sellers as independent learning agents. Note that, to preserve the privacy and avoid the conflict between prosumers, these agents do not

Algorithm 2: PQR

```
/* Pricing with Risk-sensitive Q-learning                                    */
```
1 Collect transaction information for each prosumer from *DEbATE* (Alg. 1) for current timestep t;
2 **for** *each* $i \in S$ **do**
3 Select an action, $a \in \{+\delta, -\delta, 0\}$ based on the ϵ-greedy strategy ;
4 $s = \rho_i; s_{new} = s + a; e_i = s_{new} \sum_{j \in B} x_{ij} d_j$;
5 $\rho_i = s_{new}$;
6 Update $Q(s, a)$ as in Eq. (3) and (4);
7 Send information on updated price ρ_i to seller i;
8 **end**

have access to information about other sellers or buyers. The state space (s) of the Markov Decision Process, in the Q-learning formulation, consists of the prices between the grid buying/selling ρ_{gb} and ρ_{gs}, discretized by a step size, δ, i.e., $\rho_i \in \{\rho_{gb},\ \rho_{gb} + \delta,\ \rho_{gb} + 2\delta,\ \ldots,\ \rho_{gb} + \big(\frac{\rho_{gs} - \rho_{gb}}{\delta} - 1\big)\delta,\ \rho_{gs}\}$.

The action space consists of a price increasing action, price decreasing action, and no change action, i.e., $a \in \{+\delta, -\delta, 0\}$ where δ is the amount by which price is increased or decreased. The agent goes to a new state after taking action a which is referred as s_{new}. Seller i reward function is the total revenue generated at the current trading period i.e., $e_i = (\rho_i + a) \sum_{j \in B} x_{ij} d_j$. For updating Q-values, we modify the approach proposed in [39] by considering the following Q-learning update rule that includes the PT-based perceived utility of sellers.

$$Q(s, a) \leftarrow Q(s, a) + \alpha v(y_i) \tag{3}$$

$$v(y_i) = \begin{cases} k_{+,i}(y_i)^{\zeta_{+,i}}, & y_i > 0 \\ -k_{-,i}(-y_i)^{\zeta_{-,i}}, & y_i \leq 0 \end{cases} \tag{4}$$

where, $y_i = e_i + \gamma \max_a Q(s_{new}, a) - Q(s, a)$ is the Temporal Difference (TD) error of i^{th} seller for current iteration, and $v(y_i)$ is transformation of TD error to capture each seller's personalized perceived utility on loss. α refers to the learning rate for updating Q-values in Eq. (3).

PQR, as summarized in Alg. 2, initially collects the current trading information from *DEbATE* in line 1. The subsequent $for-$loop (lines 2 – 8) updates the selling price for each seller. At each iteration, a seller $i \in S$ is considered. For that seller, the action (whether to increase/decrease the price, or no change) is selected based on a *ϵ-greedy* exploration-exploitation strategy [43] (line 3). Specifically, ϵ refers to the probability of exploration and it is initially set to 1. It is then decreased over time using an *$\epsilon-$decay factor*, that is ϵ = decay factor $\times\ \epsilon$. This way, exploitation gains more importance as the system learns the optimal policy. The algorithm returns an action a, that is used to update the current state s into the new state s_{new}, and to update the reward e_i (line 4). Additionally, the Q-value is updated accordingly (line 6) . The updated selling price is then sent to the respective seller i for the next trading period in line 7.

As discussed in the experimental section, *PQR* is able to correctly learn the optimal policy (or sellers' prices, in our case). However, it needs to be noted that in a P2P energy trading model, like in most realistic scenarios, the state spaces can be very large and multidimensional. In fact, since the Q-learning must maintain Q-values for each state-action pair, even with just three actions, a finely discretized state space could lead to a huge number of state-action pairs that needs to be stored and updated continually. This is worsened by increasing the number of agents (sellers). As a result, *PQR* may suffer from severe scalability issues, due to its tabular approach of

determining Q-values, as the system grows. Therefore, in the next subsection, we utilize a widely employed deep neural network-based *function approximator* that can be used to predict the Q-values using a learned function given state-action pairs.

3.3 ProDQN

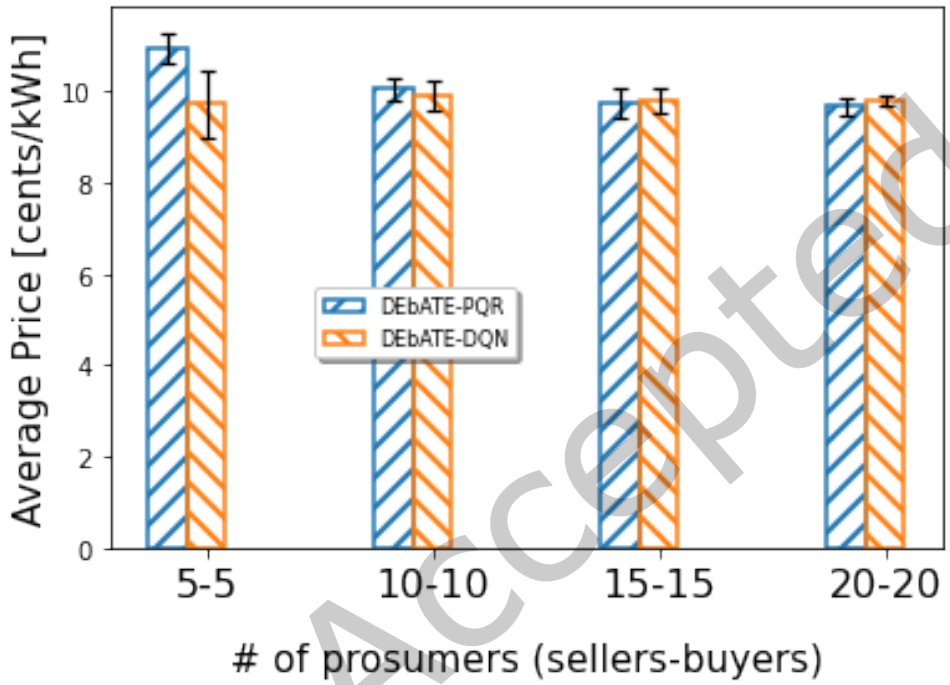

Fig. 3. Overview of the ProDQN approach

In this subsection, we adopt a reinforcement learning approach based on Deep Q-Network (*DQN*) [31], for learning the seller's optimal price, in order to overcome the scalability limitations of *PQR*. DQN is a reinforcement learning paradigm that exploits a deep neural network, called Q-network, as a non-linear function approximator. Its parameters (or *weights*) are denoted by $\boldsymbol{\theta}$, thus the Q-value function $Q(s, a)$ becomes $Q(s, a; \boldsymbol{\theta})$. Note that, approximating the function through a neural network allows to not only represent the Q-values in a compressed form compared to the tabular Q-learning algorithm, but also to generalize over similar states.

In this paper, we extend *DQN* to incorporate Prospect Theory elements, in order to devise a perceived utility-based pricing mechanism. We refer to our heuristic as *Pro*spect theory-based *DQN* (*ProDQN*). An overview of the heuristic is presented in Fig. 3. It is important to note that, using a single Q-network for reinforcement learning may result in instability. This is due to the need of training the neural network itself while using it as a Q-function approximator. This is known as the issue of *moving target*, where the target, i.e., the expected optimal price in our application, is varying after each training period. *DQN* solves this problem by utilizing *two different networks*. One network is used for learning, while the other one to determine the target respectively. The first network is the *learning network*, denoted by $Q(s, a; \boldsymbol{\theta})$, which is used to take the best action given the current state. Secondly, we have a *target network*, denoted by $\hat{Q}(s, a; \boldsymbol{\theta}^{-})$, which is used to determine how close the output

of the learning network is. The main difference between these network is that the learning network is updated after every training period while the target network is updated less frequently. Thanks to these less frequent updates, the target network is kept relatively stable, and thus the overall learning also becomes also stable.

An overview of *ProDQN* is shown in Fig. 3. *ProDQN* also employs two Q-networks – learning and target networks. Similar to *PQR*, each seller is represented by an individual *ProDQN* agent, and these agents do not share any information with each other to avoid conflicts and preserve privacy. Additionally the state spaces and action spaces are also the same as considered in *PQR*. As shown in Fig. 3, the learning network Q is used with parameters $\boldsymbol{\theta}$ to predict the current action a, given the current state $s = \rho_i$ as input. The action is either to increase ($+\delta$), decrease ($-\delta$), or no change (item 1). Following this, the action is executed and the consequence of action taken is observed (i.e., new state s_{new} and new price ρ_{new}, and the resulting reward (R) is observed (item 2). The transition tuple $< s, a, R, s_{new} >$ is stored in a Replay Buffer D. A minibatch of transitions $[z]$ of size z is randomly sampled from D (item 4) and passed to the target network $\hat{Q}$. The reason behind random sampling is to avoid bias due to high correlation in subsequent tuples. The target network returns a target value for each tuple (item 5), which is used to determine the error (or loss) in learning (item 6) and finally update the learning network parameters $\boldsymbol{\theta}$.

Specifically, for each sample $m = \{s^{(m)}, a^{(m)}, R^{(m)}, s^{(m)}_{new}\}$ in the minibatch $[z]$, the target value $Y^{(m)}$ is given by

$$Y^{(m)} = R^{(m)} + \gamma \max_{a'} \hat{Q}(s^{(m)}_{new}, a'; \boldsymbol{\theta}^-) \tag{5}$$

The computation of term $\max_{a'} \hat{Q}(s^{(m)}_{new}, a'; \boldsymbol{\theta}^-)$ is obtained from a single forward pass in the target network $\hat{Q}$ for a given state $s^{(m)}_{new}$. According to the original version of *DQN* [31], given the target values as in Eq. (5), the parameters $\boldsymbol{\theta}$ of the learning network $Q(s, a; \boldsymbol{\theta})$ are updated through stochastic gradient descent by minimizing a standard loss function, usually the square loss. Conversely, in our work we propose a novel loss function in Eq. (6) based on PT-value function similar to the one proposed in Eqs. (1) and (4). Specifically, given the target value $Y^{(m)}$ of tuple m, the PT-based loss function $\mathcal{L}^{(m)}$ is defined as:

$$\mathcal{L}^{(m)} = \begin{cases} k_{+,\cdot}(Y^{(m)} - Q(s, a; \boldsymbol{\theta}))^{\zeta_{+,\cdot}}, & Y^{(m)} > Q(s, a; \boldsymbol{\theta}) \\ -k_{-,\cdot}(Q(s, a; \boldsymbol{\theta}) - Y^{(m)})^{\zeta_{-,\cdot}}, & Y^{(m)} \leq Q(s, a; \boldsymbol{\theta}) \end{cases} \tag{6}$$

Recall that $k_{+,\cdot}, k_{-,\cdot}, \zeta_{+,\cdot}, \zeta_{-,\cdot}$ are the PT parameters that quantify the perceived utility. After calculating the loss for each sample, the mean loss is determined by averaging the loss for all samples in the minibatch i.e., $\mathcal{L} = \frac{1}{z} \sum_m \mathcal{L}^{(m)}$. The learning network's parameters are then updated by performing gradient descent step on network parameters $\boldsymbol{\theta}$, using the newly calculated loss $\mathcal{L}$ as follows:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha . \frac{1}{z} \sum_{m \in [z]} \left[(Y^{(m)} - Q(s, a; \boldsymbol{\theta})) \right] \nabla_{\boldsymbol{\theta}} Q(s, a; \boldsymbol{\theta}) \tag{7}$$

Finally, the parameters $\boldsymbol{\theta}^-$ of the target network are updated through *soft updates*. Specifically, an exponential moving average with parameter τ is used as follows:

$$\boldsymbol{\theta}^-_i \leftarrow \tau * \boldsymbol{\theta}_i + (1 - \tau) * \boldsymbol{\theta}^-_i \tag{8}$$

This process is then repeated for all the sellers to adjust their selling price in an automated manner similar to *PQR* algorithm (Alg. 2).

To the best of our knowledge, this is the first work using a PT value function-based loss calculation to update the Q-network parameters. This loss function is specially suited in our application scenario as it provides a way to capture the perceived utility of sellers based on the deviation from target values. It is to be noted that, with this PT-based loss function, the prediction of the Q-network $Q(s, a; \boldsymbol{\theta})$ tends to the target value (Y) (and therefore, to optimal Q-function i.e., $Q^*(s, a)$), transformed by the perceived utility of sellers as we update the parameters $\boldsymbol{\theta}$.

The system runs the algorithms *DEbATE* and *PQR*/*ProDQN* sequentially at every trading period. The input of *DEbATE* is updated based on the prices calculated by *PQR* or *ProDQN*, while *PQR*/*ProDQN* take as input the reward resulting from the energy transactions executed by *DEbATE*.

4 EXPERIMENTAL RESULTS

In this section, we discuss the experimental setup, the comparison approaches, and then provide a detailed discussion on performance of both of the proposed solutions versus the comparisons. In the following, we refer to *DEbATE* paired with *PQR* as $DEbATE - PQR$, and similarly to *DEbATE* paired with *ProDQN* as $DEbATE - DQN$.

4.1 Experimental Setup

The experimental setup consists of a system with 40 prosumers, split evenly as buyers and sellers. This is considered a representative number of prosumers in a microgrid or set of houses supplied by a single distribution transformer. We use a realistic dataset for buyer's energy consumption obtained from [23]. Similarly, we consider sellers equipped with $4kW$ rooftop solar located in Lexington, Kentucky, USA. The energy generated is estimated using NREL's PVWatts Calculator [34] given the solar irradiance in Lexington and size of solar panels. Losses are assigned uniformly at random (UAR) from set $\{1\%, 2\%, 3\%, 4\%\}$ and maximum loss threshold $L_{max} = 2.5\%$, while the minimum amount of energy to be exchanged (μ_j) is set to 50Wh for each buyer.

We assume that prosumers complete a survey before joining the system to estimate their individual prospect theory parameters, similar to [6, 20, 38]. For the purpose of experimentation, we use realistic prospect theory parameters from [6, 20, 38]. Specifically, the risk-averting parameter for gains $(\zeta_+) \in [0.60, 0.88]$, the risk-seeking parameter for losses $(\zeta_-) \in [0.52, 1.0]$, the loss-aversion parameters for gain and loss $(k_+), (k_-) \in [2.10, 2.61]$ for each individual prosumer. The grid or utility company generally sells energy at a higher price compared to the price it purchases energy. Therefore, we set the price at which the grid buys energy to $\rho_{gb} = \$0.06$ and the price at which it sells energy to $\rho_{gs} = \$0.12$, based on Kentucky's average net-metering rate and energy selling price. With the P2P energy trading paradigm, sellers and buyers can exploit this gap to sell/buy energy among each other at a more convenient price than the grid. Therefore, we set grid's energy selling price as upper bound for reference price for energy sellers and grid's buying price as lower bound for reference price for energy buyers, respectively. Specifically, the reference price for each seller is initially randomly sampled from range $[0.09, 0.12]$. It is then updated using either *PQR* or *ProDQN* at each iteration. The reference price for each buyer is selected in the range $[0.06, 0.10]$ and considered static for the duration of experiments, which is 365 days. The parameters for *PQR* algorithm are set as follows: learning rate $\alpha = 10^{-4}$, step size for discretizing state space $\delta = \$0.001$, and $\epsilon-$decay is set 0.965.

ProDQN uses two Q-networks, learning and target network. Each of these consists of an input and output layer connected by two hidden layers with 64 nodes each. The input layer consists of a single node for state and output layer consists of three nodes for three actions. Other hyperparamters are set as follows: learning rate $\alpha = 0.0075$, replay buffer size $|D| = 1000$, minibatch size $z = 4$, discount factor $\gamma = 0.8$, soft update rate for target network $\tau = 0.01$. These hyperparameters were chosen manually for best results. Hyperparameter optimization techniques could also be adopted such as grid search, Bayesian search, and population-based evolutionary search for further fine-tuning. We developed the P2P energy trading simulation environment and implemented the algorithms in Python using SciPy and PyTorch libraries. [1]

4.2 Comparison Approaches

In order to highlight the efficacy of our proposed approaches $DEbATE - PQR$ and $DEbATE - DQN$, we compare their performance against two recently proposed state-of-the-art approaches. The first approach, referred to as

[1]Scripts for the simulation can be found at this Github link: https://github.com/ashutoshtmlsna/P2P_energy_trading

Rule, is proposed in [4]. *Rule* allocates energy using a greedy heuristic that assigns cheapest sellers to buyers. Buyers are selected based on their registration order to the system. A mid-market pricing strategy is employed, i.e., the final price of a transaction is the mid value of seller's and buyer's asking price.

The second approach has been proposed in [55], to which we refer as *Zhu* from the name of the first author. This approach has been proposed to minimize the loss of the energy transactions. It employs a greedy algorithm to assign the energy among buyers and sellers. The algorithm considers buyers in decreasing order of demand. At each iteration, a buyer is selected and assigned to the sellers with the smallest loss for that buyers, until the demand of the buyer is satisfied. In this approach, the transaction price is given by the seller's asking price.

It is worth nothing that, both approaches do not consider the perceived utility of the buyers and they do not dynamically adjust the price of sellers. As discussed in the following, our approach matches demand and production by generating a market in which both the needs and perceptions of buyers and sellers are taken into account.

4.3 Results

We consider several experimental scenarios and performance metrics, as discussed in the following.

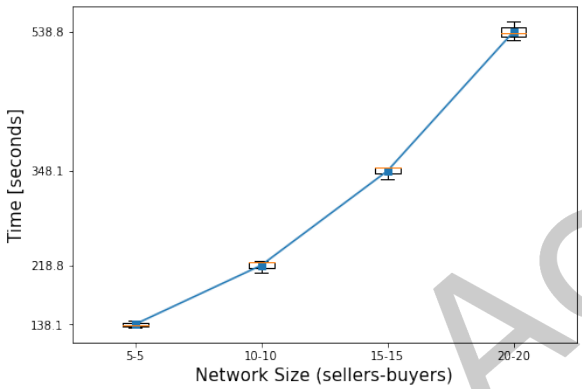

Fig. 4. Normalized objective value vs. number of iterations for varied population sizes.

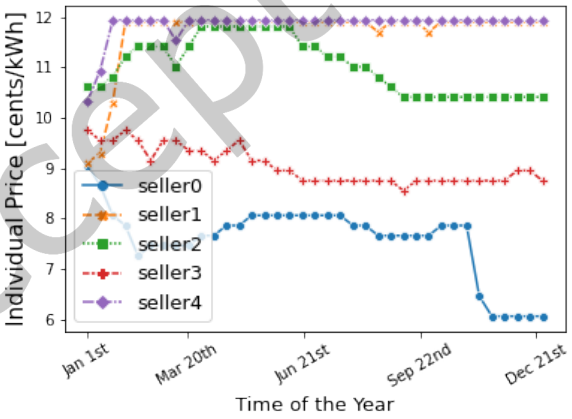

Fig. 5. Computation time vs. population sizes.

Experimental Scenario 1: We first run two experiments to study the evolutionary aspects and convergence of *DEbATE*. Specifically, we want to know the impact of the number of generations (G_{max}) and population size (NP) on the quality of the solution, i.e., on the value of the objective function. We first study the impact of the population size NP on the value of the objective function of the optimization problem in Eq. 2 and on the computational time. Specifically, we vary NP from 5 to 25. In this experiments, we consider a system with 5 sellers and 5 buyers. Fig. 4 shows the respective plot averaged over 10 runs. It can be seen that, as the population size is increased, *DEbATE* is able to find a better solution. Additionally, with all the considered population sizes, *DEbATE* is able to quickly converge towards a good solution with few iterations. We show in Fig. 5 the computational time versus the population size along with error bars. The figure clearly shows that the computation time grows linearly with the population size. This is in accordance with the complexity derived in Lemma A.2.

We now consider the impact of the number of generations (iterations) G_{max} on the quality of the solution and execution time. We consider different network sizes by linearly scaling the number of sellers and buyers from $|S| = |B| = 5$ to $|S| = |B| = 20$, and setting $NP = 20$. Fig. 6 shows the normalized objective value as a function of

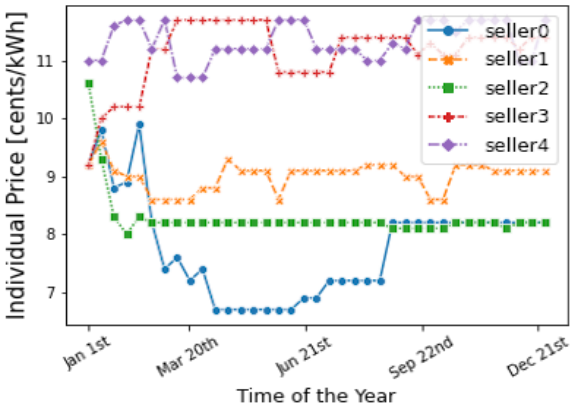

Fig. 6. Normalized objective value vs. number of iterations for varied network sizes.

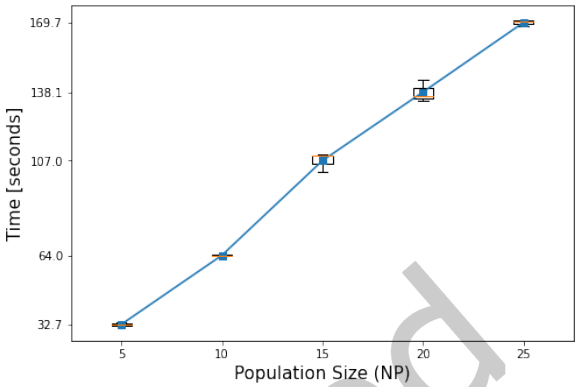

Fig. 7. Computation time vs. network sizes.

G_{max}. The results show that by setting $G_{max} = 10,000$ iterations is sufficient for the algorithm to converge in the considered settings. We plot in Fig. 7 the computation time of *DEbATE* by increasing the system size along with the error bars. According to Lemma A.2, the computation complexity is proportional to $|S| \times |B|$. As a result, by increasing both buyers and sellers linearly, we incur in a quadratic increase in the execution time.

Given the results of the aforementioned experiments, in the following, we select a trade-off between computation time and quality of the solution. For the remaining of the experiments, we therefore set the population size $NP = 20$, since it yields a solution with similar objective value while requiring 22% less execution time, and set $G_{max} = 10,000$. This helps to ensure that the algorithm will generate a quality solution.

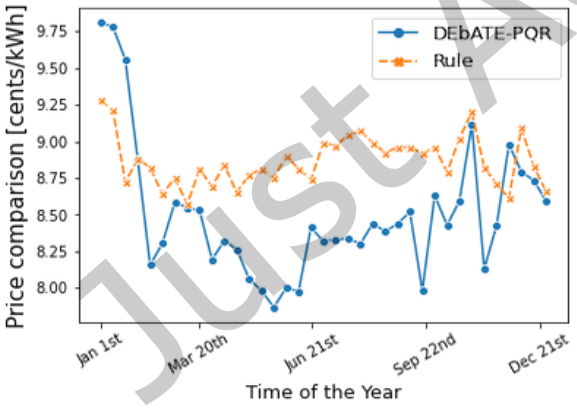

Fig. 8. Buyers' perceived values.

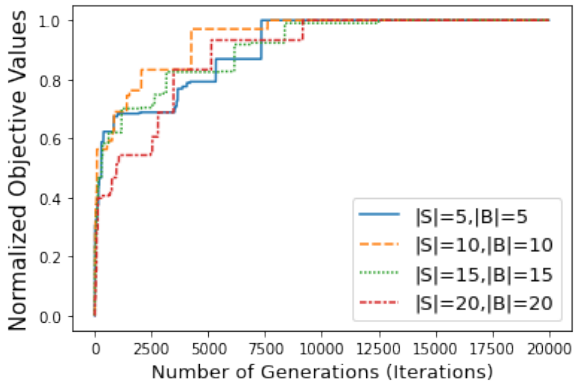

Fig. 9. Sellers' cumulative reward.

Experimental Scenario 2: In the second experimental scenario we study the performance of the considered approaches over time. Two performance metrics are considered, namely the buyers' objective value and the sellers' cumulative reward. These are represented in Figs. 8 and 9, respectively, with a moving average of 10 days. In these experiments, we consider 15 buyers and 15 sellers. Note that, the buyers' objective values are

Table 1. Statistical Analysis of experimental result in Figs. 8 and 9

Metric	Approach	DEbATE-PQR	DEbATE-DQN	Rule	Zhu
Obj. Val	Mean	**-83.182**	-97.818	-128.599	-155.312
	Std. Deviation	**34.738**	37.800	60.288	61.983
Reward	Mean	197.9	**202.3**	191.8	**215.8**
	Std. Deviation	**79.778**	82.963	79.916	90.198

negative because they are paying higher prices than their reference purchase price. Therefore, transactions are seen as loss from a prospect theory perspective. Clearly, $DEbATE-PQR$ and $DEbATE-DQN$ perform better than *Rule* in both metrics. The greedy nature of *Rule* penalizes the quality of the resulting matching, significantly reducing the buyers' perceived value while both our approaches optimize the energy assignments to maximize the buyers perceived utility. Additionally, our approaches are able to generate higher rewards than *Rule* by dynamically learning the prices for sellers. *Zhu* however performs the worst in terms of the buyers' objective values, but performs the best in terms of cumulative reward. This is because the energy assignment is driven by loss minimization, not taking into consideration the buyers' reference price. This, paired with the trading price set as the sellers' asking price, results in a market heavily biased towards sellers, achieving a very low perceived utility for buyers. We present a statistical analysis of experimental results in Figs. 8 and 9 with respect to both mean and standard deviation in table 1. As the table shows, the mean objective value is significantly higher for $DEbATE-PQR$ and $DEbATE-DQN$, with respect to the comparison approaches. This is also paired with a lower standard deviation, which implies more stable system performance. $DEbATE-DQN$ produces a slightly higher mean reward and a higher standard deviation. This is due to higher randomness engraved in deep learning frameworks. In line with our observation from Fig. 9, *Zhu* highly favors sellers with respect to buyers.

It is worth noting that, the benefits of $DEbATE-PQR$ and $DEbATE-DQN$ over *Rule* and *Zhu* are more prominent from April through October, when the energy demand and production are higher. Note that, the energy consumption is higher during summer months due to the higher use of air conditioning equipment. Similarly, the energy production is higher due to the increased solar radiation in these months. Comparing $DEbATE-PQR$ and $DEbATE-DQN$ we notice that *ProDQN* slightly penalizes buyers (lower utility) in favor of sellers (higher rewards). This slight imbalance is however compensated by the better scalability of *ProDQN*. In general, the sellers' reward decreases after mid-September for all four approaches, due to the reduced energy production during winter.

Experimental Scenario 3: We further study the performance over time by considering the evolution of individual sellers' prices. We consider a smaller system of 5 sellers and 5 buyers for ease of representation of the results. Fig. 10 shows the individual prices set by *ProDQN* algorithm while Fig. 11 shows the individual prices by *PQR*. Both approaches proposed in our system are able to learn and adjust the price over time to improve the buyers' perceived value while considering the sellers' competitiveness. The competitiveness of a seller is a function of buyers' reference prices, the seller production, and their location in the system (e.g., loss w.r.t. buyers). Note that, although both algorithms adjust prices based on the output of the transactions, which indirectly reflects the sellers' competitiveness, the evolution of prices under *ProDQN* and *PQR* may differ. When taken collectively, both algorithms are able to find a balance between buyers perceived utility and sellers reward. To support this statement, we show in Fig. 12 the average sellers' prices, after a year of execution of the algorithms, with different system sizes. Both approaches converge towards similar prices, with negligible differences as the system grows.

Experimental Scenario 4: In this scenario we test the scalability of the proposed approaches with respect to the system size through a year-long aggregated analysis. Specifically, we increase the system proportionately from $|S| = 5$ sellers and $|B| = 5$ buyers to $|S| = 20$ sellers and $|B| = 20$ buyers. Figs. (13)-(14) show the buyers' total

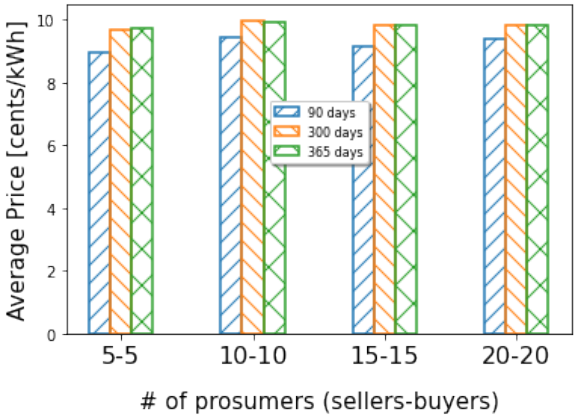

Fig. 10. Individual prices for *ProDQN*.

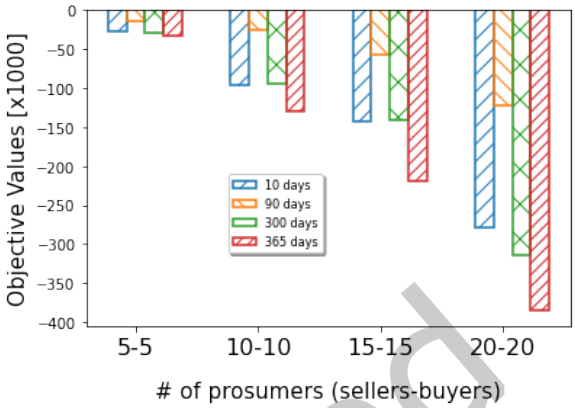

Fig. 11. Individual prices for *PQR*.

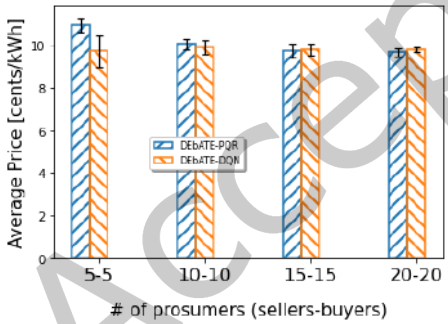

Fig. 12. Avg. price comparison for different network sizes between *PQR* and *ProDQN*.

perceived value and the sellers' reward, respectively, over a year. By considering the loss-averse and risk-seeking PT-value functions, *DEbATE − PQR* and *DEbATE − DQN* achieve an increasing advantage as the system size increases compared to *Rule*, for both sellers and buyers. *Zhu*, as previously discussed, creates a heavily biased market that penalizes buyers and favors sellers. As a numerical example, *DEbATE − PQR* achieves as much as 26% increase in buyers' perceived value while ensuring 7% profit improvement for sellers compared to *Rule*. Similarly, *DEbATE − DQN* achieves 8% more profit for sellers with almost 23% more in buyer's perceived utility.

5 LITERATURE REVIEW

In the recent years, P2P energy trading has attracted significant attention from the research community [26, 37]. In this section we review the main modeling techniques, from an optimization perspective, adopted in the domain of P2P energy trading. The authors of [30] study a P2P energy sharing model with price-based demand response. The approach introduces an energy sharing provider that coordinates the sharing activities, including a dynamical pricing model. The paper formulates a bi-level optimization problem. The upper level finds a fixed price point for each prosumer. The lower level is a minimization problem, solved individually by each prosumer, with the objective of minimizing the cost.

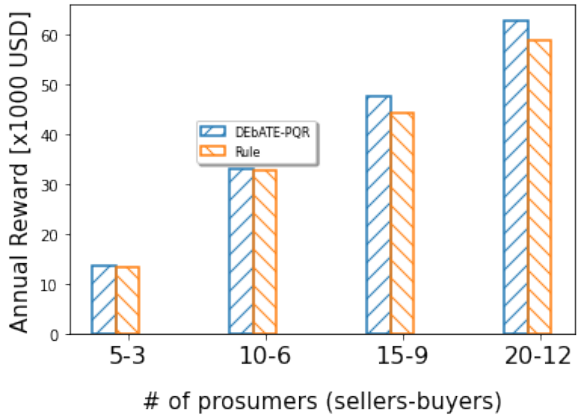

Fig. 13. Objective values for buyer vs. network size.

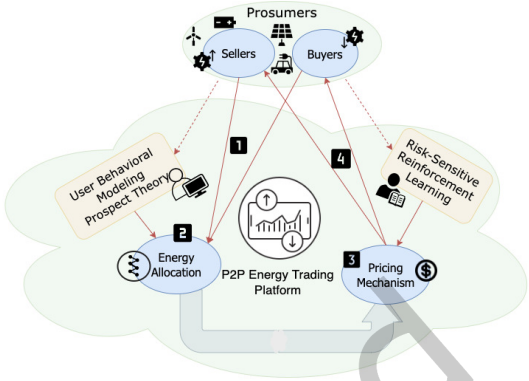

Fig. 14. Total rewards for sellers vs. network size.

More recent works, adopt *game-theory* to optimize the system operation [47, 49, 51]. Specifically, an auction theory based approach is proposed in [51]. Each participant submits a bid price to the auctioneer, and an optimal allocation is produced. Based on the allocation, each user determines the optimal bid price. In [47], a canonical coalition game is devised to share energy among groups of peers through social coalition for ensuring users' sustainable participation in the market. The approach defines a value function to evaluate the worth of a coalition. The objective is to incentive the formation of stable coalitions with appropriate pricing depending on energy demand and production. The authors of [49] introduce a cooperative Stackelberg game, where the grid is the leader that sets the prices and prosumers are the followers. Logarithm functions are used to model the utility of prosumers, while simple derivates are adopted to maximize such utility. The grid sets the price to incentivize P2P trading during peak hours.

In [5] a P2P trading scheme for a voltage-constrained network is proposed. Prosumers can form coalitions to negotiate and decide the energy trading parameters. Then, the *Myerson value rule* is used to allocate energy and price among prosumers. The distribution network is modeled through a graph where nodes are characterized by voltage parameters that should not be exceeded. Thus, the optimization objective is to minimize the total local power exported to the prosumers, under a linearization of the voltage constraints. Other works adopt a similar modeling of the physical aspects of P2P energy trading, such as power loss minimization, voltage regulation, and network constraints [5, 22, 33, 36].

The works mentioned above, however, largely overlook the user behavioral aspects in designing their solutions. The optimization problems assume that users are rational and constantly interacting with the system. A few recent works take into account users' preferences and limited capabilities [1, 44]. In these papers, the energy allocation problem is modeled as a matching problem, and specifically as an extension of the general assignment problem. Users have preferences which are learned using reinforcement learning. Following the bounded rationality principle, only a limited number of potential exchanges are given to the users, based on the learned preferences. Conversely, an automated approach is proposed in [4] to allocate energy and price between prosumers. Here, the objective is to match demand and production with a low computational complexity solution. Therefore, the authors propose a greedy algorithm that sorts sellers buy selling price, and matches buyers in the order of registration. We use this approach as a comparison in the experimental results.

Recently, there have been a few efforts in integrating *Prospect Theory* (PT) [25] in energy trading, with the objective to better modeling users' behavior and perception. The study in [16] formulates an optimization problem

that maximizes the sum of the users' prospect theory utilities under techno-economic constraints. Although the problem is a mixed-integer-linear programming (MILP) problem, and thus could have a high complexity, the authors adopt a MILP solver to find the optimal solution. The authors of [53] develop a nested market clearing algorithm for inter- and intra-microgrid energy trading. The prospect theory utility function is divided into subjective and objective uncertainty, and modeled accordingly. The optimization problem is formulated as a distributionally robust model which aims at minimizing the sum of the expected value of the costs and the prospect theory functions for users, under various system's and users' constraints. The problem is solved using linearization and an iterative algorithm that exploits Lagrangian functions.

Although these papers address the user behavioral modeling in some ways, they require active participation from users and also assume that such behavior (e.g., the prospect theory parameters) is homogeneous for all the users. Social science studies, have shown that such assumptions do not hold in practice [13]. In this paper, in light of the above mentioned limitations, we propose a novel demand/production matching problem, and a differential evolution algorithm to solve it, that take into account the users' individual perceptions through the perceived utility, modeled through prospect theory. Additionally, we define a Q-learning based algorithms that learn competitive prices while taking into account the sellers' perceived utility.

6 CONCLUDING REMARKS

In this paper, we bring together the concept of perceived utility from behavioral economics and reinforcement learning into P2P energy trading. Unlike existing literature, we propose an automated and dynamic P2P energy trading problem that maximizes the perceived value for buyers while simultaneously learning the optimal selling price for sellers. Given the non-linear and non-convex nature of the problem, we propose a novel Differential Evolution-based metaheuristic algorithm, called *DEbATE*. *DEbATE* is paired with a prospect theory enhanced Q-learning algorithm, called *PQR*, to adjust the selling price over time. Given the limitations of the tabular $Q-learning$ approach of *PQR*, we propose a Deep Q-Network-based algorithm called *ProDQN* that proposes a novel loss function based on PT value function to model the seller's perceived utility. Results show the advantages of the proposed approaches with respect to a state of the art solution using real energy consumption and production data.

This work shows that integrating concepts from behavioral economics and reinforcement learning can lead to more efficient and effective energy exchange in peer-to-peer (P2P) energy trading systems. It is also supported by the results showing how the proposed algorithms, i.e., *DEbATE*, *PQR*, and *ProDQN*, outperform existing solutions in maximizing perceived value for buyers as well as learning the optimal selling prices for sellers. In our future work we will extend the proposed approaches. Specifically, we will consider reward signals that allow agents to converge faster towards the optimal policy. Additionally, we will consider privacy concerns through the use of blockchain technology. This will provide a secure trading platform for participating prosumers.

ACKNOWLEDGMENTS

This work is partially supported by the NSF grants EPCN-1936131 and the NSF CAREER grant CPS-1943035.

REFERENCES

[1] Vincenzo Agate, Atieh R Khamesi, Simone Silvestri, and Salvatore Gaglio. 2020. Enabling peer-to-peer user-preference-aware energy sharing through reinforcement learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 1–7.

[2] International Energy Agency. 2022. IEA: Electricity Information Overview. https://www.iea.org/reports/electricity-information-overview/electricity-production

[3] Rosemary Alden, Ashutosh Timilsina, Simone Silvestri, and Dan Ionel. 2023. V2G Optimization for Dispatchable Residential Load Operation and Minimal Utility Cost. In *Transportation Electrification Conference & Expo (ITEC)*. IEEE.

[4] M Imran Azim, SA Pourmousavi, Wayes Tushar, and Tapan K Saha. 2019. Feasibility study of financial P2P energy trading in a grid-tied power network. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, Atlanta,GA,USA, 1–5.

[5] M Imran Azim, Wayes Tushar, and Tapan Kumar Saha. 2021. Coalition graph game-based P2P energy trading with local voltage management. *IEEE Transactions on Smart Grid* 12, 5 (2021), 4389–4402.
[6] Vladimír Baláž, Viera Bačová, Eva Drobná, Katarína Dudeková, and Kamil Adamík. 2013. Testing prospect theory parameters. *Ekonomicky časopis* 61 (2013), 655–671.
[7] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. 2013. *Nonlinear programming: theory and algorithms*. John Wiley & Sons.
[8] Shameek Bhattacharjee, Venkata Praveen Kumar Madhavarapu, Simone Silvestri, and Sajal K Das. 2021. Attack context embedded data driven trust diagnostics in smart metering infrastructure. *ACM Transactions on Privacy and Security (TOPS)* 24, 2 (2021), 1–36.
[9] Enrico Casella, Atieh R Khamesi, Simone Silvestri, Denise A Baker, and Sajal K Das. 2022. Hvac power conservation through reverse auctions and machine learning. In *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 89–100.
[10] Enrico Casella, Eleanor Sudduth, and Simone Silvestri. 2022. Dissecting the Problem of Individual Home Power Consumption Prediction using Machine Learning. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, Finland, 156–158.
[11] Stefano Ciavarella, Jhi-Young Joo, and Simone Silvestri. 2016. Managing contingencies in smart grids via the internet of things. *IEEE Transactions on Smart Grid* 7, 4 (2016), 2134–2141.
[12] Jackson Codispoti, Atieh R Khamesi, Nelson Penn, Simone Silvestri, and Eura Shin. 2022. Learning from non-experts: an interactive and adaptive learning approach for appliance recognition in smart homes. *ACM Transactions on Cyber-Physical Systems (TCPS)* 6, 2 (2022), 1–22.
[13] Davide Contu, Elisabetta Strazzera, and Susana Mourato. 2016. Modeling individual preferences for energy sources: The case of IV generation nuclear energy in Italy. *Ecological Economics* 127 (2016), 37–58.
[14] Thomas H Cormen. 2009. *Introduction to algorithms*. MIT press.
[15] Valeria Dolce, Courtney Jackson, Simone Silvestri, Denise Baker, and Alessandra De Paola. 2018. Social-behavioral aware optimization of energy consumption in smart homes. In *2018 14th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 163–172.
[16] Sobhan Dorahaki, Masoud Rashidinejad, Seyed Farshad Fatemi Ardestani, Amir Abdollahi, and Mohammad Reza Salehizadeh. 2021. A Peer-to-Peer energy trading market model based on time-driven prospect theory in a smart and sustainable energy community. *Sustainable Energy, Grids and Networks* 28 (2021), 100542.
[17] Peter E Earl. 2016. Bounded Rationality in the Digital Age. In *Minds, Models and Milieux*. Springer, England, UK, 253–271.
[18] Georges El Rahi, Walid Saad, Arnold Glass, Narayan B Mandayam, and H Vincent Poor. 2016. Prospect theory for prosumer-centric energy trading in the smart grid. In *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, ISGT, Minnesota, USA, 1–5.
[19] Lisa Fleischer, Michel X Goemans, Vahab S Mirrokni, and Maxim Sviridenko. 2006. Tight approximation algorithms for maximum general assignment problems. In *Proceedings of 17th ACM-SIAM symposium on Discrete algorithm*. Society for Industrial and Applied Mathematics, Philadelphia, PA, United States, 611–620.
[20] Craig R Fox and Russell A Poldrack. 2009. Prospect theory and the brain. In *Neuroeconomics*. Elsevier, London,UK, 145–173.
[21] Gerd Gigerenzer and Reinhard Selten. 2002. *Bounded rationality: The adaptive toolbox*. MIT press, MA,USA.
[22] Jaysson Guerrero, Archie C Chapman, and Gregor Verbič. 2018. Decentralized P2P energy trading under network constraints in a low-voltage network. *IEEE Transactions on Smart Grid* 10, 5 (2018), 5163–5173.
[23] Pecan Street Inc. 2019. www.pecanstreet.org
[24] Olamide Jogunola, Augustine Ikpehai, Kelvin Anoh, Bamidele Adebisi, Mohammad Hammoudeh, Sung-Yong Son, and Georgina Harris. 2017. State-of-the-art and prospects for peer-to-peer transaction-based energy system. *Energies* 10, 12 (2017).
[25] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, Singapore, 99–127.
[26] Dileep Kalathil, Chenye Wu, Kameshwar Poolla, and Pravin Varaiya. 2017. The sharing economy for the electricity storage. *IEEE Transactions on Smart Grid* 10, 1 (2017), 556–567.
[27] Atieh R Khamesi and Simone Silvestri. 2020. Reverse auction-based demand response program: A truthful mutually beneficial mechanism. In *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 427–436.
[28] Atieh R Khamesi, Simone Silvestri, Denise A Baker, and Alessandra De Paola. 2020. Perceived-Value-driven Optimization of Energy Consumption in Smart Homes. *ACM Transactions on Internet of Things* 1, 2 (2020), 1–26.
[29] Weicong Kong, Zhao Yang Dong, Youwei Jia, David J Hill, Yan Xu, and Yuan Zhang. 2017. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid* 10, 1 (2017), 841–851.
[30] N. Liu, X. Yu, C. Wang, C. Li, L. Ma, and J. Lei. 2017. Energy-Sharing Model With Price-Based Demand Response for Microgrids of Peer-to-Peer Prosumers. *IEEE Transactions on Power Systems* 32, 5 (Sep. 2017), 3569–3583.
[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015),

529–533.
[32] Michele Moretti, S Njakou Djomo, Hossein Azadi, Kristof May, Kristof De Vos, Steven Van Passel, and Nele Witters. 2017. A systematic review of environmental and economic impacts of smart grids. *Renewable and Sustainable Energy Reviews* 68 (2017), 888–898.
[33] Mohammad Nasimifar, Vahid Vahidinasab, and Mohammad Sadegh Ghazizadeh. 2019. A peer-to-peer electricity marketplace for simultaneous congestion management and power loss reduction. In *2019 Smart Grid Conference (SGC)*. IEEE, Tehran, Iran, 1–6.
[34] NREL. 2019. Solar Resource Data. pvwatts.nrel.gov/pvwatts.php
[35] Yael Parag and Benjamin Sovacool. 2016. Electricity market design for the prosumer era. *Nature Energy* 1 (March 2016), 16032. https://doi.org/10.1038/nenergy.2016.32
[36] Amrit Paudel, LPMI Sampath, Jiawei Yang, and Hoay Beng Gooi. 2020. Peer-to-peer energy trading in smart grid considering power losses and network fees. *IEEE Transactions on Smart Grid* 11, 6 (2020), 4727–4737.
[37] Frederik Plewnia. 2019. The energy system and the sharing economy: Interfaces and overlaps and what to learn from them. *Energies* 12, 3 (2019), 339.
[38] Marc Oliver Rieger, Mei Wang, and Thorsten Hens. 2017. Estimating cumulative prospect theory parameters from an international survey. *Theory and Decision* 82, 4 (2017), 567–596.
[39] Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. 2014. Risk-sensitive reinforcement learning. *Neural computation* 26, 7 (2014), 1298–1328.
[40] Esteban A Soto, Lisa B Bosman, Ebisa Wollega, and Walter D Leon-Salas. 2021. Peer-to-peer energy trading: A review of the literature. *Applied Energy* 283 (2021), 116268.
[41] Rainer Storn and Kenneth Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11, 4 (1997), 341–359.
[42] Wadim Strielkowski. 2019. *Social Impacts of Smart Grids: The Future of Smart Grids and Energy Market Design.* Elsevier, Netherlands.
[43] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press, Cambridge,MA,USA.
[44] Ashutosh Timilsina, Atieh R Khamesi, Vincenzo Agate, and Simone Silvestri. 2021. A Reinforcement Learning Approach for User Preference-Aware Energy Sharing Systems. *IEEE Transactions on Green Communications and Networking* 5, 3 (2021), 1138–1153.
[45] Ashutosh Timilsina and Simone Silvestri. 2022. Prospect Theory-inspired Automated P2P Energy Trading with Q-learning-based Dynamic Pricing. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 4836–4841.
[46] Wayes Tushar, Tapan Kumar Saha, Chau Yuen, M Imran Azim, Thomas Morstyn, H Vincent Poor, Dustin Niyato, and Richard Bean. 2020. A coalition formation game framework for peer-to-peer energy trading. *Applied Energy* 261 (2020), 114436.
[47] Wayes Tushar, Tapan Kumar Saha, Chau Yuen, Paul Liddell, Richard Bean, and H Vincent Poor. 2018. Peer-to-peer energy trading with sustainable user participation: A game theoretic approach. *IEEE Access* 6 (2018), 62932–62943.
[48] Wayes Tushar, Tapan Kumar Saha, Chau Yuen, Thomas Morstyn, Malcolm D McCulloch, H Vincent Poor, and Kristin L Wood. 2019. A motivational game-theoretic approach for peer-to-peer energy trading in the smart grid. *Applied energy* 243 (2019), 10–20.
[49] Wayes Tushar, Tapan Kumar Saha, Chau Yuen, Thomas Morstyn, H Vincent Poor, Richard Bean, et al. 2019. Grid influenced peer-to-peer energy trading. *IEEE Transactions on Smart Grid* 11, 2 (2019), 1407–1418.
[50] Wayes Tushar, Tapan K Saha, Chau Yuen, David Smith, and H Vincent Poor. 2020. Peer-to-peer trading in electricity networks: an overview. *IEEE Transactions on Smart Grid* 11, 4 (2020), 3185–3200.
[51] Wayes Tushar, Chau Yuen, Hamed Mohsenian-Rad, Tapan Saha, H Vincent Poor, and Kristin L Wood. 2018. Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches. *IEEE Signal Processing Magazine* 35, 4 (2018), 90–111.
[52] U.S. Department of Energy. 2008. *The Smart Grid: An Introduction.* Technical Report. U.S. Department of Energy.
[53] Yuanxing Xia, Qingshan Xu, Yu Huang, Yihan Liu, and Fangxing Li. 2022. Preserving privacy in nested peer-to-peer energy trading in networked microgrids considering incomplete rationality. *IEEE Transactions on Smart Grid* 14, 1 (2022), 606–622.
[54] Yunting Yao, Ciwei Gao, Tao Chen, Jianlin Yang, and Songsong Chen. 2021. Distributed electric energy trading model and strategy analysis based on prospect theory. *International Journal of Electrical Power & Energy Systems* 131 (2021), 106865.
[55] T. Zhu, Z. Huang, A. Sharma, J. Su, D. Irwin, A. Mishra, D. Menasche, and P. Shenoy. 2013. Sharing renewable energy in smart microgrids. In *2013 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, PA,USA, 219–228.

A APPENDIX

Theorem A.1. *The optimization problem in Eq.* (2) *is NP-hard.*

Proof. We present a reduction from the Generalized Assignment Problem (GAP) [19] as a proof of NP-hardness of our optimization problem in Eqs. (2)- (2f). In a *general instance* of GAP, there are n tasks and m processors. Each processor i has a resource budget given by e_i. By assigning task j to processor i, we obtain a profit p_{ij} while

consuming g_{ij} amount of resources. A task can only be assigned to a single process, and therefore, the goal is to find the assignment that provides maximum profit given the resource budget of the processors. The GAP can be formulated as an integer linear programming problem:

$$\max \quad \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}x_{ij} \tag{9}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} g_{ij}x_{ij} \leq e_i, \qquad \forall i \tag{9a}$$

$$\sum_{i=1}^{m} x_{ij} = 1, \qquad \forall j \tag{9b}$$

$$x_{ij} \in \{0,1\} \qquad \forall i,j \tag{9c}$$

From a general GAP instance, we can create a *reduced instance* of our problem as follows. We create a buyer for each task and a seller for each processor. We set $(1+\ell_{ij})d_j = g_{ij}$ and set the energy production of a seller i to e_i. We also set $l_{max} = \infty$ so that all exchanges are possible (i.e., a task can be assigned to any processor). An important difference between our reduced problem and the GAP is that the decision variables x_{ij} are continuous instead of discrete. However, infinitesimal exchanges are not allowed in our system, as they need to be greater than or equal to μ_j. By setting $\mu_j = d_j$, the constraint in Eq. (2e) forces the decision variable x_{ij} to be either 0 or 1, same as binary decision variable z_{ij}. Additionally, it also forces the system to assign a buyer (i.e., a task in the GAP problem) to a single seller.

We set all the loss-aversion parameters $(k_{+,.}, k_{-,.})$ to 1 and the risk-sensitive parameters $(\zeta_{+,.}, \zeta_{-,.})$ to 1. We also set $\rho_j = 0$ for all sellers in S. In summary, the objective function becomes linear, i.e.,

$$\sum_{j\in B}\sum_{i\in S}(\rho_{gs} - \rho_i)d_j x_{ij} - \sum_{j\in B}\sum_{i\in S}\rho_{gs}d_j$$

The term $-\sum_{j\in B}\sum_{i\in S}\rho_{gs}d_j$ is just a constant, and can thus be ignored for the purpose of the maximization problem. Now, by setting $(\rho_{gs} - \rho_i)d_j = p_{ij}$, we have successfully reduced the objective function to

$$\sum_{j\in B}\sum_{i\in S} p_{ij}x_{ij}$$

As a result, the solution of our reduced problem provides the assignment that maximizes the profit within the constrained processors' resources. Therefore, our problem is at least as hard as GAP, and thus it is NP-Hard. □

Lemma A.2. *The time complexity of the DEbATE algorithm is* $O(G_{max} \times NP \times |S||B|)$.

Proof. The complexity is dominated by the *while* loop (lines 3 – 19), which is executed G_{max} times. Within this loop, the for–loop (lines 4 – 17) does $|\mathcal{X}| = NP$ total iterations. In each iteration, the inner for–loop (lines 7 – 12) iterates over the sets S and B, and only contains constant operations. Similarly, checking the constraints (lines 13 – 15) requires to iterate over the same sets. Finally, calculating the function $f(.)$ (line 16) has cost $|B|$. Overall, the complexity is $O(G_{max} \times NP \times (|S||B| + 3|S||B| + |B|)) = O(G_{max} \times NP \times |S||B|)$

□