energies

MDPI

Article

Cost-Optimized Microgrid Coalitions Using Bayesian Reinforcement Learning

Mohammad Sadeghi, Shahram Mollahasani and Melike Erol-Kantarci *

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada; msade033@uottawa.ca (M.S.); smollah2@uottawa.ca (S.M.)

* Correspondence: melike.erolkantarci@uottawa.ca; Tel.: +1-613-562-5800 (ext. 6693)

Abstract: Microgrids are empowered by the advances in renewable energy generation, which enable the microgrids to generate the required energy for supplying their loads and trade the surplus energy to other microgrids or the macrogrid. Microgrids need to optimize the scheduling of their demands and energy levels while trading their surplus with others to minimize the overall cost. This can be affected by various factors such as variations in demand, energy generation, and competition among microgrids due to their dynamic nature. Thus, reaching optimal scheduling is challenging due to the uncertainty caused by the generation/consumption of renewable energy and the complexity of interconnected microgrids and their interplay. Previous works mainly rely on modeling-based approaches and the availability of precise information on microgrid dynamics. This paper addresses the energy trading problem among microgrids by minimizing the cost while uncertainty exists in microgrid generation and demand. To this end, a Bayesian coalitional reinforcement learning-based model is introduced to minimize the energy trading cost among microgrids by forming stable coalitions. The results show that the proposed model can minimize the cost up to 23% with respect to the coalitional game theory model.

Keywords: machine learning; Bayesian reinforcement learning; microgrid; smart grid

Citation: Sadeghi, M.; Mollahasani, S.; Erol-Kantarci, M. Cost-Optimized Microgrid Coalitions Using Bayesian Reinforcement Learning. *Energies* **2021**, *14*, 7481. https://doi.org/10.3390/en14227481

Academic Editor: Moez Esseghir

Received: 21 September 2021
Accepted: 27 October 2021
Published: 9 November 2021

1. Introduction

The overall demand for energy consumption has drastically increased over recent years, and it is also expected to reach up to 1000 Exajoule by the end of 2050 [1]. Governments are trying to enhance their energy generation capabilities by considering green and smart models to satisfy the massive energy demand. Therefore, energy generation and consumption models require a fundamental transformation in order to employ these capabilities in traditional power systems. By using smart grids and the advances in information and communications technologies (ICTs), a strong foundation is generated for transforming unidirectional power and information flow into a distributed bidirectional power and information system known as a transactive energy framework [2,3]. Transactive energy can be categorized into (i) transactive network management for organizing the energy supply chain, (ii) transactive control for controlling and managing the energy generation/consumption rate and (iii) peer-to-peer (p2p) energy market for allowing customers to trading energy among themselves [4].

One of the promising characteristics of microgrids is the possibility of p2p energy trading with each other or the utility grid. Energy trading can be carried out by transferring surplus energy from a microgrid to a close-by microgrid, which has been a well-known research topic in field of the smart grid since the 2010s [5]. Generally, we can describe the energy trading problem as a group of interconnected microgrids exchanging their surplus energy to serve the loads in other microgrids. The microgrids are also connected to the macrogrid, and energy trading can be conducted between microgrids and the macrogrid or among themselves. Some microgrids may have surplus energy at different time intervals

and prefer to sell their energy, while others suffer from a lack of supply and wish to buy it. This system can be modeled as a game theory problem and tackled with game-theoretical or learning approaches [6–8].

Although energy trading has been explored to some degree, energy trading under uncertainty has been less explored. This paper investigates the energy trading problem among microgrids where each microgrid has different levels of energy surplus or demand in each epoch. Additionally, the dynamic nature of the energy levels causes uncertainty in our system. We employ a Bayesian reinforcement-based coalition formation scheme for energy trading among microgrids to deal with this uncertainty. This algorithm was first introduced in [9], and an application of this model was also developed for device-to-device communications in wireless networks [10]. In this work, we develop the Bayesian reinforcement learning model, which enhances the conventional Bayesian coalition formation by learning from past observations and experiences. We then employ this approach in the energy trading problem among microgrids under uncertainty. We compared the proposed method with two Bayesian reinforcement learning-based models [10], Q-learning [11], Bayesian coalition formation [11], and conventional coalition formation game theory [6]. The results show up to 23% improvements in cost minimization compared to the coalitional game theory-based method.

The rest of this work is organized as follows. Recent works are summarized in Section 2. In Section 3, the system model is demonstrated. In Section 4, the Bayesian coalition formation game (BCG) scheme is illustrated. In Section 5, the Bayesian coalitional reinforcement learning (BCRL) based scheme is proposed. In Section 6, the numerical results are evaluated, and finally, the conclusions are presented at the end.

2. Related Work

Game-theoretic methods have been widely employed for energy trading in microgrids. In [12], a game-theoretic approach is proposed for distributed energy trading between microgrids. In this study, a set of interconnected microgrids aim to exchange energy with each other and also with the macrogrid. In [13], a priority-based energy trading game is proposed in which buyers are prioritized according to the past contributions of the buyers and their current demands. In [14], a Stackelberg game is designed with a central power station as the single leader and multiple followers who want to sell their extra energy to this central station. In [15], the authors develop a model based on the repeated game, which lets microgrids choose a strategy with a probability for trading energy in the market in a way that their average revenue is maximized.

Coalitional game theory is a subset of game theory in which players cooperate to maximize a shared payoff and then distribute the received payoff among the players. In several studies, the energy trading problem among microgrids has been modeled as a coalitional game theory problem. In this approach, microgrids can cooperate by forming coalitions for a specific period in which some microgrids with surplus energy supply the others that require energy. Table 1 summarizes the research attempts that investigate the energy trading problem using coalitional game theory.

In [6], for the first time, the energy trading problem in the microgrid community is investigated using coalitional game theory to minimize the power loss. The idea of coordinated operation of cooperative microgrids is studied in [16]. Although [6], only focuses on energy loss, the authors expanded the objective functions to maximize microgrids' expected profits and usage while minimizing power loss and consumers' discomfort. In [17], the authors proposed a nucleolus-based approach to fairly distribute the payoff among microgrids for transactive energy management in microgrid communities locally. In [18], the authors proposed coalitional-based energy trading where, in each coalition, an auction-based matching is employed to calculate the utility of the coalition, and then the coalition formation technique is used to partition the microgrids into coalitions. In [19], the authors designed the energy trading scheme in two stages. First, microgrids form coalitions, and then a matching game is used to schedule energy exchange in each coalition.

Machine learning algorithms have proven useful in a wide range of applications such as computer vision, sentiment analysis, self-organized systems, and robotics. However, it is not straightforward to use the same algorithms in AI-enabled smart grids. Existing machine learning techniques need to be tailored to meet the smart grid and microgrids' needs. In [20], the authors propose two learning automata-based methods for optimal power management in smart grids. In [21], the authors propose a dynamic demand response and distributed generation management method for a residential microgrid community. In [22], a fully distributed learning approach is proposed for optimal reactive power dispatch. In this method, a multi-agent Q-learning algorithm is employed that minimizes the active power loss and satisfies the bus voltage range and reactive power generation constraints. In [23], the temporal difference reinforcement learning approach is used to achieve the optimal control policy for residential energy storage. The problem of dynamic pricing in smart grids with reinforcement learning methods is addressed in [24]. The authors propose reinforcement-based dynamic pricing and energy consumption scheduling to help energy providers and consumers learn their best strategies.

Energy trading is also among the problems that can be tackled with machine learning approaches, specifically using reinforcement learning models such as Q-learning, Bayesian reinforcement learning, and deep reinforcement learning. In [8], a hot-booting Q-learning-based approach is implemented to achieve the Nash equilibrium of the dynamic repeated energy trading game. In [25], the authors improve [8] by designing a deep Q-network-based approach. In our prior work, [11] we proposed a Bayesian coalitional algorithm that helps agents make a system of beliefs about the types of other agents. In contrast, in [26], agents can learn from their experience by using a Bayesian reinforcement learning technique; however, the proposed model suffers from the lack of a belief system. In this study, we propose a comprehensive Bayesian reinforcement learning framework for the problem of coalition formation in microgrid communities, which helps agents make a system of beliefs about the types of other agents and learn from their past experiences simultaneously.

Table 1. A sumary of energy trading studies.

Paper	Objective	Elements	Uncertainty	Methods
[6]	Power Loss	Macrogrid, microgrids	✗	CG
[16]	Energy Management	Macrogrid, microgrids	✗	CG
[17]	Energy Management	Macrogrid, microgrids	✗	CG
[18]	Cost	Macrogrid, microgrids	✗	CG
[19]	Cost	Macrogrid, microgrids	✗	CG
[11]	Power Loss	Macrogrid, microgrids, EVs	✓	BCG
[26]	Power Loss	Macrogrid, microgrids	✓	CG, BRL

3. System Model

In this work, we consider a network of M interconnected microgrids while each microgrid is also connected to the main utility grid known as the macrogrid as shown in Figure 1. The amount of generated energy by microgrid $m \in M$ and its demand are presented by g_m and d_m, respectively. Therefore, we can find the total surplus or shortage energy of each microgrid as $q_m = g_m - d_m$, which represents the energy that each microgrid is required to export to or import from the network. As a result, microgrids initiate an energy trading process among each other and with the macrogrid to satisfy their export/import requirements. Due to the dynamicity of the system, each microgrid can be either a seller or buyer of energy during each epoch.

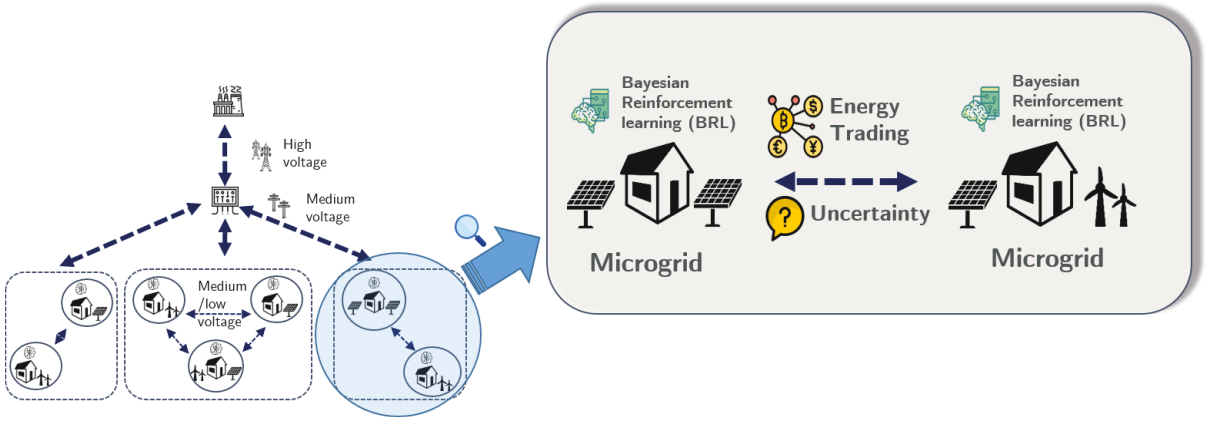

Figure 1. Block diagram of a system of microgrids. The figure illustrates that due to the dynamicity of the system, different coalitions can be formed during each epoch.

The process of energy trading, either among microgrids or with the macrogrid, imposes a variety of costs. In the proposed system, we assume that each energy transaction is associated with two sets of costs. The first set corresponds to the costs resulting from the power loss in the line, transformer loss from high to medium or low voltages, maintenance cost, etc. We call this set of costs as operational costs. Additionally, there are other hidden costs associated with energy trading among microgrids—for example, a case where two far away microgrids want to trade energy among themselves. This energy trading is not always feasible through the direct power line since it is not possible to have direct lines between all microgrids in practice. Therefore, we assume that direct lines are only deployed between close-by microgrids. When there are no direct links, microgrids have to make their energy transactions through a series of intermediate microgrid links or transfer their energy through the macrogrid (the seller microgrid sends its surplus energy to the macrogrid, and the buyer microgrid receives the transferred energy from the macrogrid). One of the interesting capabilities of microgrids is trading energy only among themselves without relying on a central macrogrid while in islanding mode. Islanding will be at risk by any reliance on energy trading with the macrogrid, resulting in unexpected costs. We address all of these third parties involved in energy trading with the unpredicted costs (virtual costs) as the second set of costs. Therefore, the total cost of power transaction E_{mn} from m-th microgrid to n-th microgrid is given by:

$$S_{mn} = wd_{mn}E_{mn} + \delta PL(E_{mn}) \tag{1}$$

where d_{mn} represents the length of power lines that need to be used for transferring energy between the m-th and n-th microgrids and δ shows the scaling factor. E_{mn} is the power that is being traded plus the loss that happens during trading. $PL(E_{mn})$ denotes the power loss in trading energy between m-th and n-th microgrids and scale. w represents a weighting coefficient associated with the virtual cost and can be calculated as:

$$w = \begin{cases} w_s & m, n \neq 0 \ and \ d_{mn} \leq d_{tr} \\ w_l & m, n \neq 0 \ and \ d_{mn} > d_{tr}. \\ w_0 & m = 0 \ or \ n = 0 \end{cases} \tag{2}$$

The virtual cost is a function of distance and energy that is weighted by the parameter w. This parameter is fixed to a lower value w_s for energy trading among microgrids which are closer than threshold d_{tr}, and a higher value w_l for the rest [27]. We assume that distant microgrids that are further than the threshold have no direct link in between. Consequently, the virtual cost increases compared to close-by microgrids. w_0 is the weight factor for energy transactions with the macrogrid.

Power loss is defined as below, in which R_{mn} represents the resistance of line per km in energy trading between microgrids m and n [6].

$$PL(E_{mn}) = R_{mn} d_{mn} \frac{{E_{mn}}^2}{U_m^2} + \rho E_{mn}, \tag{3}$$

where U_m and ρ denote voltage and the fraction of power loss in the transformer at the interconnection point between the microgrids and the macrogrid (macro station), respectively. E_{mn} is the trading power required to deliver the total surplus/demand q_j of microgrid n to microgrid m and can be obtained as follows:

$$E_{mn} = q_n + PL(E_{mn}) \tag{4}$$

The main objective of this system is to minimize the total cost. Therefore:

$$\begin{gathered} \min \sum_{m=0}^{M} \sum_{n=0}^{M} S_{mn} \\ s.t: \sum_{m=1}^{M} q_m + \sum_{m=0}^{M} \sum_{n=0}^{M} P(E_{mn}) = \sum_{n=0}^{M} E_{0n} - \sum_{m=0}^{M} E_{m0} \end{gathered} \tag{5}$$

Energy trading among nearby microgrids decreases the total cost with respect to trade energy among distant microgrids and macrogrid. Therefore, forming groups of close-by microgrids (also known as coalition formation) to trade energy in their groups is a promising approach that reduces the overall cost. In addition, there is no transformer loss in the operating range of energy trading among microgrids (low or medium voltage) [28].

We can formulate a coalition with a pair (C, v_C). C expresses the coalition in which coalition members cooperate to gain a higher coalitional value v_C [29]. In this paper, we consider the total cost of energy trading among coalition members plus the cost of trading extra energy with the macrogrid as the coalition value. The objective is to minimize the cost. Therefore, the coalition value as the negative form of cost can be formulated as follows:

$$v_C = -\sum_{m=0}^{|C|} \sum_{n=0}^{|C|} S_{mn}, \tag{6}$$

where $|C|$ shows the number of members of coalition C. Index 0 expresses any transactions with the main grid. When the coalition is formed, energy transfer among the coalition members needs to be scheduled to minimize the total cost in the coalition. Therefore, the coalition payoff in our considered system is defined as the maximum achievable coalition value $v_{max}(C)$ which is given by:

$$v_C^{max} = \max\{-\sum_{i=0}^{|C|} \sum_{j=0}^{|C|} S_{mn}\}. \tag{7}$$

4. Bayesian Coalition Formation Game

In this section, we propose a Bayesian coalition formation game (BCFG) that tackles the uncertainty in the power level of the microgrids.

4.1. Game Formulation

The Bayesian coalition formation game can be characterized by a set of agents (M), a set of agent types ($T^m \in \vec{T}$), a set of agent beliefs (B^m), a set of coalition actions ($\vec{A}^{C_k}$), a set of outcomes known as states ($\vec{S}$), and the reward functions (u^m).

To employ BCFG in our problem, we can describe BCFG as a cost minimization model in a microgrid community where a set of M rational microgrid agents is involved in the coalition formation game. A coalition C_k represents a set of microgrids that allows them to trade energy among themselves. The m-th microgrid's type T^m stands for the microgrid's

power level. Each microgrid is only aware of its type (T^m) but not the types of other microgrids. The m-th microgrid's beliefs about the types of other players are denoted by $B^m(T^{-m})$ that consists of a joint distribution over $\vec{T}^{-m}$ which is the probability assigned to other agents about their types. We assume that any coalition of microgrids has a restricted set of coalition actions $\vec{A}^{C_k}$. A collective coalition action α^{C_k} is an action that is approved by all coalition members of C_k about a new member to join their coalition. The coalition action is only observable for the coalition members and hidden from agents in other coalitions.

We consider the coalition tag of each microgrid as its state in each iteration of the game. Therefore, agents' state vectors can be defined as $\vec{s} = (s^1, ..., s^M)$. Any state vector s corresponds to a joint reward $R^{C_k}(\vec{s}, T^{C_k})$ which is calculated as:

$$R^{C_k}(\vec{s}, T^{C_k}) = v_{C_k}^{max} = \sum_{m \in C_k} r^m(\vec{s}, T^{C_k}) \tag{8}$$

We use the proportional fair division method to distribute the coalition reward among coalition members, allocating each member a share of the coalition reward proportionate to their cost. Therefore, $r^m(s, T^{C_k})$ is defined as [29]:

$$r^m(\vec{s}, T^{C_k}) = \zeta_m \left(v_{C_k}^{max} - \sum_{n \in C_k} v(\{n\}) \right) + v(\{m\}). \tag{9}$$

where ζ_m is equal to $\frac{v(\{i\})}{\sum_{j \in C} v(\{j\})}$ and demonstrates the relative contribution of each microgrid. $v(\{j\})$ and v_C^{max} show a single member coalition.

4.2. Stability Notation

Like all cooperative games, in the coalitional game theory, players with a common interest or members of a specific coalition maximize their joint reward, known as the coalition value. We compute the value of coalition C_k with the members of type T^{C_k} as follows:

$$V(C_k|T^{C_k}) = \max_{\alpha^{C_k} \in \vec{A}^{C_k}} \sum_{s \in S} \Pr\{s|C_k, \alpha^{C_k}, T^{C_k}\} r^{C_k}(s, T^{C_k}) = \max_{\alpha^{C_k} \in \vec{A}^{C_k}} \sum_{s \in S} Q(C_k, \alpha^{C_k}|T^{C_k}) \tag{10}$$

where $\Pr\{s|C_k, \alpha^{C_k}, T^{C_k}\}$ represents the probability of transitioning to state s in coalition C_k with members of type T^{C_k} when taking action α^{C_k}. $Q(C_k, \alpha^{C_k}|T^{C_k})$ shows the long-term action value. It can be seen that $V(C_k|T^{C_k})$ is a function of the actual "type" of coalition members while the "type" of a microgrid is not known by the other microgrids inside the coalition. Therefore, to estimate the coalition value C_k, coalition members need to rely on their beliefs about the "type" of other players. We call this estimation the expected value of coalition C_k and coalition member m can compute its expected coalition value according to its beliefs B^m as follows:

$$V(C_k, B^m) = \max_{T^{C_k} \in \vec{T}^{C_k}} \sum_{s \in S} B^m(T^{C_k}) \sum_{s \in S} Q(C_k, \alpha^{C_k}|T^{C_k}) = \max_{\alpha^{C_k} \in \vec{A}^{C_k}} \sum_{s \in S} Q(C_k, \alpha^{C_k}, B^m) \tag{11}$$

where $Q(C_k, \alpha^{C_k}, B^m)$ demonstrates the expected value of coalition C_k when action α^{C_k} is taken while the system's belief is equal to B^m. Since all the microgrids have their specific systems of beliefs, it is common for microgrids to end up with different estimations of $Q(C_k, \alpha^{C_k}, B^m)$ and consequently $V(C_k, B^m)$. Therefore, none of the microgrids can reach the accurate estimations about the coalitional reward R^{C_k} and their share of reward r^m. To this end, players need a system to estimate their achievable rewards for cooperating in coalitional activity. We define demand D^m as the share of the coalitional value that microgrid m believes in receiving in the coalition. Having the coalition structure C_k with the

demand vector $\vec{D}^{C_k} = (D^1, D^2, ..., D^M)$, microgrid m's belief about the expected reward of microgrid j by taking action α^{C_k} can be estimated by:

$$Q_j^m(C_k, \alpha^{C_k}, \vec{D}^{C_k}) = \frac{D^j Q(C_k, \alpha^{C_k}, B^m)}{\sum\limits_{i \in C_k} D^i} \tag{12}$$

microgrid m expects a long-term reward by taking action α^{C_k} and it expects the demand vector as $\vec{D}^{C_k}$ which is defined by $Q^m(C_k, \alpha^{C_k}, \vec{D}^{C_k})$.

Considering the above-mentioned definitions, the concept of a strong Bayesian core (SBC) can be defined as follows [9].

Definition 1. *We assume that a tuple of a specific coalitional structure and a specific demand vector* $(C_k, \vec{D}^{C_k})$ *are in the SBC of a Bayesian coalition formation game if:*

- *No player believes there exists a better tuple than* $(C_{k'}, \vec{D}^{C'_k})$.
- *All of the coalition members accept it according to their beliefs about the expected rewards of other players.*

This definition can be formulated as follows:

$$Q^m(C_{k'}, a^{C_{k'}}, \vec{D}^{C_{k'}}) > Q^m(C_k, \alpha^{C_k}, \vec{D}^{C_k}) \tag{13}$$

and

$$Q_j^m(C_{k'}, a^{C_{k'}}, \vec{D}^{C_{k'}}) > Q_j^m(C_k, \alpha^{C_k}, \vec{D}^{C_k}) \tag{14}$$

where $m \in M$. Equation (13) demonstrates the preference of microgrid m for itself and (14) shows the preference of microgrid j believed by microgrid m.

4.3. Coalition Formation

In this section, we define the Bayesian coalition formation process that we present in this paper. We assume that negotiations among the microgrids to merge and split from coalitions happen over an infinite number of iterations. At every iteration, there is a pairing the of coalitional structure and demand vector named the coalitional agreement $(CS, \vec{D})$, which all players agree on. All the microgrids have the chance to modify this coalition agreement concerning their utility (a rational player changes the agreement to improve its utility). We call the microgrid m who attempt to change the coalitional agreement a proposer since it proposes to change the agreement in either one of the following ways:

- A proposer can stay in its current coalition C_k and propose a new demand D^m from the coalition.
- A proposer can decide to split from its current coalition and propose merging to other coalition $C_{k'}$ with new demand D^m.

The microgrids have the following finite set of actions (or negotiations options): (1) if a microgrid is a proposer the action is to make proposal $\pi_m^k = (C_k, \{D^i\}_{i \in C_k} \overset{\circ}{} D^m)$ which means joining (or staying in) coalition C_k with the new demand D^m. (2) If a microgrid is a responder to a proposal, it has the following action options: (i) either accept ($\kappa_k^m = 1$) or (ii) reject ($\kappa_k^m = 1$), in response to the presented proposal. We can summarize the proposition procedure as follows. At the beginning of every iteration, a proposer m is chosen randomly from all the microgrids with an equal probability of $1/M$. Then, the proposer presents the proposal π_m^k to join or stay in the coalition C_k with demand D^m. After that, members of the coalition C_k independently accept or reject the offered proposal without having any information regarding the action of other members. All the individual responder actions need to be unified in a single coalitional action to respond to the proposal. To this end, we

introduce function f, which maps all the responding actions into the coalitional action α^{C_k}. We define this coalitional action as follows:

$$\alpha^{C_k} = \begin{cases} f(\pi_k^m), & if \prod_{i \in C_k \setminus \{m\}} \kappa_k^m = 1 \\ f(C_k, \vec{D}^{C_k}), & otherwise \end{cases}. \quad (15)$$

This means that coalition members accept a proposal if all coalition members approve it; otherwise, the proposal will be rejected, and the existing coalitional agreement will be in effect.

We assume that all the players are rational, which means that the proposer submits a proposal that maximizes its expected reward. Meanwhile, because the other players are rational as well, they only accept a proposal that does not degrade their expected reward. Therefore, a rational proposal is to offer the maximum possible demand D_{max}^m that does not degrade the expected reward of other players according to the beliefs of the proposer about other players. This particular proposal is achievable for the proposed microgrid if:

$$\frac{D^j Q(C_k \cup \{m\}, \alpha^{C_k}, B^m)}{\sum_{i \in C_k \cup \{m\}} D^i} \geq Q_j^m, \ \forall j \in C_k \quad (16)$$

where $\alpha^{C_k} = f(\pi_k^m)$ and Q_j^m is the expected reward of microgrid n believed by microgrid m. If proposer microgrid m finds π_k^m to be feasible, it expects all the responders to accept the proposal according to its system of beliefs about others. It should be noted that this feasibility is just an expectation, and the proposer is not sure that the proposal will be accepted or rejected since it does not know what is best for the responder. Considering (16), the proposer can estimate $D_{\max}^m$ as follows:

$$X_{\max}^m(C_k) = \min_{j \in C_k} \frac{D^j Q(C_k \cup \{m\}, \alpha^{C_k}, B^m) - Q_j^m \sum_{i \in C_k \cup \{m\}} D^i}{Q_j^m} \quad (17)$$

The requested demand by the proposer is restricted to the interval $[0, D_{\max}^m(C_k)]$. To simplify the search for a proper demand, we define a unit Δ, making the proposer to propose a demand as integral multiplies of Δ. Therefore we can define the possible demand vector as $[0, \Delta, 2\Delta, ..., \lfloor D_{\max}^m(C_k)/\Delta \rfloor \Delta, D_{\max}^m(C_k)]$.

5. Bayesian Reinforcement Learning Coalition Formation

Types of players in a coalition dynamically change since microgrids' generation and demand vary in time. As a consequence, the coalition values change, which imposes uncertainty on the system. Combining Bayesian learning (RL) with the Coalition formation game gives the players the chance to learn about other players and eliminate uncertainties about them through interactions in the Bayesian Coalition formation process. In this section, first, we explain the conventional Bayesian reinforcement learning (RL) framework for a single agent, then we present the cooperative multi-user Bayesian learning framework suitable for the coalition formation process in microgrids, which is called as Bayesian Learning-based Coalition formation.

5.1. Conventional Bayesian RL

In the following, we briefly explain single-agent Bayesian reinforcement learning. We first need to define the Markov decision process (MDP) as an essential part of the reinforcement learning [30]. An MDP consist of four elements (S, A, Pr, r), where S is a vector of all possible states s. A is a set of all possible actions. Pr is a vector of all transition probabilities, and $Pr\{s'|s, a\}$ shows the chance of transition from state s to state s' while taking action a. $r(s, a)$ expresses the reward that the agent receives by taking action a in the state s. The RL problem can be defined as the problem of finding the optimal mapping

strategy from actions to states $\sigma : S \rightarrow A$ for the MDP with the known or unknown transition probabilities. In the Bayesian RL algorithm, first, a prior distribution is assigned to the initial beliefs of the agent about the values of the unknowns in the system. This belief will be updated continuously as the agent observes the unknown parameters. Considering the partially observable nature of MDP in Bayesian reinforcement learning, in this work, we employ the partially observable MDP (POMDP) technique in our model [30].

A POMDP consist of the following elements $(S_p, A_p, O_p, Pr_p, z_p, r_p)$, in which S_p denotes the set of states consisting of $S_p = S \times \{T_a^{s,s'}\}$, where $T_a^{s,s'}$ shows the unknown transition dynamics, $A_p = A$ represents set of actions, and $O_p = S$ shows the observation space similar to the state space in the general MDP. $Pr_p(s', T'|s, T, a)$, $z_p(s', T', a, o)$ and $r(s, T, a)$ represents state the transition probabilities, observation space and reward function, respectively.

In POMDP, the strategy is to map from beliefs to actions as $\sigma : B \rightarrow A$. We can calculate the value of a specific policy σ as the expected sum of discounted reward over infinite time in the future given by:

$$V^{\sigma}(B) = \sum_{t=0}^{+\infty} \gamma^t r(s_t, \sigma(B_t)) \tag{18}$$

where γ, s_t and B_t expresses the discount factor, state, and belief at time t. We are interested in finding the optimal policy σ^*. The optimal policy has the highest value for all the belief states, i.e., $V^{\sigma *}(B) > V^{\sigma}(B)$ and the corresponding value function of the optimal policy satisfies the Bellman equation as follows:

$$\begin{aligned} V^* &= \max_{a \in A} Q(s, B, a) \\ &= \max_{a \in A} \sum_{o \in O_p} \Pr\{o|s, B, a\}[u(s, B, a) + \gamma V^*(B_a^o)] \\ &= \max_{a \in A} \sum_{s' \in S_p} \Pr\{s'|s, B, a\}[u(s, B, a) + \gamma V^*(B_a^{s,s'})]. \end{aligned} \tag{19}$$

$Q(s, B, a)$ represents the action value in the case of taking action a in state s. B is the current belief about unknown parameter and $B_a^o = B_a^{s,s'}$ is the updated belief which is a probability density function (PDF). The PDF can be update using the Bayesian Theorem and observed transition (s, a, S') as follows:

$$B_a^{s,s'}(T) = \psi \Pr\{s'|s, B, a\} B(T) = \psi T_a^{s,s'} B(T) \tag{20}$$

where ψ is a normalizing constant.

5.2. Bayesian Reinforcement Learning Coalition Formation

In the following, we extend the previously discussed conventional single-agent BRL to the case of multiple agents in a coalition formation game. Our goal is to find the optimal coalition formation in the Bayesian coalition formation game that can be modeled as a POMDP.

Let us assume that the initial belief of the microgrid m is denoted by $B^m = B^m(T^{C_k})$ where T^{C_k} shows the types of players in the coalition C_k, similar to the unknown in the conventional BRL. Each microgrid m in the coalition C_k with the coalition action α^{C_k} can compute a long-term expected action value according to its beliefs B^m at each time slot t as follows:

$$\begin{aligned} &Q_t^m(C_k, \alpha^{C_k}, B^m) = \\ &\sum_{s^m \in S^m} \Pr\{s^m|C_k, \alpha^{C_k}, B^m\}(u^m(t) + \gamma V^m(C_k, B^m(T^{C_k}))) \\ &\qquad = \sum_{T^{C_k} \in \vec{T}^{C_k}} B^m(T^{C_k}) Q_t^m(C_k, \alpha^{C_k}, B^m|T^{C_k}) \end{aligned} \tag{21}$$

and

$$Q_t^m(C_k, \alpha^{C_k}, B^m | T^{C_k}) = \sum_{s^m \in S^m} \Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\}(u^m(t) + \gamma V^m(C_k, B^m_{s'^m}(T^{C_k}))) \quad (22)$$

where $u^m(t) = u^m(s', T^{C_k})$ expresses the reward that microgrid m receives at time t in the coalition C_k with the members of type T^{C_k} in the current state s'. The probability of transition from the current state s' to next state s^m by taking coalitional action α^{C_k} by members of type T^{C_k} is denoted by $\Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\} = \Pr\{s^m | s', C_k, \alpha^{C_k}, T^{C_k}\}$. $B^m_{s^m}(T^{C_k})$ expresses the updated belief after transition to the next state s^m about the types of other coalition members, T_{C_k}, which can be estimated using the Bayesian theorem as follows (same as single-agent belief update):

$$B^m_{s^m}(T^{C_k}) = \psi \Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\} B(T^{C_k}) \quad (23)$$

Consequently, we can find the optimal value-function V^m with a modified Bellman equation as follows:

$$V^m(C_k, T^{C_k}) = \sum_{C_k | m \in C_k, \vec{D}^{C_k}} \Pr\{C_k, \alpha^{C_k}, \vec{D}^{C_k} | B^m\} \times Q^m_{t-1}(C_k, \alpha^{C_k}, B^m) \quad (24)$$

Unlike the original form of the Bellman equation, in our problem, microgrid m cannot find the optimal V^m by maximizing Q^m_t as the coalitional process does not have full control of the coalition formation process. Therefore, microgrid m should estimate the probability $\Pr\{C_k, \alpha^{C_k}, \vec{D}^{C_k} | B^m\}$ instead tp find a specific coalition agreement $(C_k, \vec{D}^{C_k})$ that all coalition members will accept. Therefore, by considering (21) and (22) and the belief update (23), each microgrid can learn the long-term value of any agreement (CS, D) to find the optimal decision with respect to its beliefs about the types of other microgrids.

5.3. Computational Approximations

As has been mentioned in the previous part, it is not straightforward to estimate (24), since, on the one hand, we need to approximate the transition $\Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\}$ and the acceptance $\Pr\{C_k, \alpha^{C_k}, \vec{D}^{C_k} | B^m\}$ probabilities. On the other hand, by considering the size of type and state space, it is not possible to directly compute (21), (22), and (24). Therefore, a realistic simplification is needed to approximate (24). To this end, we employ the Bayesian exploration bonus to estimate the transition probability $\Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\}$[31]. In this method, we deploy counter parameters to determine how many times each transition occurs at each iteration t. The exploration bonus is used in order to put more weight on the paths that are not visited enough. Let us define the total number of transitions as:

$$\mu^m_0(C_k, \alpha^{C_k}, T^{C_k}) = \sum_{s^m \in S^m} \mu^m(s^m, C_k, \alpha^{C_k}, T^{C_k}) \quad (25)$$

here, $\mu^m(s^m, C_k, \alpha^{C_k}, T^{C_k})$ is a counter that shows how many time transition to s^m is accrued. Then, we can calculate $\Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\}$ as:

$$\Pr\{s^m | C_k, \alpha^{C_k}, T^{C_k}\} = \frac{\sum_{s^m \in S^m} \mu^m(s^m, C_k, \alpha^{C_k}, T^{C_k})}{\mu^m_0(C_k, \alpha^{C_k}, T^{C_k})} \quad (26)$$

Therefore, we can estimate the action value in (21) as follows:

$$Q_t^m(C_k, \alpha^{C_k}, B^m) = \sum_{T^{C_k} \in T^{C_k}} B^m(T^{C_k}) \sum_{s^m \in S^m} \tilde{\Pr}\{s^m | C_k, \alpha^{C_k}, T^{C_k}\} \times (u^m(t) + BEB + \gamma V^m(C_k, \tilde{B}^m_{s^m}(T^{C_k}))) \quad (27)$$

where BEB is given by:

$$BEB = \frac{\xi}{1 + \mu^m_0(C_k, \alpha^{C_k}, T^{C_k})} \quad (28)$$

ξ is a tuning parameter to adjust the chance of exploring less-visited transitions in transition probability. $\xi = 0$ means we skip the effect of BEB in our calculations. To estimate the acceptance probability $\Pr\{C_k, \alpha^{C_k}, \vec{D}^{C_k} | B^m\}$, we need $\lambda_0^m(C_k, \vec{D}^{C_k})$, which defines the times that agreement $(C_k, \vec{D}^{C_k})$ has been proposed, and $\lambda^m(C_k, \vec{D}^{C_k})$ shows how many times this agreement has been accepted. Therefore, we can estimate the $\Pr\{C_k, \alpha^{C_k}, \vec{D}^{C_k} | B^m\}$ as follows:

$$\tilde{\Pr}\{C_k, \alpha^{C_k}, \vec{D}^{C_k} | B^m\} = 0.5 + \zeta \frac{\lambda^m(C_k, \vec{D}^{C_k})}{\lambda_0^m(C_k, \vec{D}^{C_k})} \tag{29}$$

It should be noted that 0.5 is the initial value that is set for the acceptance probability. Additionally, we assume that $0 < \zeta < 1$.

The BRLC algorithm as applied to our microgrid coalition formation problem is given in Algorithm 1. The algorithm is divided into an initialization step and the main loop. Each microgrid's initial power level, location, and coalition are assigned randomly in the initialization step. Then the initial demand of each microgrid is derived concerning their direct power loss in the case that they only perform energy transactions with the macrogrid. After all initial steps, the (C_k, D_m, T^m) tuple will be transferred to all microgrids.

The main loop consist of two phases: the learning phase (lines 5–7) and the coalition formation phase. In the learning phase, the action values of all coalitions, the current reward of each microgrid, transition probabilities, and the action-value function of each microgrid will be updated. Then, in the coalition formation phase (lines 7–15), we assume that each time the power level of one random microgrid is changing and that specific microgrid is given a chance to propose. The proposer microgrid makes a proposal $\pi_k^i = (C_k, \vec{D}^{C_k})$ in which it decides about the coalition to join (or stay in the same coalition) and proposes a new demand in a way that maximize its own belief about Q_t^i. The proposal will then be transmitted to the member of the target coalition. Suppose all the members in the coalition find that their action-value function will be higher considering the new proposal. In this case, the proposal will be accepted, and the proposer microgrid will join/stay in the targeted coalition with the new demand. Otherwise, the proposal will be rejected, and the proposer microgrid stays in its previous coalition with the previous demand. After forming the new coalition structure, each coalition uses a greedy algorithm, introduced in [11], to exchange energy among the members of the coalition. If the coalition has a surplus or shortage of energy, then the coalition will transfer the surplus to or import this shortage from the macrogrid.

6. Performance Evaluation

In this section, at first, we briefly introduce our benchmark models and then examine the performance of the proposed model with respect to our benchmarks.

6.1. Benchmarks

6.1.1. Maximum a Posterior Estimation (MAPE)

In this model, the estimation of the action-value function is simplified to the most probable belief type, believed by agent m based on its current belief vector B^m as follows:

$$\tilde{T}_m^{C_k} = \underset{T^j \in \vec{T}^j}{\arg\max}\{B^m(T^j)\} \ \ , \forall j \in C_k \tag{30}$$

The main advantage of MAPE with respect to BCRL is its lower complexity due to ignoring the expected coalition value microgrids. To this end, in MAPE, microgrids reduce their action-value function as follows:

$$\tilde{Q}_t^m(C_k, \alpha^{C_k} | \tilde{T}_m^{C_k}) = \sum_{s^m \in S^m} \tilde{\Pr}\{s^m | C_k, \alpha^{C_k}, \tilde{T}_m^{C_k}\}(u^m(t) + BEB) \tag{31}$$

Since this method is a relaxed estimation of the proposed BCRL, we call it BCRLMAPE in the rest of the paper.

Algorithm 1: Coalition formation with BCRL for distributed energy trading among microgrids

Initialization:
for all microgrid m, $i \in M$ **do**
end
Randomly assigns the power level.
Randomly assign the location
Randomly assign to the coalition C_k
initializes demand D^m using direct power loss to macrogrid
Broadcast (C_k, D^m, T^m) to all microgrids and set ${Q^m}_t = 0$
Main Loop: time slot $t = 1 : iterations$ all microgrid m, $i \in M$
Update coalition action $\alpha^{C_k} \rightarrow f(C_k, \vec{D}^{C_k})$ according to the agreement (CS, X).
Update current reward $u^m(t)$.
Update transition probabilities and beliefs.
Estimate (27)
BR Coalition Formation:
Randomly selects a proposer microgrid m with the probability $1/M$.
Make a proposal $\pi_k^i = (C_k, \vec{D}^{C_k})$ which maximize microgrid m beliefs about Q_t^i.
Send $\pi_k^i = (C_k, \vec{D}^{C_k})$ to all microgrid n, $j \in C_k$.
for all microgrid j, $j \in C_k$ **do**
end
if $Q_t^i(C_k, \alpha^{C_k}, \vec{D}^{C_k} \circ D^i) \geq Q_t^i(C_k, \alpha^{C_k}, \vec{D}^{C_k})$ **then**
end
    set a response $\Omega_k^j = 1$ and send (Ω_k^j, D_m, T_m) to microgrid m
else
end
    set a response $\Omega_k^j = 0$
if $\prod_{m \in C_k} \Omega_k^j = 1$ **then**
end
    Update agreement $(C_k, \vec{D}^{C_k}) \rightarrow (C_k, \vec{D}^{C_k} \circ D^i)$
    set the state $s^i \rightarrow C_k$
    set the type T^i
    broadcast T^i to all microgrid j, $j \in C_k$

6.1.2. Fully Myopic Estimation (FME)

Similar to BCRLMAPE, the FME model has a lower complexity since in this model, only the instantaneous action-value function is considered, and the experience history is discarded. The action-value function is given by:

$$\tilde{Q}_t^m(C_k, \alpha^{C_k}, B^m) = \sum_{T^{C_k} \in \vec{T}^{C_k}} B^m(T^{C_k}) \sum_{s^m \in S^m} \tilde{\Pr}\{s^m | C_k, \alpha^{C_k}, T^{C_k}\}(u^m(t) + BEB) \tag{32}$$

The FME model, same as MAPE, is the reduced version of the BCRL; therefore, in this paper, we call this model BCRLFME.

6.1.3. Q-Learning Based Method

We compare our work with the Q-learning-based algorithm developed in [11]. Q-learning aims to reach a sub-optimal policy by choosing actions that maximize the expected current and future rewards. We assume that microgrids are agents and an agent's action is to refuse or accept the proposition of another microgrid to join their coalition. The state

is the vector of coalition memberships, and the reward function is the same as (9). The ϵ-greedy method is employed to consider action exploration.

6.1.4. Bayesian Coalitional Game Theory (BCG)

We implemented a Bayesian coalitional game theory-based approach for coalition formation in [11]. In this scheme, each microgrid makes a belief system about the types of other agents; however, agents do not learn from past experiences.

6.1.5. Coalitional Game Theory (CG)

A game theory-based coalition formation approach has been proposed in [6]. In this scheme, by employing a random merge and split technique, the system reaches a stable coalition formation which is not necessarily optimal or sub-optimal.

We refer the readers to [6,11] for more information on BCG and CG benchmarks.

Note that the proposed method, BCRLMAPE, BCRLFME, and Q-learning benchmarks use the ϵ-greedy policy to increase the chance of exploration. The ϵ-greedy policy helps with the trade-off between exploration and exploitation. Agents attempt to improve their long-term benefits through exploration, while exploitation can be achieved by performing greedy actions. Algorithm 1 is also used for BCRLMAPE, BCRLFME, and Q-learning benchmarks.

6.2. Numerical Results and Discussions

In this section, for numerical evaluation, we consider a network of 4 to 10 microgrids within an area of 20 km by 20 km where microgrids and macrogrid interconnections are located randomly. We divided the full day into 240 time slots, where the load and generation patterns are randomly generated, and this procedure was periodically repeated every day with slight variations as in [6].

We compare the proposed BCRL with BCRLMAPE, BCRLFME, Q-learning, BCG, and CG benchmarks. The results are averaged over ten runs. The simulation parameters are presented in Table 2.

Table 2. Summary of simulation parameters

Parameters	Value
Line Resistance (R_{mn})	0.2
Medium Voltage (U_0)	50 kV
Low voltage (U_i)	22 kV
Transformer loss fraction (ρ)	0.02
Threshold distance (D_{tr})	5 km
Virtual cost parameter (w_s)	0.02
Virtual cost parameter (w_l)	0.04
Virtual cost parameter (w_0)	0.08
Scaling parameter (δ)	0.95

In Figure 2, we present the average cost per user versus the number of microgrids ranging from 4 to 10. As expected, increasing the number of microgrids will reduce the cost since microgrids have more chance to make local coalitions in a dense network, resulting in less power transmission with the macrogrid and by microgrids, resulting in lower costs. Moreover, since BCRL is designed to overcome the uncertainty, it demonstrates more cost results than the other algorithms. The proposed algorithm shows 4% to 16% improvement compared to BCG and the sub-optimal BCRLMAPE, respectively.

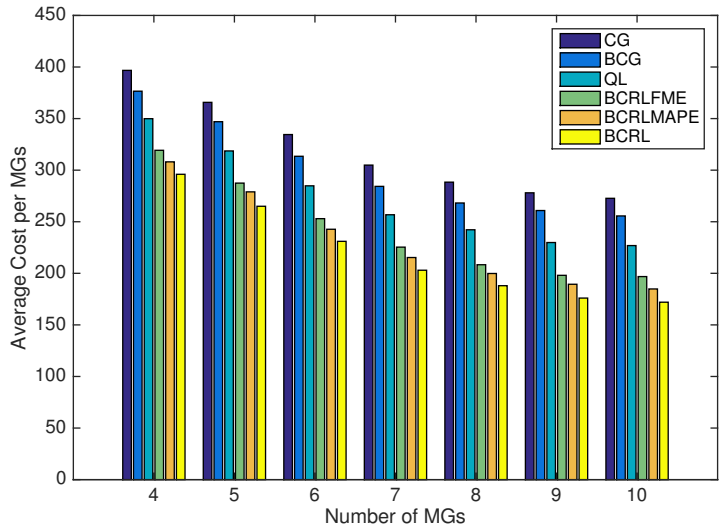

Figure 2. Average cost versus number of microgrids.

In Figure 3, to evaluate the effect of increasing power levels, we demonstrate the average cost per user versus the power levels. It should be noted that, in this figure, BCG and CG models cannot be compared with other models, since the power levels are not considered in these models. As is shown, when the power levels increase, the average cost decreases as expected. As we increase the power level, the quantization error will be reduced, and as a result, all the approaches perform better. As we can see in Figure 3, at different power levels, BCRL reduces the average cost per microgrid to 5% and 15% compared to the BCRLMAPE and BCG methods, respectively.

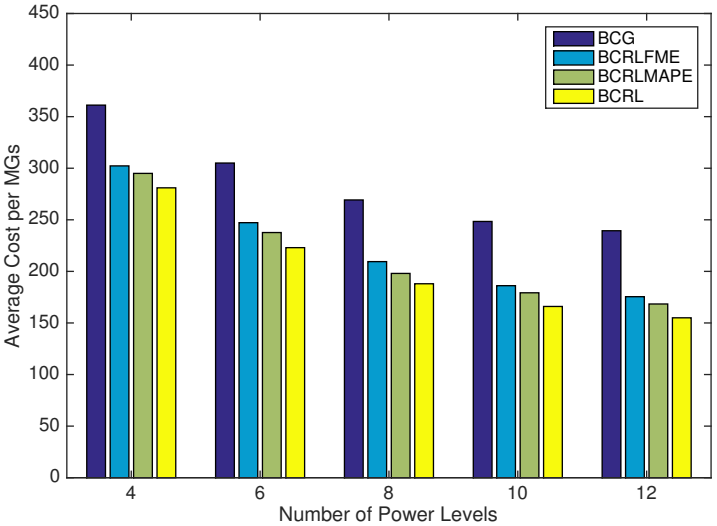

Figure 3. Average cost per microgrid versus the power levels.

In Figure 4, we present the average power loss per user versus the number of microgrids. As the number of microgrids increases, the distance between microgrids will be reduced, reducing the power loss in the system. Moreover, since BCRL is designed to overcome the uncertainty, it demonstrates better power loss results than benchmark approaches, with an up to 50% improvement with respect to conventional CG. While Q-learning benefits from past experience to make the best decision, BCG relies on the beliefs about the types of other players. The CG method only performs based on the random join and split iterations in coalitions to reach a stable coalition formation, which is not necessarily optimal.

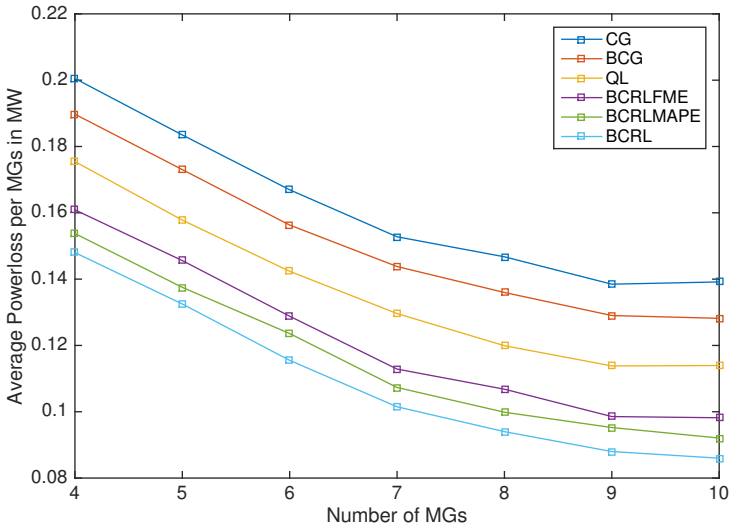

Figure 4. Average power loss versus number of microgrids.

Figure 5 shows the average power loss per user versus the number of power levels. As expected, by increasing the number of power levels, the power loss will be reduced due to the lower quantization error. As can be seen, the BCRL method is less prone to quantization errors due to its comprehensive estimation model for the expected action value. The BCRL method gained up to 20% on average compared to the BCG method.

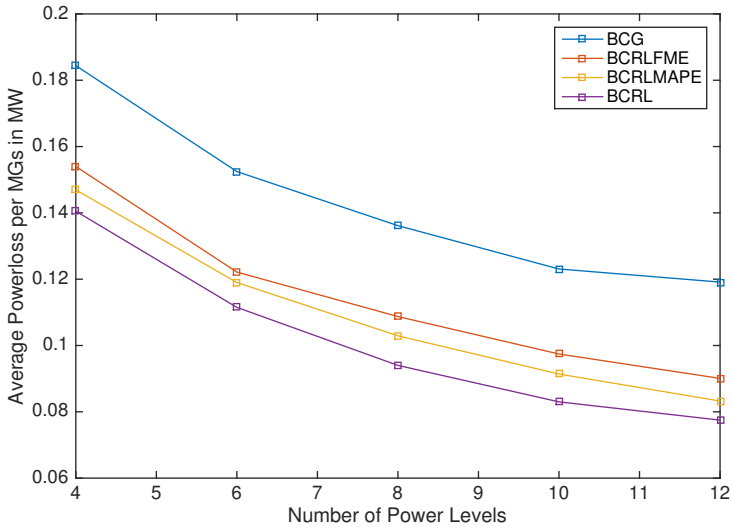

Figure 5. Average power loss per microgrid versus the number of power levels.

In Figure 6, the average amount of energy transferred to the macrogrid versus the number of microgrids is presented. As we can see, BCRL requires a lower amount of energy exportation to or importation from the macrogrid compared to the benchmark techniques. Additionally, due to the lower power loss between nearby microgrids, the probability of joining nearby microgrids to the same coalition will increase by increasing the number of microgrids, which can reduce the power exported to the macrogrid, as well.

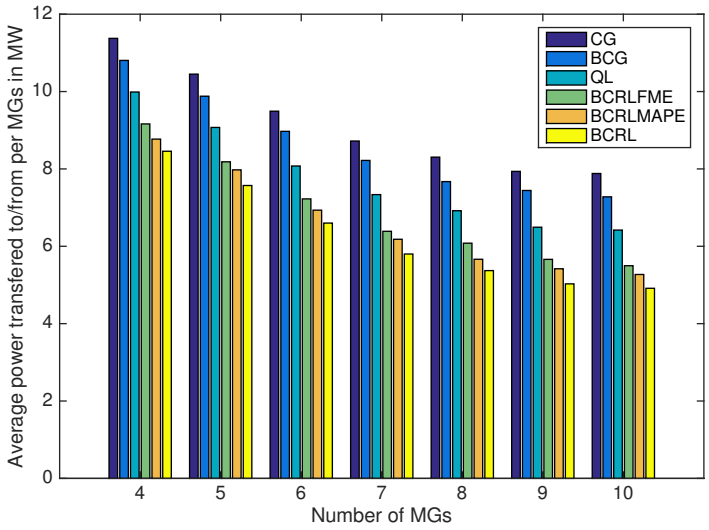

Figure 6. Average energy transfer with macrogrid versus number of microgrids.

Figure 7 shows the impact of increasing the cost of transferring energy with the macrogrid versus the average energy transferred with macrogrid. Here, the range of weighting parameter w_0 varied between 0.02 to 0.22. As we can see, when w_0 increases, the average energy transfer with macrogrid decreases, giving a chance to the coalition of microgrids to operate in islanding mode. We can see that the proposed BCRL model always performs better in making independent coalitions that rely less on macrogrid. BCRL decreases the exported power to the macrogrid up to 10% in comparison with the CG technique.

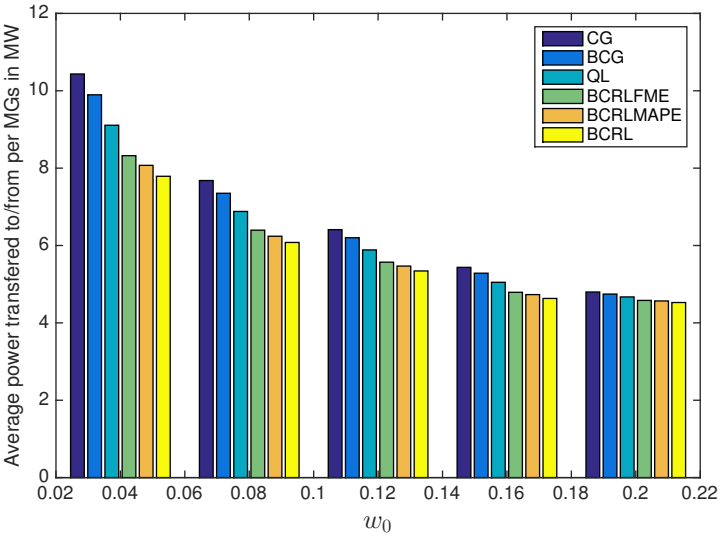

Figure 7. The average energy transfer with macrogrid versus virtual cost weighting parameter w_0.

Figure 8 shows the convergence of the BCRL technique in terms of the average power loss per user. As is shown, the proposed model will be converged after 12,000 iterations.

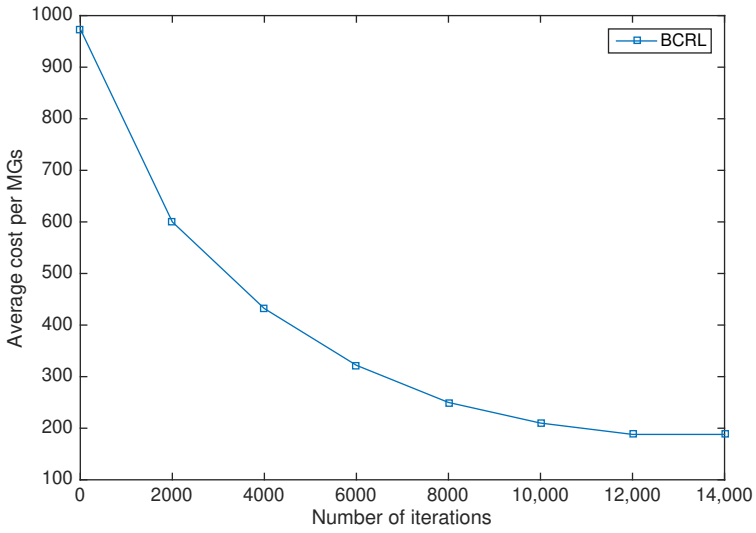

Figure 8. Convergence of the average cost per user versus the number of iteration for BCRL.

In Figure 9, we demonstrated the average number of iterations that are needed for the convergence of accumulative average power loss as the number of power levels increases in the BCRL scheme.

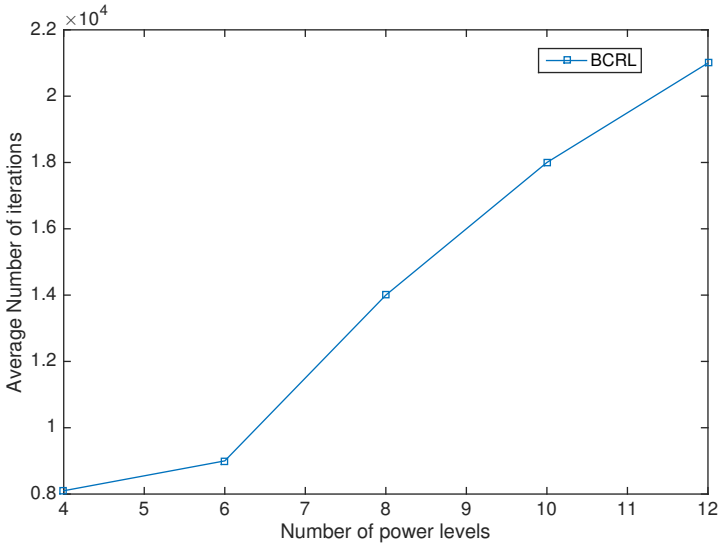

Figure 9. Number of iterations to convergence versus the power levels.

7. Conclusions

In this paper, we propose a Bayesian coalitional reinforcement learning-based approach for learning the optimal policy to minimize the cost for distributed energy trading among microgrids. In this work, each microgrid is modeled as an agent that can compete and cooperate with other agents. We model this problem as a Markov game, which aims to maximize the reward for each agent to overcome the uncertainties that are caused by microgrids based on their generation and demand. With the proposed scheme, microgrids reach stable coalitions where the energy export to the macrogrid or distant microgrids are reduced in the system. We introduced an algorithm that helps each agent systematically propose joining a new coalition and give the coalition members the chance to accept or reject the proposal according to their expected long-term rewards. To evaluate the performance of the proposed model, we compared our results with five benchmark schemes and showed that our scheme reduced the cost and power loss more than the others, reaching to 23% reduction in cost and a 28% reduction in power loss.

Author Contributions: M.S. is a student author who implemented the proposed scheme. S.M. helped with finalizing the manuscript. M.E.-K. is the supervisor of both students and contributed to the concepts, methodology, and manuscript preparation. Conceptualization, methodology, and writing—original draft, M.S., S.M. and M.E.-K.; software, M.S.; supervision, M.E.-K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSERC Discovery program.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ICT	Information and communications technologies
p2p	peer to peer
microgrid	Micro grid
BCFG	Bayesian coalition formation game
BCRL	Bayesian coalition reinforcement learning
CG	Coalitional game theory
EV	Electrical Vehicle
AI	Artificial intelligence
SBC	Strong Bayesian Core
MDP	Markov decision process
RL	Reinforcement learning
POMDP	partially observable MDP
M	Number of microgrids
g_m	The amount of generated energy by microgrid $m \in M$
d_m	microgrid m demand
q_m	Surplus energy of microgrid m
S_{mn}	Total cost of power transaction from m-th microgrid to n-th microgrid
d_{mn}	Length of power lines between m-th and n-th microgrids
δ	Scaling factor
E_{mn}	The power that is being traded plus the loss that happens during trading
$PL(E_{mn})$	Power loss in trading energy between m-th and n-th microgrids
w	Weighting coefficient
w_s	Lower bound for w
d_{tr}	Distance threshold
w_l	Higher bound for w
w_0	weight factor for energy transactions with the macrogrid
R_{mn}	Resistance of line
PL	Power loss
U_m	Voltage in transformer
ρ	Fraction of power loss in transformer
C	Coalition
v_C	Coalitional value
$\|C\|$	Number of members of coalition C
$v_{max}(C)$	Maximum achievable coalition value
$\vec{T}$	Set of agent types
B^m	Set of agent beliefs
$\vec{A}^{C_k}$	Set of coalition action
$\vec{S}$	Set of states
u^m	Reward function
$B^m(T^{-m})$	m-th microgrid's beliefs about the types of other players
$\vec{T}^{-m}$	Probability assigned to other agents about their types
$\vec{A}^{C_k}$	Set of coalition actions
$v(j)$	Value of Member j in the coalition
v_C^{max}	Coalition C value

D^m	Demand
$\vec{D}^{C_k}$	Demand vector of coalition C_k
π_m^k	Proposal by prosper m
γ	discount factor
s_t	State at time t
B_t	Belief at time t

References

1. Amin, W.; Huang, Q.; Umer, K.; Zhang, Z.; Afzal, M.; Khan, A.A.; Ahmed, S.A. A motivational game-theoretic approach for peer-to-peer energy trading in islanded and grid-connected microgrid. *Int. J. Electr. Power Energy Syst.* **2020**, *123*, 106307. [CrossRef]
2. Brolin, M.; Pihl, H. Design of a local energy market with multiple energy carriers. *Int. J. Electr. Power Energy Syst.* **2020**, *118*, 105739. [CrossRef]
3. Zhang, W.; Maleki, A.; Rosen, M.A. A heuristic-based approach for optimizing a small independent solar and wind hybrid power scheme incorporating load forecasting. *J. Clean. Prod.* **2019**, *241*, 117920. [CrossRef]
4. Abrishambaf, O.; Lezama, F.; Faria, P.; Vale, Z. Towards transactive energy systems: An analysis on current trends. *Energy Strategy Rev.* **2019**, *26*, 100418. [CrossRef]
5. Erol-Kantarci, M.; Kantarci, B.; Mouftah, H.T. Reliable overlay topology design for the smart microgrid network. *IEEE Netw.* **2011**, *25*, 38–43. [CrossRef]
6. Saad, W.; Han, Z.; Poor, H.V. Coalitional Game Theory for Cooperative Micro-Grid Distribution Networks. In Proceedings of the IEEE International Conference on Communications Workshops, Kyoto, Japan, 5–9 June 2011.
7. Zhou, H.; Erol-Kantarci, M. Correlated Deep Q-learning based Microgrid Energy Management. In Proceedings of the 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Pisa, Italy, 14–16 September 2020; pp. 1–6. [CrossRef]
8. Xiao, X.; Dai, C.; Li, Y.; Zhou, C.; Xiao, L. Energy Trading Game for Microgrids Using Reinforcement Learning. In *Game Theory for Networks*; Springer International Publishing: Cham, Switzerland, 2017; pp. 131–140.
9. Chalkiadakis, G.; Boutilier, C. Bayesian reinforcement learning for coalition formation under uncertainty. In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, New York, NY, USA, 19–23 July 2004; Volume 3, pp. 1090–1097.
10. Asheralieva, A. Bayesian Reinforcement Learning-Based Coalition Formation for Distributed Resource Sharing by Device-to-Device Users in Heterogeneous Cellular Networks. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 5016–5032. [CrossRef]
11. Sadeghi, M.; Mollahasani, S.; Erol-Kantarci, M. Power Loss-Aware Transactive Microgrid Coalitions under Uncertainty. *Energies* **2020**, *13*, 5782. [CrossRef]
12. Lee, J.; Guo, J.; Choi, J.K.; Zukerman, M. Distributed Energy Trading in Microgrids: A Game-Theoretic Model and Its Equilibrium Analysis. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3524–3533. [CrossRef]
13. Jadhav, A.; Patne, N. Priority Based Energy Scheduling in a Smart Distributed Network with Multiple Microgrids. *IEEE Trans. Ind. Inform.* **2017**, *13*, 1. [CrossRef]
14. Tushar, W.; Zhang, J.A.; Smith, D.; Poor, V.H.; Thiebaux, S. Prioritizing Consumers in Smart Grid: A Game Theoretic Approach. *IEEE Trans. Smart Grid* **2014**, *5*, 1429–1438. [CrossRef]
15. Wang, H.; Huang, T.; Liao, X.; Abu-Rub, H.; Chen, G. Reinforcement Learning in Energy Trading Game Among Smart Microgrids. *IEEE Trans. Ind. Electron.* **2016**, *63*, 5109–5119. [CrossRef]
16. Feng, C.; Wen, F.; You, S.; Li, Z.; Shahnia, F.; Shahidehpour, M. Coalitional Game Based Transactive Energy Management in Local Energy Communities. *IEEE Trans. Power Syst.* **2019**, *35*, 1729–1740. [CrossRef]
17. Lahon, R.; Gupta, C.P.; Fernandez, E. Coalition formation strategies for cooperative operation of multiple microgrids. *IET Gener. Transm. Distrib.* **2019**, *13*, 3661–3672. [CrossRef]
18. Mei, J.; Chen, C.; Wang, J.; Kirtley, J.L. Coalitional game theory based local power exchange algorithm for networked microgrids. *Appl. Energy* **2019**, *239*, 133–141. [CrossRef]
19. Essayeh, C.; El Fenni, M.R.; Dahmouni, H. Optimization of energy exchange in microgrid networks: A coalition formation approach. *Prot. Control. Mod. Power Syst.* **2019**, *4*, 24. [CrossRef]
20. Misra, S.; Krishna, P.V.; Saritha, V.; Obaidat, M.S. Learning automata as a utility for power management in smart grids. *IEEE Commun. Mag.* **2013**, *51*, 98–104. [CrossRef]
21. Jiang, B.; Fei, Y. Dynamic Residential Demand Response and Distributed Generation Management in Smart Microgrid with Hierarchical Agents. *Energy Procedia* **2011**, *12*, 76–90. [CrossRef]
22. Xu, Y.; Zhang, W.; Liu, W.; Ferrese, F. Multiagent-Based Reinforcement Learning for Optimal Reactive Power Dispatch. *IEEE Trans. Syst. Man Cybern. Part A (Appl. Rev.)* **2012**, *42*, 1742–1751. [CrossRef]
23. Guan, C.; Wang, Y.; Lin, X.; Nazarian, S.; Pedram, M. Reinforcement learning-based control of residential energy storage systems for electric bill minimization. In Proceedings of the 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2015; pp. 637–642.

24. Kim, B.G.; Zhang, Y.; van der Schaar, M.; Lee, J.W. Dynamic pricing for smart grid with reinforcement learning. In Proceedings of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 27 April–2 May 2014; pp. 640–645.
25. Xiao, L. Reinforcement Learning-Based Energy Trading for Microgrids. Available online: https://arxiv.org/abs/1801.06285 (accessed on 5 November 2021).
26. Sadeghi, M.; Erol-Kantarci, M. Power Loss Minimization in Microgrids Using Bayesian Reinforcement Learning with Coalition Formation. In Proceedings of the 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, 8–11 September 2019; pp. 1–6.
27. Sadeghi, M.; Mollahasani, S.; Erol-Kantarci, M. Cost-Aware Dynamic Bayesian Coalitional Game for Energy Trading among Microgrids. In Proceedings of the IEEE International Conference on Communications Workshops (ICC), Montreal, QC, Canada, 14–23 June 2021; pp. 1–6.
28. Machowski, J.; Lubosny, Z.; Bialek, J.W.; Bumby, J.R. *Power System Dynamics: Stability and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2020.
29. Ieong, S.; Shoham, Y. Bayesian Coalitional Games. In Proceedings of the 23rd National Conference Artificial Intelligence, Chicago, IL, USA, 13–17 July 2008; AAAI: Palo Alto, CA, USA, 2008; pp. 95–100.
30. Duff, M.O.; Barto, A. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*; University of Massachusetts Amherst: Amherst, MA, USA, 2002.
31. Kolter, J.Z.; Ng, A.Y. Near-Bayesian exploration in polynomial time. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 513–520.