The Electricity Journal 35 (2022) 107129

Contents lists available at ScienceDirect

The Electricity Journal

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

# A multi-agent deep reinforcement learning based energy management for behind-the-meter resources

Patrick Wilk [a], Ning Wang [b], Jie Li [a,*]

[a] Electrical and Computer Engineering Department, Rowan University, Glassboro, NJ, USA
[b] Computer Science Department, Rowan University, Glassboro, NJ, USA

ARTICLE INFO

*Keywords:*
Energy management
Multi-agent deep reinforcement learning
Behind-the-meter resources

ABSTRACT

The future communities are becoming more and more electrically connected via increased penetrations of behind-the-meter (BTM) resources, specifically, electric vehicles (EVs), smart buildings (SBs), and distributed renewables. The electricity infrastructure is thus seeing increased challenges in its reliable, secure, and economic operation and control with increased and hard to predict demands (due to EV charging and demand management of SBs), fluctuating generation from renewables, as well as their plug-N-play dynamics. Reinforcement learning has been extensively used to enable network entities to obtain optimal policies. The recent development of deep learning has enabled deep reinforcement learning (DRL) to drive optimal policies for sophisticated and capable agents, which can outperform conventional rule-based operation policies in applications such as games, natural language processing, and biology. Furthermore, DRL has shown promising results in many resource management tasks. Numerous studies have been conducted on the application of single-agent DRL to energy management. In this paper, a fully distributed energy management framework based on multi-agent deep reinforcement learning (MADRL) is proposed to optimize the BTM resource operations and improve essential service delivery to community residents.

## 1. Introduction

The future communities are becoming more and more electrically connected via increased penetrations of behind-the-meter (BTM) resources, specifically, electric vehicles (EVs), smart buildings (SBs), and distributed renewables. Transportation Electrification (TE), fueled by carbon-free electric energy, is regarded as one of the major contributors in reducing petroleum use, meeting air quality standards, improving public health, and achieving greenhouse gas emissions reduction goals (Air Quality,). Although promising, once implemented, simultaneously charging a group of EVs concentrated in a limited number of charging stations could exacerbate an undesirable peak demand, impacting the reliable operation of the electricity distribution system (Clement-Nyns et al., 2010). According to the Oak Ridge National Laboratory (Harley and Tsvetkova, 2008), if not planned and managed properly, most regions would need to invest in additional generation capacities to meet the new demand for EV charging. Buildings across university campuses, governments, schools, and residential, industrial, and commercial sectors represent key elements in a modern community ecosystem, as well as major consumers of the electric energy systems. More and more newly built and retrofitted buildings are deployed with smart building energy management systems (BEMSs) to strategically manage the operation of HVAC, Heating, Cooling, and Lighting facilities for enhanced building energy efficiency, while in response to the electric utilities' demand response (DR) requests for economic benefits (Kolokotsa and Kampelis, 2021). It is also well-recognized that a group of energy efficient SBs with flexible end-use equipment and onsite distributed generations can collectively work to maximize building and grid efficiency without compromising occupant comforts (Grid-interactive efficient buildings,), thus playing a critical role in mitigating environmental impacts and enhancing community energy efficiency and reliability. Furthermore, in pursuing a sustainable energy future, an increasing penetration of distributed renewables has been observed to reform the power flow of electricity distribution systems, transforming electric distribution line flows from unidirectional to multidirectional, along with higher probabilities in line congestions (Hadusha and Meeus, 2018), which is rarely observed in the systems of the past. The electricity infrastructure is thus seeing increased challenges in its reliable, secure, and economic operation with increased and hard to predict demands (due to EV charging and demand management of SBs), fluctuating generation resources from

* Corresponding author.
*E-mail address:* lijie@rowan.edu (J. Li).

https://doi.org/10.1016/j.tej.2022.107129
Received 28 March 2022; Accepted 12 April 2022
Available online 4 May 2022

renewables, as well as their plug-N-play dynamics. In this paper, a feasible, efficient, and scalable energy management system (EMS) solution is explored to facilitate the plug-N-play of a massive number of heterogeneous BTM resources, providing valuable decision-making tools to community facilities.

Reinforcement learning (RL) (Kohl and Stone, 2004; Tesauro, 2018) have been extensively used to enable network entities to obtain optimal policies (e.g., operation decisions or scheduling actions). The recent advancement of deep learning has enabled deep reinforcement learning (DRL) (Bengio et al., 2013; Arulkumaran et al., 2017) to drive optimal policies for sophisticated and capable agents, which can outperform conventional rule-based operation policies in applications such as games (e.g., Go, WarCraft) (Mnih et al., 2015; Bellemare et al., 2013), natural language processing (Bahdanau et al., 2017; Ranzato, 2016), and biology (e.g., AphaFold). Furthermore, DRL has shown promising results in many control and resource management problems, such as task scheduling, resource allocation, communication, and control (Frikha, 2021; Chen, 2021). We thus see great potential in leveraging DRL in optimizing the management of massive networked BTM resources to enhance the electricity infrastructure operations and guarantee the essential services provision.

The remaining of this paper is organized as follows: Section 2 summarizes the major challenges and possible technologies for building a feasible, efficient, and scalable EMS for BTM resources; Section 3 provides a detailed literature survey of RL based solutions for energy management; Section 4 proposes a multi-agent deep reinforcement learning (MADRL) based EMS, and detailed functional components are discussed; Section 5 summarizes this paper and presents some future works.

## 2. Challenges and potential technologies

The operation of BTM resources, residing in future community energy systems, constitute temporally and spatially coupled cyber and physical constraints, meaning a current system operation decision may affect its future decisions, and behaviors among different entities may impact each other (Yu et al., 2021). It becomes difficult to create an explicit mathematical model that is accurate and efficient enough to perform a real-time energy management towards operation optimization, due to the complexities and uncertainties associated with the energy systems of future communities. For example, SBs with controllable loads create complex operation modes, considering different entities with different consumptions, generations, and flexibilities profiles, while dispersedly deployed PV panels exhibit fluctuations due to many uncertainties such as temperature, shading, wind, etc., creating a high data dimension. The unpredictable nature of renewables, uncertainties of EV charging stations (EVCSs), time-varying loads, together with changing energy prices, introduce challenges for a feasible, efficient, and scalable EMS solution. A solution that can monitor, predict, schedule, learn, and make decisions in real-time is of essence. In addition, in a multi-device environment, a good operation strategy should cooperate multiple entities to maximize the operation efficiency and find the balancing point between their individual benefits and the system benefit.

### *2.1. Mathematical modeling-based methods*

Many studies exist using mathematical modeling-based methods, such as dynamic programming (DP), linear programming (LP), and their derivatives to perform the operation optimization of an energy management system. Mathematical models are rigorous and real-time management could be realized. However, they rely on explicit formulation of objective functions and system constraints, which are difficult to abstract from real-world environments (Zhang et al., 2019). Nevertheless, these approaches fall short due to their infamous curse of dimensionality and lack of ability to adapt to the stochasticity of the environment, and thus have limited scalability and versatility (Arwa and Folly, 2020).

Rule-based heuristic methods have been proposed for EMS as well, constructed by predefined policies, heuristics, or human expertise to estimate optimal solutions. Although highly reliable and robust, these methods lack the adaptability and flexibility to the frequent dynamics in energy systems (Abdullah et al., 2021).

Global search methods such as genetic algorithm and swarm intelligence have gained popularity and recognition in solving non-convex energy management problems with large problem scales. However, these methods are less robust, cannot be proofed rigorously, are generally slow, and without a learning component. Thus, they cannot operate online and instead must solve optimization iterations every time new data is introduced, proving computationally expensive (Zhang et al., 2019).

The energy system moving towards more economic and environment friendly is asking for more feasible, efficient, and scalable EMS solutions. Conventional methods introduced above are seeing bottlenecks when solving such complex control problems due to the increased complexity, uncertainty, and high dimension data acquisition.

### *2.2. Learning-based method*

Machine Learning (ML) algorithms learn from experience, by finding trends and patterns in the training dataset with a goal to make accurate predictions and decisions. ML algorithms can be categorized into supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, labeled data is used to evaluate and improve the learned model, thus it cannot learn by itself and extract unknown information from the training data. Unsupervised learning, with training data unlabeled, allows the extraction of hidden patterns from data without human involvement. Both learnings, unsupervised and supervised, require static datasets to train a model, resulting in a static trained model. No matter how complicated the relationship representing the dataset, the learned model is predefined for a particular dataset. For example, a predictive EVCS model trained on a specific dataset cannot be reused on another EVCS. Unlike supervised or unsupervised learning, RL is trained on a dynamic dataset upon trial and error and finds a dynamic policy proving more robust. In addition, supervised learning assumes events are independent of previous events which is not an appropriate assumption in energy management problems that involves temporally/spatially dependent events. For instance, the optimal charging for an EV now depends on its strategies of previous/future times, as well as other EVs. In contrast, RL algorithms use a sequential design, and thus are more appropriate for solving complex scheduling problems and are constantly learning and adapting to the changing environments.

RL has proved effective in non-stationary environments that change over time (Kim et al., 2018; Chiş et al., 2017; Yoldas et al., 2020; Remani et al., 2019). A RL problem is usually formulated as a Markov Decision Process (MDP), a sequential decision task (Abdullah et al., 2021). A MDP consists of well-defined state-space, possible actions within each specific state, a state transition function or probability, and rewards (Arwa and Folly, 2020). In RL, the agent learns optimal policy by interacting with the environment, where a policy is a mapping from states to the probabilities of selecting every possible action (Yu et al., 2021). A software agent can be thought of a decision/action-maker that learns through repeated trial and error, defined by a reward scheme with a goal of maximizing total reward over an extended time. The reward is used to communicate how well the agent is learning. Therefore, reward shaping is important to achieve the desired performance. The advantages of RL algorithms can be summarized as: i) eliminate the need to iterate during online operation, since they can be trained offline where optimal solutions are retrieved for the whole optimization horizon (Arwa and Folly, 2020). After training and during execution, computational complexity is very low. If a high-dimensional state is given, the optimal action can be determined quickly (i.e., 10 ms) (Yu et al., 2021). ii) Unlike mathematical modeling-based methods, RL does not require an accurate model of the environment to achieve the optimal solution and can learn optimal policies by trial-and-error interacting with the environment. iii) RL can complete complicated tasks with lower prior knowledge thanks to its ability to learn different levels of abstractions from data, can handle high dimensional data, can make real-time and online decisions, and is even more robust compared to heuristic methods (Zhang et al., 2019). iv) The application of a deep neural network (DNN) has the capability to make accurate predictions and so there is no need for a separate forecasting model like in global search methods (Arwa and Folly, 2020). Deep Neural Networks (DNN) are also used to approximate a value function, supporting high-dimension feature extraction and learning (Schulze et al., 2016). Deep Learning can be further evolved to include RL, allowing the estimation from DL and rewarded actions from RL. Compared to global search methods, DRL agents do not require forecasting or statistical information of the environment, but continuously learn and improve via online learning.

For all these reasons, RL is suitable for a dynamic environment sequential decision-making, especially by integrating with deep learning for automatic feature extraction from data and improved scalability. Energy management problems usually involve high-dimensional and continuous state or action spaces, that cannot be stored in a table or function. DNNs are function approximators, which are particularly useful in RL when the state space or action space are too large to be completely known or stored. It is expected that a DRL based EMS solution can achieve online optimization and real-time control for the energy management of massive number of heterogeneous BTM resources, improve efficiency of energy utilization, reduce operating costs, and increase overall community benefits.

There are two types of DRL methods, model-based or model-free (Fernandez et al., 2020) shown in Fig. 1. A model-based DRL does not necessarily mean a mathematical model must be provided, but instead agents learn a model based on observing how states in an environment change with certain actions, and then use the learned model. Model-based methods are more complex having more assumptions and approximations compared to model-free methods and therefore, may be limited to specific tasks. Model-based methods outperform model-free methods in the sense of sample complexity. However, model-based methods often see the challenge of obtaining an accurate model from the environment especially in the complex and uncertain energy systems. Model-free DRL, either value-based or policy-based, do not need to develop a model from the environment, but instead directly learn a policy or state-action value. Value-based methods learn an approximation of optimal policy function (indirectly) while policy-based methods learn an approximation of optimal policy (directly). Typically, value-based methods update value function in an “off-policy” manner, which means the previously collected experience transitions in the same environment can be used for training, and high data efficiency can be achieved. In contrast, “on-policy” makes all the updates using data from the trajectory distribution generated by the current policy. Thus, “on-policy” methods are more stable but less data-efficient compared with “off-policy” methods (Shin et al., 2020).

Deep Q-Network (DQN), Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), and Asynchronous Advantage Actor Critic (A3C) are typical model-free DRL algorithms. DQN is value-based off-policy algorithm and only supports discrete action space, while DDPG is policy-based off-policy algorithm and only supports continuous action space. Experience replay is adopted by DQN and DDPG, which makes them have higher data efficiency compared to on-policy algorithms. However, they tend to overestimate the value function and generate sub-optimal policies. PPO and A2C/A3C, two policy-based on-policy algorithms, can support both discrete and continuous actions spaces. PPO can support stable learning by controlling the similarity between the current and old policies and is robust to hyperparameters and network architectures. A2C/A3C can support reliable and parallel learning on a single multi-core CPU, but it is sensitive to employed hyperparameters (Yu et al., 2021).

DRL methods are also categorized according to the number of agents, i.e., single-agent and multi-agent. A common and straightforward way to implement multi-agent DRL is to extend single-agent approaches. The diversity of BTM sources, temporally and spatially coupled constraints along with various user priorities, constitutes the allocation of multiple agents for decision making in EMS solutions. Spatially coupled constraints can be guaranteed with a multi-agent DRL that has a proper reward function and observation space, allowing system coordination. Temporally coupled constraints can also be satisfied by designing efficient reward function to incite the agent to take reasonable actions. Furthermore, assigning multiple agents increases the speed of problem-solving, enables self-learning for each agent, and increase the solution

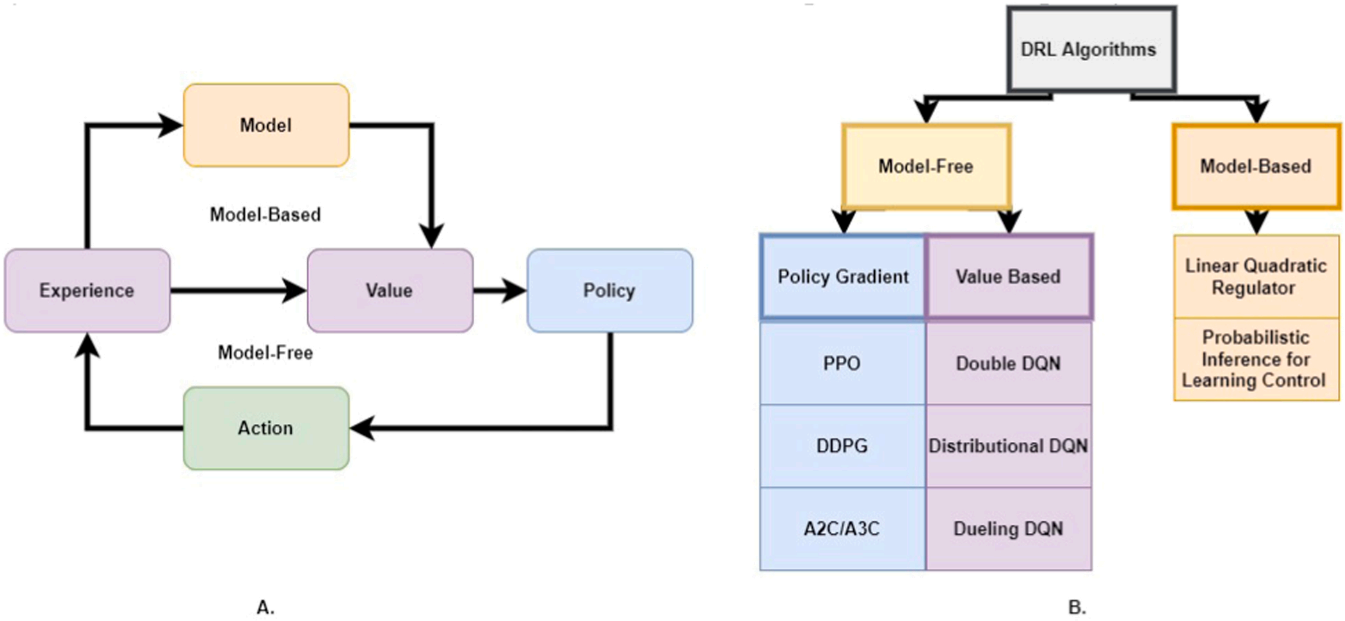

**Fig. 1.** DRL Algorithm Breakdown (A.) with Examples (B.).

reliability because of redundant agents (Ahrarinouri et al., 2021).

## 3. Literature review on RL in energy managing

In this section an extensive review of scientific literatures of RL based EMS solutions are presented and compared. A systematic review is conducted based on a methodology originally performed on medical research and first outlined for the field of organization studies by Tranfield et al. (2003). The aim of the review is to locate relevant studies, evaluate their contributions, and compare the conclusions with regards to further research specifically multi-agent RL frameworks in managing the BTM resources. A total of 62 journal articles were identified, and of which 21 investigated Multi-agent Reinforcement Learning (MARL) solutions. This review mainly focuses on EMS solutions on both the micro level (e.g., buildings/communities/sites/etc.) and the macro level (e.g., grids/regions/etc.). To combine the evidence of the articles reviewed, all the articles are analyzed to identify important concepts and themes within literature and categorized into following groups.

### *3.1. Q-learning*

Q-learning, a value-based method, is the most widely used off-policy RL algorithm due to its simplicity (Arwa and Folly, 2020). Q-learning does not require a model of the environment to define relationship between an environment and problem features and to establish state transitional probabilities (Ahrarinouri et al., 2021). Q-learning use a Q-table to keep track of the learning process. However, as state-action pairs increase, the Q-table also increase, thus, Q-learning suffers from the curse of dimensionality just like DP, which prevents the use of Q-learning in complex problems with large state-action spaces. In addition, Q-function is deterministic and cannot handle stochastic policies (Arwa and Folly, 2020). Standard Q-learning applies only to discrete action spaces, which limits the algorithm from achieving optimal actions by forcing the agent to select actions from a pre-defined action set. However, discretization of the state and action space can be solved by using a function approximation called W-learning (Abdullah et al., 2021). Q-learning is effective when state space is low-dimensional, but a nonlinear approximator such as a NN can be used to represent the action-value function more efficiently to support high-dimensional state space (Yu et al., 2021).

In (Foruzan et al., 2018) an adaptive EMS framework is proposed considering the variability and stochastic entities in a microgrid by applying a distributed multi-agent Q-learning algorithm, offering a scalable solution without the need of excessive communication to a central controller or other agents. A total of five as self-interested agents were modeled that can learn to adapt to each other without prior knowledge of a stochastic environment. A random model was included to represent the uncertainties associated with each agent, which used Q-learning to converge to a Nash equilibrium. The author evaluated and proved the cooperation among agents with a “fairness factor” that compared the profit of generation agents to the cost of customer agents. In (Ahrarinouri et al., 2021) a Q-learning approach is explored to achieve the optimal solution for a residential EMS. Both deterministic and stochastic environments were tested to justify the effectiveness and robustness of the method. Similar to (Foruzan et al., 2018), the interoperability among agents was measured using “no-regret” learning, which calculated the distance between the sum of the rewards earned by each agent under an optimal policy and the existing cooperative policy. The agents cooperated well when the sum of these differences was minimized. In (Diyan et al., 2020) the authors proposed a Q-learning algorithm for a smart building scheduling considering a comfort penalty. The proposed method was proved to reduce energy consumption and user discomfort compared to Least Slack Time-based scheduling. In (Xu et al., 2020) the author proposed a fully distributed multi-agent Q-learning algorithm with no observation for hour-ahead energy consumption and EV charging decisions. The Q-learning algorithm was proved to outperform a GA based solution since the RL agents consider both the current and future rewards. A feed forward DNN is added in the architecture to predict future trends of electricity price and solar generation according to real-world data. As MARL is computationally expensive, the hardware must be efficient with desired performance. The hardware issue is addressed in (Xiongfeng Zhang et al., 2021) where the authors explored the feasibility of a practical implementation of a multi-agent Q-learning framework in a smart grid to optimize energy consumption of various devices. The author provided guidelines on how to implement an experimental testbed to validate the MARL algorithm. And based on the experimental results, they concluded that the MARL-Q-learning was able to achieve optimal load control on a hardware EMS containing several LEDs and motors. In (Chen et al., 2021) the author proposed the use of a Preference Based multi-objective RL algorithm that applies Q-tables corresponding to the number of home appliances to expediate the learning process in response to possible changes of user preference, in conjunction with Q-learning for battery system control. Five agents, an ESS and four appliances, achieved fast convergence to a steady objective considering price and renewable uncertainty, by shifting or reducing energy consumption in a SB. The author in (Hao et al., 2020) proposed a finite non-cooperative MARL algorithm with a discounted Q factor hybrid architecture for SBs to control their HVAC system with the objective of minimizing cost constrained by a temperature boundary.

### *3.2. Deep Q-network*

DQN, the first DRL algorithm, combines RL with DNNs to overcome the instability and divergence issues of Q-learning. Several techniques were adopted to stabilize the learning process, e.g., experience replay and target network. Experience replay stores the experience transitions on a replay memory and draws samples of them uniformly at random for training, establishing greater data efficiency when compared to the standard Q-learning algorithm. In addition, randomizing the samples contributes to the reductions of their correlations and the variance of updating DNN weights. DQN commonly utilizes two NNs, a prediction and a target network. The target network is adopted to improve stability of the training process by copying a separate network with a longer update period for the computation of the target value (Yu et al., 2021). DQN can only deal with discrete action spaces, and thus is not practical for regulation tasks in energy systems that require continuous action spaces. DQN may also suffer from instability issues as dimensionality increases because it assigns a value to every possible action and then selects the action with the highest value. Selecting an action with the highest value is very difficult if not impossible if the action space is continuous or very large. The DQN replay buffer further complicates efficiency, since it takes a long time to train and is limited to off-policy methods. Overestimation also may occur, as a result from the same NN being used for both policy estimation and evaluation. Double-deep Q-network (DDQN) solves this issue by using separate NNs for action selection and action evaluation.

In (Xiaohan Fang et al., 2021) the author proposed a multi-agent DQN (MADQN) for distributed energy management for a double-auction microgrid market. Optimal equilibrium in RL iteration guarantees that all agents benefit from fairness, which consequently improves algorithm convergence. The MADQN has no dependence on accurate model or parameter estimation and unlike centralized scheduling methods, and it solves the constraint in distributed manner for individual agents. However, scalability is not considered, and aggregation of resources is implemented instead. Not always is a homogeneous DRL method used for MARL. In (Nie et al., 2020) the author proposed a double DRL algorithm, both DQN and DDPG, for a load agent and a generation agent. The agents interact with the environment independently gathering their own rewards, with no communication. However, to interact with and influence each other, to one agent, the other agent becomes part of the environment. Although the agent realizes a dual control, i.e., energy storage management on the source side and load shedding on the load side, the proposed algorithm lacks scalability and can improve control performance with a communication layer between both agents. Similarly, the author in (Chenyu Guo et al., 2021) proposed a bi-level distributed optimal EMS controlled with DRL Dueling-DQN and DDPG algorithms, with prioritized experience replay. The upper-level does not require detailed private information from the lower-level agents and only key information is transmitted using a communication channel. After training the agents off-line with historical data, the model-free DRL can adapt to different environments effortlessly with high efficiency and favorable scalability.

### 3.3. Actor critic

Experience replay has been applied to improve stability in value-based DRL methods, but it introduces variance in policy-based methods. Separation of the policy and the value function networks produces better results. By using hybrid policy gradients with value-based methods, Actor-Critic (AC) architecture turns out to be more robust in EMS (Arwa and Folly, 2020). The actor-critic method involves two DNNs, the policy network (Actor) and the value function estimator (Critic). The actor takes an action based on the input environment states, while the critic returns the estimated value of an action based on its observation of the environment states and the reward from the actor's action. The actor uses a gradient ascent method to maximize objective reward while the critic uses gradient descent to minimize error in the value function estimation. Instability with Q-learning recursions when applied to DNN can be further addressed with an asynchronous advantage actor-critic (A3C) algorithm. A3C trains several agents with different copies of the environment asynchronously. A2C is a synchronous advantage actor critic technique that can achieve the same or even better results than A3C (Arwa and Folly, 2020).

In (Dong et al., 2021) the author compared an A3C with DQN, DDPG framework for a distribution system's economic dispatch. The modeled distribution network includes an external power grid, a Wind Turbine, a PV system, an energy storage system (ESS), natural gas stations, gas loads, heat loads, and an electric heating furnace, etc. A3C shortened training time by 30% and 37% and reduced daily operating cost by 5.2% and 3% compared to DQN and DDPQ, respectively. One challenge of MARL is that multi-agent domains are nonstationary from an agent's perspectives since other agents' interactions with the environment. Furthermore, the action space grows exponentially with the increased number of agents, and the learning becomes very difficult due to partial observability or limited communication. Thus, a centralized Critic and a decentralized Actor, or a linear decomposition of the joint value function across agents were explored. In (Shin et al., 2020) the author proposed an actor and critic DRL model to manage PV/ESS EVCSs in a distributed manner. These agents could communicate their embedded state information, making convergence stable. To further prove performance, the method was evaluated on a large-scale data set, and it was confirmed to achieve desired performance. A multi-energy management framework with decentralized execution and centralized training, formulated as a partially observable MDP was proposed in (Dafeng et al., 2022). Soft actor-critic with an attention mechanism was adopted to enhance policy stability and encourage agents to focus on important energy related information, improving exploration efficiency and robustness. In addition, a novel reward based on the Lagrange multiplier method to ensure capacity constraints of ESSs was implemented. Simulations and results based on actual data set verified high scalability and the algorithms optimization.

### 3.4. Deep deterministic policy gradient

If the policy is deterministic, the AC method is also called deep deterministic policy gradient (DDPG). The methods, AC and DDPG, perform well online because of their high efficiency and speed. In DDPG, the action-value is used to update the critic network while in other AC cases, the state-value is used to update the critic network, otherwise the architecture is the same. DDPG performs well for continuous action spaces, however, DDPG directly chooses deterministic actions ignoring uncertainty and randomness. Random fluctuations and incomplete modeling in an EMS burden the application of deterministic methods. It is more practical to use a probabilistic control policy for BTM energy management (Lee et al., 2020). Another issue with policy gradient-based algorithms is that they can only handle one action at a time. The actor may only return one action or the probability of taking one action at a particular state (Arwa and Folly, 2020).

In (Li and Yu, 2020) the author proposed a centralized training/decentralized implementation DRL framework for an optimal automatic generation control (AGC). A multi-agent distributed multiple improved deep deterministic policy gradient (MADMI-TD3) algorithm, an extension of DDPG combining several RL techniques, is proposed to improve stability and training efficiency. The AGC framework included various units including distributed generations and flexible loads to solve a coordinated control and dispatch for an electrical system. MADMI-TD3 is proved to be effective in global search and optimizing speed in a random environment, which employs different parameters of multiple actor networks for distributed optimization of control performance and economic benefits.

### 3.5. PPO

AC algorithms change their policy according to a gradient descent update, which introduces a challenge in selecting the step size for the updates. Large step sizes can cause high performance variation between iterations, resulting in instability during training. This is detrimental in agents that have a high probability of gathering bad data. Policy optimization methods use a probability ratio between old and new policies to tackle this challenge (Arwa and Folly, 2020). Trust Region Policy Optimization (TRPO)'s updates are limited to a "trust region" to avoid misleading observations. PPO is a simpler version of TRPO and has shown to perform better than most algorithms with an AC architecture in solving multi-dimensional continuous environments. PPO's linearization of both the objective function and step size makes it simpler and more robust to solve the issue of curse of dimensionality, due to their ability to easily optimize objectives more efficiently in highly uncertain, continuous, and multi-dimensional environments (Arwa and Folly, 2020). Most BTM EMSs are multidimensional and have a continuous state space, challenging PPO to be the best algorithmic choice. PPO is proven to have more stable convergence than DDPG methods in some studies (Lee et al., 2020).

## 4. Proposed solution

The increasing complexity and need for a practically feasible, efficient, and scalable BTM EMS of future communities cannot be satisfied by the existing centralized methods (Sun and Yang, 2019). Many state of the art solutions are either based on a single entity or central aggregate control of multiple entities which selects and sends translated decisions to individual entities (Abdullah et al., 2021). A BTM EMS may contain numerous heterogeneous entities; therefore, is exposed to frequent and various system dynamics. The lack of scalability of centralized algorithms makes it expensive, if not unfeasible to obtain a global optimal solution with large data acquirement requirements. There are some research focused on decentralized strategies or hybrid centralized-decentralized management for EMSs (Abdullah et al., 2021). Unlike centralized methods, decentralized coordination algorithms can handle high uncertainty and flexibility. As we analyzed extensively above, DRL is promising in energy management for BTM sources, proving in general less computationally heavy compared to conventional methods, as well as faster converge to acceptable near optimal solution due to offline training. However, they still need great computational power, that may cost energy management systems to place machines at every entity.

### 4.1. MARL methods

MARL could be categorized into three frameworks, fully centralized training and execution, fully decentralized training and execution, and centralized training with decentralized execution. In a fully centralized strategy, the agent collects aggregated information from all entities to decide on a joint action set for all entities. In a fully decentralized strategy or distributed strategy each entity is an agent and chooses an action to optimize their own reward independently, assuming there is no central controller (Abdullah et al., 2021). Centralized training and decentralized execution (CTDE) improve decentralized RL by using an actor-critic structure and learning a centralized critic to reduce variance (Lee et al., 2020).

### 4.2. Proposed multi-agent actor-critic methods

The simplest approach to learning multi-agent environments is to independently train agents. Q-learning is most commonly used but does not perform well (Ryan Lowe et al., 2017). Q-learning is not capable to train centrally and execute in a decentralized manner without making assumptions about the environment, since Q-learning uses the same information for training and execution. In an ideal MARL based EMS for various BTM sources, i) each agent should be capable of choosing an action based on its local observation; ii) different agents can coordinate or compete with each other; iii) agents do not have to have communication channels among themselves; iv) their reward functions could be dependent on the future states and actions; and v) each agent's state and observation spaces are flexible to be continuous, discrete or mixed. These requirements and the non-stationary environment mean that the value-based algorithms such as Q-learning are no longer capable because they depend on the Markov assumption that the state transition and reward function are dependent only on the state and action of a single agent at the last time-step. This makes the use of past experience replay, which is critical for stabilizing DQN, unstable. Furthermore, most often, an accurate model of the dynamic environment is not given or hard to formulate mathematically, leaving out model-based algorithms as a solution. Policy gradient suffers from a variance that increases with the number of agents grow. Thus, we propose two extensions of actor-critic algorithms, multi-agent PPO and multi-agent DDPG, as potential solutions to address these challenges, under a centralized training and decentralized execution framework as shown in Fig. 2.

The lack of visibility regarding each agent's strategy creates a local impression of a non-stationary environment. Thus, a centralized critic network is proposed to allow the aggregator to provide a certain information transparency between agents to guide the learning process. During training the critic acts as a central coordinator. The advantage of a decentralized energy management control, after the training phase, is proposed to allow each agent to act independently in the execution phase without communicating with other agents. In addition, by assuming a centralized value function, the full global state reduces a partially observable MDP to a fully observable MDP, which guarantees a quicker and easier value learning. MAPPO addresses the scenarios of cooperative learning with shared utility, but also proves to adapt to non-cooperative goals. Similar, but not as accurate and quick performance can be achieved by learning MAPPO in a fully-distributed manner. This would be most useful, since connectivity of smart devices is limited, and local computing power is necessary.

MADDPG has the capability to learn policies using their own observations only at execution time, without assuming a differentiable model of the environment dynamics or a communication structure between agents, Thus, it is applicable to both cooperative and competitive interactions. In addition, it can act in a mixed cooperative-competitive

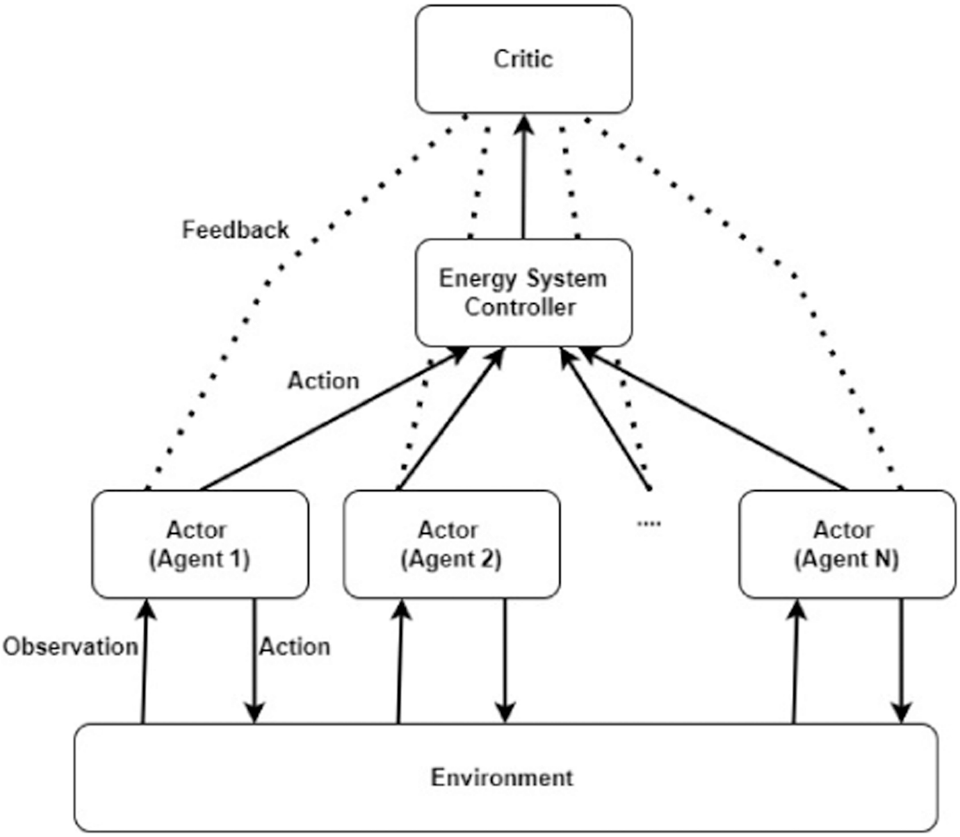

**Fig. 2.** Centralized training and Decentralized execution EMS for BTM resources.

environment involving both physical and communicative behavior. Similar to MAPPO, MADDPG performs centralized training, the critic acquires extra information to eases training, and decentralized execution, each actor only uses its local information to keep data private. After the training, the decentralized actors can be applied to a cooperative or competitive setting. Agents are capable of learning approximate models of other agents online and use them in their own policy learning procedure. Some argue that MAPPO is slower than MADDPG, but (Chao Yu et al., 2021) proves that it can perform significantly faster than off-policy MARL methods for both cooperative and competitive environments. The proposed actor-critic methods consider action policies of other agents and are expected to successfully learn policies that require complex multi-agent coordination (Ryan Lowe et al., 2017; Chao Yu et al., 2021).

## 5. Discussion and future directions

Unlike games (Go) that have strict rules and clear rewards, the energy management problem for the BTM resources is much more complicated and uncertain. For example, renewables cannot be predicted accurately, equipment may fail at any time, and in various locations there may be different EV charging/discharging behaviors and load management strategies (Zhang et al., 2019). State and action spaces are not clearly defined and has high complexity, thus requiring thoughtful planning, initialization, and critical algorithm parameters must be set to carefully balance exploration and exploitation (Zhang et al., 2019). Coordination among agents representing BTM resources is difficult to implement due to dynamic environment and heterogeneous agent models. Compared to a single-agent RL algorithm, MARL allows for more complex environments with a high feature dimension, high action dimension, and continuous or mixed space for both states and actions. Each agent has its own reward function and learns independently from other agents, usually with partially or fully observable information among agents. Multiple agents interacting in the environment can overcome the issue of dimensionality and the discretization problem a single-agent faces. However, MARL are not widely used in literature due to their complexity and non-stationary issue. Complexity is a consequence of a high dimensional environment and the need to train several agents simultaneously, which is computationally expensive to implement. If several agents are learning independently in the same environment, it becomes non-stationary, meaning the agents' version of the environment is not fixed due to other agents regularly altering the state of the environment. In simpler terms, the state of the environment changes based on all actions taken by all agents, not a single agent. Another challenge is adjusting to the dynamic behavior of other agents or measure the cooperation among agents to see how well stability has occurred (Ahrarinouri et al., 2021). Multi-agent DRL energy management algorithms with complex reward components should be designed to efficiently promote the coordination. Lastly, when the similarity gap is large (e.g., the dimensions of state spaces and action spaces in two MDPs are different), how to design efficient inter-task mapping function and select proper form of the transferred knowledge, especially for multi-agent DRL-based problems (Yu et al., 2021) is a big challenge.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Abdullah, H.M., Gastli, A., Ben-Brahim, L., 2021. Reinforcement learning based EV charging management systems–a review. IEEE Access 9, 41506–41531. https://doi.org/10.1109/ACCESS.2021.3064354.

Ahrarinouri, M., Rastegar, M., Seifi, A.R., 2021. Multiagent reinforcement learning for energy management in residential buildings, in. IEEE Trans. Ind. Inform. 17 (1), 659–666. https://doi.org/10.1109/TII.2020.2977104.

Air Quality, Energy & Sustainability, New Jersey Department of Environmental Protection, Available at: ⟨https://www.state.nj.us/dep/aqes⟩/.

Arulkumaran, K., Deisenroth, Ma.P., Brundage, M., Bharath, A.A., 2017. A brief survey of deep reinforcement learning. Sp. Issue Deep Learn. Image Underst. 37.

Arwa, E.O., Folly, K.A., 2020. Reinforcement learning techniques for optimal power control in grid-connected microgrids: a Comprehensive Review, in. IEEE Access 8, 208992–209007. https://doi.org/10.1109/ACCESS.2020.3038735.

Bahdanau, D , An actor-critic algorithm for sequence prediction, 2017.

Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M., 2013. The arcade learning environment: an evaluation platform for general agents. Jou. Artif. Intel. Res. 47 https://doi.org/10.1613/jair.3912.

Chao Yu, 2021. The surprising effectiveness of MAPPO in cooperative, multi-agent games. CoRR abs/2103. 01955.

Chen, S.-J., Chiu, W.-Y., Liu, W.-J., 2021. User preference-based demand response for smart home energy management using multiobjective reinforcement learning, in. IEEE Access 9, 161627–161637. https://doi.org/10.1109/ACCESS.2021.3132962.

Chen, Wuhui,etc., Deep Reinforcement Learning for Internet of Things: A Comprehensive Survey. IEEE Communications Surveys & Tutorials (2021).

Chenyu Guo, 2021. Optimal energy management of multi-microgrids connected to distribution system based on deep reinforcement learning. Int. J. Elect. Power Energy Syst. 131, 107048.

Chiş, A., Lundén, J., Koivunen, V., 2017. Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Trans. Vehicular Technol. 66 (5), 3674–3684. https://doi.org/10.1109/TVT.2016.2603536.

Clement-Nyns, K., Haesen, E., Driesen, J., 2010. The Impact of charging plug-in hybrid electric vehicles on a residential distribution grid. IEEE Trans. Power Syst. 25 (1), 371–380.

Dafeng, Zhu, 2022. Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park. Appl. Energy 311, 118636.

Diyan, D., Muhammad, 2020. A Multi-Objective Approach for Optimal Energy Management in Smart Home Using the Reinforcement Learning. In: Sensors, 20. Basel, Switzerland, p. 3450. https://doi.org/10.3390/s20123450.

Dong, J., Wang, H., Yang, J., Lu, X., Gao, L., Zhou, X., 2021. Optimal scheduling framework of electricity-gas-heat integrated energy system based on asynchronous advantage actor-critic algorithm, in. IEEE Access 9, 139685–139696. https://doi.org/10.1109/ACCESS.2021.3114335.

Fernandez, Gabriel I., et al. Deep Reinforcement Learning with Linear Quadratic Regulator Regions. arXiv preprint arXiv:2002.09820 (2020).

Foruzan, E. , Soh, L. , Asgarpoor, S. , Reinforcement learning approach for optimal distributed energy management in a microgrid, in IEEE Transactions on Power Systems vol. 33 5 2018 5749 5758 doi: 10.1109/TPWRS.2018.2823641.

Frikha, etc., Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey. Computer Communications 2021.

Grid-interactive efficient buildings, Available at: https://www.energy.gov/eere/buildings/grid-interactive-efficient-buildings.

Hadusha, S.Y., Meeus, L., 2018. DSO-TSO cooperation issues and solutions for distribution grid congestion management. Energy Policy 120, 610–621.

Hao, J. , Gao, D.W. , Zhang, J.J. , Reinforcement learning for building energy optimization through controlling of central HVAC system IEEE Open Access Journal of Power and Energy 7 2020 320 328 doi: 10.1109/OAJPE.2020.3023916.

Harley, S.W. , Tsvetkova, A. , Potential impacts of plug-in hybrid electric vehicles on regional power generation, Technical Report, ORNL/TM-2007/150, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, USA, 2008.

Kim, Sunyong, et al., 2018. Reinforcement learning based energy management algorithm for smart energy buildings. Energies 11, 2010.

Kohl, N. , Stone, P. , Policy gradient reinforcement learning for fast quadrupedal locomotion, in Proceedings - IEEE International Conference on Robotics and Automation, 2004, vol. 2004, no. 3. doi: 10.1109/robot.2004.1307456.

Kolokotsa, Denia, Kampelis, Nikos, 2021. Smart Buildings, Smart Communities and Demand Response. Wiley-ISTE. ISBN: 978-1-119-80423-9 January.

Lee, J. , Wang, W. , Niyato, D. , Demand-Side Scheduling Based on Multi-Agent Deep Actor-Critic Learning for Smart Grids, 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 2020, pp. 1–6, doi: 10.1109/SmartGridComm47815.2020.9302935.

Li, J. , Yu, T. , Virtual Generation Alliance Automatic Generation Control Based on Deep Reinforcement Learning, in IEEE Access vol. 8 2020 182204 182217 doi: 10.1109/ACCESS.2020.3029189.

Mnih, V., et al., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540). https://doi.org/10.1038/nature14236.

Nie, H. , Chen, Y. , Xia, Y. , Huang, S. , Liu, B. , Optimizing the Post-Disaster Control of Islanded Microgrid: A Multi-Agent Deep Reinforcement Learning Approach, in IEEE Access vol. 8 2020 153455 153469 doi: 10.1109/ACCESS.2020.3018142.

Ranzato, M. , etc., Sequence level training with recurrent neural networks, 2016.

Remani, T., Jasmin, E.A., Ahamed, T.P.I., 2019. Residential load scheduling with renewable generation in the smart grid: a reinforcement learning approach. IEEE Systems J. 13 (3), 3283–3294. https://doi.org/10.1109/JSYST.2018.2855689.

Ryan Lowe, 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. CoRR abs/1706. 02275.

Schulze, Mike, et al., 2016. Energy management in industry – a systematic review of previous findings and an integrative conceptual framework. J. Clean. Product. 112, 3692–3708.

Shin, M. , Choi, D.-H. , Kim, J. , Cooperative management for PV/ESS-enabled electric vehicle charging stations: a multiagent deep reinforcement learning approach, in IEEE Transactions on Industrial Informatics vol. 16 5 2020 3493 3503 doi: 10.1109/TII.2019.2944183.

Sun, Q., Yang, L., 2019. From independence to interconnection — a review of AI technology applied in energy systems, in. CSEE J. Power Energy Syst. 5 (1), 21–34. https://doi.org/10.17775/CSEEJPES.2018.00830.

Tesauro, G., 2018. Temporal difference learning and TD-gammon. ICGA J. 18 (2) https://doi.org/10.3233/icg-1995-18207.

Tranfield, David, et al., 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Brit. J. Manag. 14, 207–222.

Xiaohan Fang, et al., 2021. Multi-agent deep reinforcement learning for distributed energy management and strategy optimization of microgrid market. Sustain. Cities Society 74, 103163.

Xiongfeng Zhang, et al., 2021. Testbed implementation of reinforcement learning-based demand response energy management system. Appl. Energy 297, 117131.

Xu, X. , Jia, Y. , Xu, Y. , Xu, Z. , Chai, S., Lai, C.S. , A multi-agent reinforcement learning-based data-driven method for home energy management, in IEEE Transactions on Smart Grid 1 4 2020 3201 3211 doi: 10.1109/TSG.2020.2971427.

Bengio, Y. Courville, A. , Vincent, P. , Representation learning: a review and new perspectives IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 35 8 2013 doi: 10.1109/TPAMI.2013.50.

Yoldas, Y., Goren, S., Onen, A., 2020. Optimal control of microgrids with multi-stage mixed-integer nonlinear programming guided $q$-learning algorithm, in. J. Modern Power Syst. Clean Energy 8 (6), 1151–1159. https://doi.org/10.35833/MPCE.2020.000506.

Yu, L., Qin, S., Zhang, M., Shen, C., Jiang, T., Guan, X., 2021. A review of deep reinforcement learning for smart building energy management, in. IEEE Inter. Things J. 8 (15), 12046–12063. https://doi.org/10.1109/JIOT.2021.3078462.

Zhang, Zidong et al. Deep reinforcement learning for power system: An overview. CSEE Journal of Power and Energy Systems (2019): n. pag.

**Patrick Wilk** received his B.S. degree in Electrical and Computer Engineering from Rowan University, USA, in 2020. Presently, he is in the process of completing his PhD in Electrical Engineering at Rowan University. His research interests and studies include electricity market bidding strategy, economic dispatch, optimizing energy system operations, and machine learning in power system operation.

**Ning Wang** is currently an assistant professor in the Department of Computer Science at Rowan University, Glassboro, NJ. He received his Ph.D. degree in the Department of Computer and Information Sciences at Temple University, Philadelphia, PA, USA, in 2018. He obtained his B.E. degree in School of Physical Electronics at University of Electronic Science and Technology of China, Chengdu, Sichuan, China, in 2013. He currently focuses on communication and computation optimization problems in Internet-of-Things systems and operation optimization in Smart Cities applications.

**Jie Li** received her B.S. degree in Electrical Engineering and M.S. degree in Systems Engineering from Xi'an Jiaotong University, China, in 2003 and 2006, respectively, and Ph. D. in EE from Illinois Institute of Technology, USA, in 2012. Presently, she is an Associate Professor in the ECE Department at Rowan University, and before that, she is working at Clarkson University. Her research interests include power system operation and planning, microgrid, and electricity market bidding strategy.