Les Cahiers du GERAD ISSN: 0711–2440

Multi-agent hierarchical reinforcement learning for energy management

I. Jendoubi, F. Bouffard

G–2021–46

August 2021

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2021
– Bibliothèque et Archives Canada, 2021

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2021
– Library and Archives Canada, 2021

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Multi-agent hierarchical reinforcement learning for energy management

Imen Jendoubi [a, b]

François Bouffard [a, b]

[a] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[b] *Department of Electrical and Computer Engineering, McGill University, Montréal (Qc), Canada, H3A 0E9*

imen.jendoubi@mail.mcgill.ca
francois.bouffard@mcgill.ca

August 2021
Les Cahiers du GERAD
G–2021–46

Abstract : The increasingly complex energy systems are turning the attention towards model-free control approaches such as reinforcement learning (RL). This work proposes novel RL-based energy management approaches for scheduling the operation of controllable devices within an electric network. The proposed approaches provide a tool for efficiently solving multi-dimensional, multi-objective and partially observable power system problems. The novelty in this work is threefold: We implement a hierarchical RL-based control strategy to solve a typical energy scheduling problem. Second, multi-agent reinforcement learning (MARL) is put forward to efficiently coordinate different units with no communication burden. Third, a control strategy that merges hierarchical RL and MARL theory is proposed for a robust control framework that can handle complex power system problems. A comparative performance evaluation of the proposed control approaches is also presented. Experimental results of two typical energy dispatch scenarios show the effectiveness of the proposed approaches.

Keywords: Energy management, hierarchical reinforcement learning, microgrid, multi-agent reinforcement learning, options framework

1 Introduction

The evolution towards decarbonized energy systems is entailing deep changes at the levels of generation and demand. From one side, the integration of more renewable energy sources is bringing stochasticity and random fluctuations to the generation side. Furthermore, emerging microgrids, and zero-emission communities are regarded as promising frameworks for harnessing the potential of renewable generation and assisting the operation of the main grid technically and economically [16]. These patterns are leading the evolution towards a decentralized grid architecture with many small interconnected generation units and bi-directional power flows. From the load side, the emergence of prosumers and the movement towards the electrification of transport and heating are increasing the uncertainty and peaks on the demand side. These changes are making the electric networks more complex with high levels of uncertainty and non-linearity due to the presence of non-linear loads and non-linear generation such as solar photovoltaics (PV). This makes the planning and operation of such systems more challenging. Meanwhile, ongoing technological advances are offering various opportunities for the current electric networks. Advances include improved information and communication technologies, smart metering and sensing technologies, power electronic devices and energy storage systems (ESSs) which are regarded as flexibility providers and supporters of grid operation. Furthermore, rapidly improving control algorithms and computational capabilities are opening up the possibility of developing and implementing more advanced control frameworks. These challenges and opportunities are calling for the development and adoption of more innovative and up-to-the-minute solutions in order to handle the emerging challenges of planning and operating energy systems more effectively and reliably.

The most common approaches that are typically used to solve power system optimal decision and control problems are model-based (e.g. dynamic programming, stochastic programming, model predictive control, heuristic approaches, etc.) In general, a model-based approach requires a precise model of the underlying system dynamics, a predictor for forecasting the sources of uncertainty, and a solver for solving the formulated problem. The main limitation of a model-based approach is its reliance on a precise system modeling which is not realistic especially in highly uncertain and non-linear environments where obtaining an accurate model is usually difficult and costly. Its performance is also dependent on the predictions' quality of the uncertainty sources (generation, demand, electricity prices, etc.) [15]. Furthermore, model-based approaches lack adaptability and may not be computationally efficient for real-time control as many approaches require the reiteration of all the computations whenever a new decision is to be taken.

To overcome these limitations, model-free and data-driven approaches have been proposed in the literature. As have been reviewed in many studies such as [2, 5, 11, 12, 23, 26, 27] , there is an increasing interest towards the application of reinforcement learning (RL), a data-driven decision making approach, in solving typical energy management problems. In general, RL is a branch of machine learning (ML) in which an agent acquires the ability of solving one or many given tasks through interaction with its surrounding environment. To learn solving a task, the agent takes actions and subsequently receives rewards from the environment, which reflects how "good" its action was. Based on the collected experience, the agent progressively learns how to take actions in the direction that maximizes the rewards obtained. The growing attention surrounding RL is motivated by the success it has achieved mainly in the fields of games and robotics [9]. A particular category of RL is deep reinforcement learning which has been receiving major research focus as it was able to reach super-human intelligence in complex ATARI video games, the game of go, and others [9]. Such features make RL algorithms potential candidates for solving complex control problems in the area of power systems.

The emerging small-scale and distributed decision makers (e.g. distributed energy resources (DERs), residential loads, electric vehicles (EVs), etc) are increasing the dimensionality of energy management problems in terms of the number of states, actions and objectives. An optimal coordination between different interconnected units linked by market transactions and physical electric flows is required to reach a sustainable, cost-effective, stable and reliable operation of a given network[23].

A typical study environment can be an eco-neighborhood or a microgrid (grid-connected or islanded) that is composed of a set of loads, DERs (renewable or non-renewable), energy storage systems, capacitor banks, etc. The objective is to find the optimal dispatch of controllable resources in order to meet the load with the minimum environmental impact, and the maximum profit for the suplier and customers. Other objectives of interest include peak shaving, power fluctuations reduction, maintaining customer comfort, voltage regulation, etc.

Therefore, multi-agent modeling of energy management problems is becoming more evident. Multi-agent RL (MARL) is a promising field that can model each decision-making unit as an autonomous agent that takes actions based on its local observation [7]. Although promising, the application of MARL in power system decision and control is still in its early stages as only a few works in this field have been proposed recently in the literature [5]. Focusing on single-agent RL (SARL)-based control strategies limits the applicability of RL to specific small-scale scenarios while the emerging electric and energy networks are becoming increasingly multi-dimensional. Putting forward MARL in energy management problems is therefore of great interest.

Another limitation of the current RL applications in power system decision and control is that the most prevailing algorithms are based on Q-learning and its variants [12]. The problem of such approaches is that they can solve problems with discrete action spaces. In the case of controllable units with continuous action spaces such as ESSs or heating ventilation and air conditioning (HVAC) units, a discretization of the action space is required which, in turn, results in sub-optimal solutions. Therefore, it would be of interest to investigate approaches that allow the use of continuous action spaces such as policy gradient (PG)-based approaches.

As the control problems are becoming increasingly complex, primitive RL may not be enough to learn sufficiently complex actions. Furthermore, some real control problems may violate the Markov property that supposes that, given the present information, the future of the given process is independent of its past. Such assumption is a basis for the development of RL algorithms, and is therefore required for their proper functioning and convergence to the optimal solution. However, no solutions have been provided in the literature to address these issues. Although RL is a wide research area with diverse sub-areas, the choice of RL algorithms in the area of power systems is focusing on a very limited choice of simple algorithms while a large part of RL's potential is still unexplored. In their thorough studies, [11] and [23] have highlighted the need for applying new and diverse RL-based approaches to solve various case studies of typical energy management problems.

There is also a lack of benchmarking through the provision of a comparative evaluation of diverse RL algorithms when solving the same energy management task [1]. In this direction, hierarchical RL, a promising sub-area of RL, can be a potential alternative in this regard as it intends to solve complicated problems by decomposing it through the learning of spatio-temporal abstract actions rather than primitive actions, and it can cope well with non-Markovian dependencies. Further, merging hierarchical RL with MARL theory can potentially give rise to performant and highly robust control strategies where different agents are coordinated and jointly learn to take spatio-temporally extended actions in highly complex and uncertain environments. In particular, the control hierarchy for generating the action is not composed of only one layer as in primitive RL, but rather it is extended to multi-layers in the space or time dimension.

In this work, we design and implement efficient and robust control strategies in the area of energy management within microgrids and eco-neighborhoods. An eco-neighborhood is a self-productive community of multi-dwelling residential buildings endowed with various DERs to reduce its reliance on conventional generation with high Greenhouse Gas Emissions. This work aims to propose a control framework that can deal with such complex and uncertain frameworks with the overall objective of promoting smart, sustainable and cost-effective operation of the studied systems.

Two typical control problems are studied as sequential decision making processes and solved by different RL-based control strategies to compare their control performances. In particular, the focus is

on highlighting the potential of hierarchical RL in the area of power systems which is investigated for the first time in this work. The potential of MARL is also put forward by designing control strategies that are able to deal with multi-objective and high dimensional state and action spaces. In particular, approaches at the intersection of hierarchical RL and MARL are proposed in order to give rise to performant and robust control strategies in the area of power systems control. Experimental results based on real data validate the performance of the proposed control strategies for solving two typical energy management problems within a microgrid and eco-neighborhood. The RL-based solutions are also compared with the solution of a deterministic optimization problem.

The rest of this work is organized as follows: Section 2 describes the studied control framework. Section 3 provides an overview about the applications of MARL in power systems and presents a formulation of the studied control problems. The proposed RL-based control strategies are described in Section 4 along with an introduction to hierarchical RL. Experimental results are reported in Section 5 with a discussion about the results and their applicability. Finally, the paper is concluded in Section 6.

2 Control problem description

In general, we consider a system composed of various interconnected electrical components including non-controllable loads and DERs, as shown in Figure 1. These DERs can be small-scale generators such as diesel generator, controllable loads (e.g. HVAC, EVs), ESSs, etc. The system can be self-sufficient relying on its internal resources for meeting the demand or can receive backup from a stiff source (e.g. electricity provider, main grid, etc.) Overall, the main objective within such a system is to optimally manage the electric flows through an optimal dispatch of controllable units in order to meet the load with an optimal operation cost and reduced negative environmental impact. Such a control framework is inherently multi-objective with high-dimensional state and action spaces as it features a set of interconnected and interacting units linked by electrical flows and market transactions.

For the sake of illustration and without loss of generality, we consider solving two power system optimal decision and control problems which are typical instantiations of the system described previously:

- Problem (A): A microgrid in islanded mode with residential load, one ESS, a diesel generator and high PV production. The objective is to implement an energy dispatch strategy for determining the optimal charging/discharging power for the ESS in order to meet the load, maximize self-consumption from PV generation, and minimize diesel generator operation cost.
- Problem (B): An eco-neighborhood consisting of two multi-dwelling residential buildings, two ESSs and a shared PV production. The objective is to find the optimal coordination among the ESSs in order to reduce the energy purchased from the electricity provider while reducing power fluctuations and peaks in the aggregated net demand power to the benefit for the electricity provider.

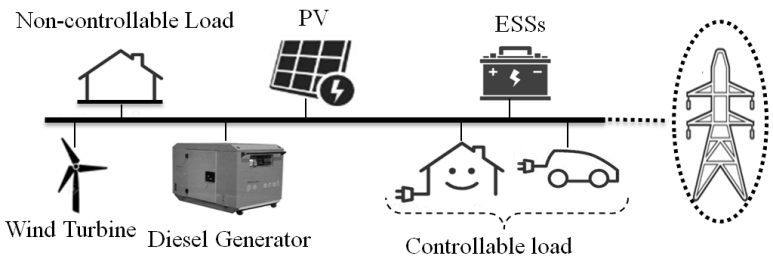

Figure 1: Control environment.

The two problems are fully cooperative as the decision-making units involved have a common goal to achieve. Further, the problems are multi-objective featuring uncertainties (from load and PV) and high-dimensional state space. Another challenge is that the control environment is partially observable because at the beginning of each time step, an observation is obtained instead of the full state of the system, as the load and renewable production cannot be exactly known in advance.

To handle the underlying complexities and uncertainties, we propose to solve such control problems by hierarchical RL and MARL-based control strategies.

3 Problem formulation

In this section, MARL and its application in the area of power system optimal decision and control are briefly reviewed. Then, details about the formulation of the studied control problem are provided.

3.1 Application of multi-agent reinforcement learning

MARL is a RL-based decision making approach for multi-agent problems. It is an active research topic that is gaining increasing attention, and is still undergoing important developments and breakthroughs. There is an increasing attention towards the application of MARL for solving typical smart grid control problems as these systems are becoming more decentralized where a coordination among homogeneous or heterogenous interconnected units is required to reach the overall desired objectives. In this case, the use of centralized approaches to solve such a control problem is usually prone to scalability and communication issues. Furthermore, the application of single-agent based approaches may not be suitable because, as the name suggests, these are intended to deal with small-scale problems involving only one type of dispatchable units, while the emerging power systems are becoming increasingly multi-dimensional and decentralized with diverse dispatchable units. Therefore, multi-agent modeling is more evident and natural for solving such energy management scenarios effectively.

3.1.1 Multi-agent reinforcement learning variants

There are many variants of MARL algorithms. The category of independent learning is the simplest as it is based on the naive extension of single-agent based approaches. Each agent learns independently and treats other agents as part of the environment. Although simple, the independent learning-based approach is computationally expensive and does not perform well in environments where complex coordination between agents is required. Furthermore, the control environment is prone to the non-stationarity issue because the agents' policies are updated simultaneously, and the policies change from the perspective of each individual agent. It is noted that the environment stationarity is a crucial assumption necessary for the convergence of RL-based algorithms. In multi-agent environments, ensuring a stationary environment is possible if each agent has knowledge about other agents' policies which is challenging as the agents are usually required to act independently based on their local observations only.

Another approach is based on centralized-training and execution as it uses a centralized agent that have common information (i.e. observations and actions of all agents) [17]. This approach is mostly suitable for applications where decentralized operation of agents is not required [17] and is applicable only in fully cooperative scenarios [19].

In some applications, the decentralized execution is prioritized in order to avoid the need for communicating information in real-time and its related problems. To this aim, centralized-training decentralized-execution (CTDE) approaches have been proposed in the literature, and can be applied to solve problems with discrete or continuous action spaces.

The idea is to introduce a central unit that has access to common information about all the agents. This unit is however only needed during the training phase to guide and ease the learning process.

Once an optimal policy is learned by each agent at the end of the training phase, it is used to generate local decisions based on local observations in a decentralized way during the real-time execution phase. Approaches based on CTDE can be value-based such as QMIX, or actor-critic based (e.g. COMA for discrete action spaces and multi-agent deep deterministic policy gradient (MADDPG) for continuous action spaces) [17, 19].

It is noted that MARL is an open research area that is continuously giving rise to diverse approaches besides the algorithms above.

3.1.2 Previous work in the area of energy management

Compared with SARL, the application of MARL in the field of energy management is still in its early stages as only very few works in this field have been recently proposed as seen in [5].

In the category of independent learning, studies such as [8, 20, 24, 25] have been proposed. A multi-agent control scheme has been developed in [24] to enable energy sharing between buildings with the overall aim of reaching a zero-energy community status. Each building, modeled by an agent, learns its control strategy independently based on deep Q-learning, and learns to cooperate with its neighbour buildings through shared rewards. Reference [20] combines a multi Q-learning based algorithm with an artificial neural network for solving a typical deep reinforcement learning (DRL)-based energy management problem in a residential house. In [25], a fully decentralized multi-agent scheme for reducing the electricity bill and user dissatisfaction through the optimal hour-ahead scheduling of home appliances and an EV load has been presented. In an energy market framework, [8] presented an energy and load management strategy based on a distributed MARL approach.

Similarly, the CTDE has been applied to solve typical problems in the area of energy management [3, 18, 28]. In [3], the energy management problem within an industrial microgrid with high renewable generation has been solved by a multi-agent Proximal Policy Optimization (PPO) approach. The objective is to coordinate the different machines to reduce current and predicted overall energy cost. In [28], a typical HVAC control problem in a multi-area commercial building is solved by an actor-critic based multi-agent DRL approach with attention mechanism. The proposed method aims to reduce the HVAC overall cost while preserving comfort and reducing Greenhouse Gas Emissions. The authors in [18] have solved a typical demand side management problem under real-time pricing by Multi-agent PPO through modeling each household as an agent.

3.2 Formulating energy management problems as Markov games

To apply MARL, the control problem is first formulated as a Markov game. Then a MARL algorithm is implemented to solve the formulated problem.

3.2.1 Markov game

A Markov game is an extension of a Markov decision process (MDP) to a multi-agent setting [19]. In particular, a Markov game of N agents is composed of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N$, and a set of observations $\mathcal{O}_1 \times \mathcal{O}_2 \times \cdots \times \mathcal{O}_N$. Each agent i uses its local policy π_i to choose an action $a_i \in \mathcal{A}_i$ based on its local observation $o_i \in \mathcal{O}_i$. When each agent takes an action, the system moves to a new global state $s' \in \mathcal{S}$ based on the transition function $\mathcal{T}\colon \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N \mapsto \mathcal{S}$. Subsequently, a reward r_i and a new observation o_i are allocated to each agent i. The reward function of agent i is such that: $\mathcal{R}_i\colon \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N \mapsto \mathbb{R}$, and the observation is correlated to the state such that: $o_i\colon \mathcal{S} \mapsto \mathcal{O}_i$. The aim of each agent i is to maximize its own total cumulative expected reward R_i defined by:

$$R_i = \sum_{k=0}^{T} \gamma^k r_{i,k} \tag{1}$$

where $r_{i,k}$ is the reward obtained by agent i during time step k, and γ is the discount factor.

It is noted that when the number of agents N is equal to 1, it is the case of an MDP, which is the modeling basis for solving single-agent RL-based problems.

3.2.2 Control problems' modeling

To formulate an energy management problem as a Markov game, each dispatchable unit is modeled as a separate agent that can take control actions based on its local observations. Then, depending on the studied system and its components, the observation space, action space and reward function of each agent are defined. Elements of Markov game modeling of each control problem vary depending on the various components that constitute the control environment. For the sake of illustration, details are provided about the formulation of problems (A) and (B) as Markov games.

Problem (A) In this problem, the objective is to find the optimal charging/discharging power for the ESS which is modeled as a separate agent. We choose the local observation of the ESS agent to be composed of its current state of charge (SOC) and the history of load consumption and PV production in the previous k time steps (k is a hyperparameter). It is noted that the use of a history of observations is a common way for dealing with partial observability. The action space of the ESS agent is the range between its maximum discharging and charging powers. The diesel generator compensates at best for any lack or surplus of capacity and energy. A penalty is then associated to the agent for any non-served load or if the diesel generator has been employed. In case of high PV production, the generation excess is curtailed. Therefore, the reward associated to each agent is simply the negative of the cost, and is defined as follows:

$$R(t) = -c_0 u_d(t) - c_1 P_{dsl}(t) - c_2 P_{dsl}^2(t) - c_{nsl} P_{nsl}(t) \tag{2}$$

where c_{nsl} is the marginal cost of non-served load while c_0, c_1 and c_2 are the components of the diesel generator operation cost and u_d is its ON/OFF status. P_{dsl} is the diesel generator power output, while P_{nsl} is the amount of non-served load.

Problem (B) The objective here is to find the optimal charging/discharging power for each of the two ESSs. In this case, we suppose that the electricity provider is a stiff source and can compensate for any lack or surplus of capacity and energy. The aim is threefold: meet the load, minimize the cost of purchased energy which is beneficial for the building, and reduce the demand fluctuations which is advantageous for the electricity provider. Therefore, the cost to be minimized can be defined by the following relationship

$$R(t) = -cP_g(t) - 2c|P_g(t) + P_g(t-1)| \tag{3}$$

where c is the operation cost and P_g is the amount of power provided by the electricity provider.

4 Proposed control approaches

Once the problem is formulated as a Markov Game as detailed in the previous section, the control algorithm is then implemented to solve the formulated problem. In this work, we design and implement RL-based control strategies based on hierarchical RL and MARL. In particular, the proposed MARL approaches are CTDE-based. In this section, more details about the principle of CTDE and its advantages are provided. Potential control strategies are then proposed for solving the formulated problems.

4.1 Centralized-training decentralized execution

Prior to describing the principle of CTDE, two elemental components of any RL algorithm are defined, namely the Q-function and V-function. The Q-function quantifies the value of taking action a under

state s at time t while following control policy π, and is defined as follows [5, 9]:

$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{k=0}^{T} \gamma^k r_{t+k} | S_t = s, a_t = a] \tag{4}$$

The V-function represents the value of a state s when following a policy π [5, 9]:

$$V^{\pi}(s) = \mathbb{E}[\sum_{k=0}^{T} \gamma^k r_{t+k} | S_0 = s] \tag{5}$$

CTDE, a common category in MARL, features various advantages when solving multi-agent problems. As any RL algorithm, it comprises a training phase in which agents are trained to generate optimal actions, and an execution phase in which each agent uses its learned control policy for taking decisions in real-time. As the name suggests, the training phase in CTDE is centralized in the sense that it uses a central unit, called critic network, that has access to information of all agents (i.e. local observation, action and reward) and guides the learning process of the agents. The use of such common information during the training phase is an implicit way for sharing information among agents, therefore addresses the non-stationarity issue and promotes a better coordination among the agents.

In this work, the focus is on actor-critic based CTDE approaches. In more detail, each agent learns two deep neural networks during the training phase:

- The actor network: is a parameterized representation of the control policy π. It is trained based on the feedback provided by a critic network in order to provide better control actions. The actor network receives a local observation as input, and outputs the optimal action.
- The critic network: is a parameterized representation of a Q or V-function which quantifies the value of an action. It is trained to learn how to generate better evaluations of the action taken in a given state. The critic network takes the global state (i.e. a grouping of all agents observations) and global action (i.e. a grouping of all agents actions), and outputs a scalar value that judges the obtained action (i.e. whether it is a good action or not).

Figure 2 illustrates the training and execution phases of an actor-critic based CTDE approach where the critic network is the parameterized representation of the Q-function.

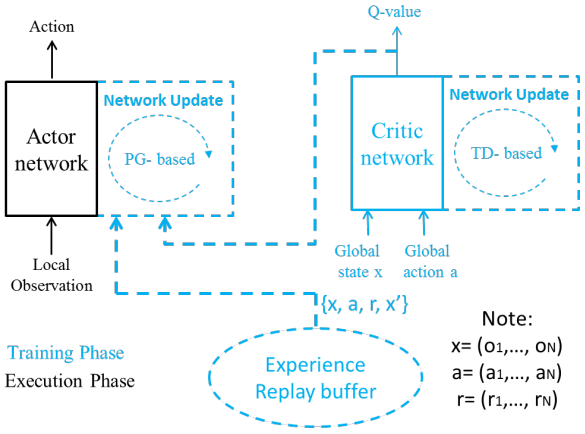

Figure 2: Actor and critic networks of one agent during training and execution phases.

In multi-agent settings, each agent usually has its own actor and critic network. Through the training process, each agent regularly updates its actor and critic networks by adjusting their weights. These updates are based on random data samples from an experience replay buffer that stores the

history of interaction of the agents with the environment (i.e. observations, actions and obtained rewards). Commonly, the critic network update is based on temporal difference (TD) while the actor network update is policy-gradient based.

At the end of the training phase, each agent has learned to generate optimal actions, and will use its learned control policy during the execution phase for taking real-time decisions. In the execution phase, the centralized critic networks are no longer needed, and each agent uses only its actor network to generate actions based on its local observations. Therefore, the execution phase is said to be decentralized because each agent only uses its local information for generating actions with no need for communication or sharing information with other agents in real-time.

Such framework is suitable for power system applications, because the emerging systems are becoming increasingly multi-dimensional and decentralized. In this case, CTDE allows to solve problems involving various interconnected decision-making units without any communication burden.

4.2 Multi-agent deep deterministic policy gradient

By reviewing a variety of MARL-based approaches, MADDPG, a multi-agent actor-critic based approach, is promoted as a primary candidate for solving the studied multi-agent problem for many reasons. MADDPG [19] is of great interest when the action space is continuous which is the case for the ESS charging/disharging power control. The use of MADDPG can coordinate effectively the actions of all participating agents without any real-time communication burden, which is made possible by the principle of CTDE. That is, common or shared information is needed only during the training phase, while the real-time execution is fully decentralized. Furthermore, MADDPG can effectively solve cooperative, competitive or mixed cooperative competitive scenarios. Finally, by a throughout review of the literature, it has been found that the application of MADDPG in power system control has been limited to frequency and voltage control. Therefore, for the first time, we intend to investigate its performance in other categories of power system decision and control problems.

MADDPG is an adaptation of the classical actor-critic method to deal with a multi-agent setting. An actor and critic model must be obtained for each agent during the training phase. In general, an actor model receives an observation and generates an action. The critic model, as its name suggests, receives the action and state as input, and generates a value that represents an evaluation of the action (whether it is a good action or not based on the generated reward). To deal with the non-stationarity in multi-agent settings, each critic network is augmented with the states and actions of all agents. However, each actor only needs its local observation as input for making decisions. During the training phase, the critic, which have more input information about the agents, acts as a guide for its corresponding actor to make it learn how to choose better actions, while the critic learns how to generate better judgements. The actor and critic networks are updated based on PG and TD-learning, respectively. It is noted that the critic networks are only needed during the training phase. During the execution phase (validation and testing), only the actor networks (which need local information only) are used to make decisions [19].

4.3 Multi-agent advantage actor critic (MA-A2C)

The second candidate method is the multi-agent advantage actor critic (MA-A2C) approach which will be applied for the first time to solve a problem in power system decision and control. The MA-A2C takes the classical A2C approach as a basis and extends it to a multi-agent setting based on the principle of CTDE. Therefore, it is expected to include the merits of the A2C approach on one hand and the CTDE from the other.

A2C, an actor-critic and single-agent based approach, is known for having a very stable learning behavior when converging to the solution, and a good bias-overfitting balance [2, 6]. This is a result of using an advantage function in which a baseline is subtracted from the expected return in order

to make the critic and actor updates more stable. However, it is mostly suitable for problems with discrete action spaces. Although A2C is a widely-used RL approach with many successful applications such as traffic light control [6], its use in the area of energy management is rare [2]. To extend A2C to a multi-agent setting, we simply follow the CTDE approach used in [19] when the DDPG approach has been extended to the MADDPG approach.

4.4 Hierarchical reinforcement learning-based control strategies

4.4.1 Hierarchical reinforcement learning

As we deal with power system problems of increasing complexity at many levels, the use of primitive RL-based approaches may not be sufficient to learn efficient control strategies. The application of Hierarchical RL, a sub-area of RL, can be a potential alternative to primitive RL as it intends to solve complicated problems by learning spatio-temporal abstract actions rather than primitive ones. In particular, the control architecture of primitive RL is composed of only one control layer, while in hierarchical RL the control hierarchy is multi-layer, and these layers can be extended in the space or time dimension. Based on temporal abstraction, the choice of the control layer that generates the action depends on the time space, while in spatial abstraction it is dependent on the state space. The concept of abstraction implies that the learned policy is composed of a set of high and low-layer control policies: A high-layer policy generates meta-actions that give instructions to a low-layer policy. The low-layer policy is reponsible for generating the primitive actions that the agent takes in its environment. There are mainly four foundational methods under hierarchical RL: Options framework, MAXQ, Feudal Reinforcement Learning, and Hierarchical Abstract Machines [14].

In this work, the focus is on the options framework; It is based on temporal abstraction in which an agent takes actions with different time scales when solving a given task. In particular, at each decision time-step, the agent chooses an option (called also skill) which is suggested by a high-layer policy (called policy-over-options) among a set of options. The agent then follows the low-layer policy (called intra-option policy) that is associated with its selected option. The low-layer policy generates the action that the agent takes in the environment. A termination function then decides whether the agent should keep the same selected option for making the next decision, or a new option should be selected. The option-critic approach has various advantages as compared with other hierarchical RL-based algorithms. It is an end-to-end approach that allows to automatically discover and learn the underlying temporal abstractions within a given task without the need for specifying sub-goals by hand as is the case in other hierarchical approaches. In the terminology of the options framework, three functions are learned: 1) The policy-over-options: chooses the option; 2) The intra-option policy: chooses the primitive action and 3) The termination function: decides whether to change the option or continue using it.

Hierarchical RL is a promising research area that has outperformed primitive RL in many applications, and has proven to be efficient in solving challenging tasks such as complex ATARI games, management of multi-agent autonomous guided vehicles, robot soccer, etc. [4, 21]. The extension of hierarchical RL applications to larger and more complex control environments is an ongoing and open research work that is receiving increasing attention. Such features motivate its application for solving problems in power system decision and control area.

4.4.2 Primitive and hierarchical advantage actor critic

The A2C approach is a primitive RL-based control strategy as it is composed of one control layer (i.e. actor network). To improve its performance, we propose for the first time a hierarchical advantage actor option critic (HA2OC) approach that combines A2C and hierarchical RL. The idea is inspired by the work presented in [13], which proposes an extension of asynchronous advantage actor critic (A3C) algorithm to the Options framework. In HA2OC, the agent learns in a similar way to the A2C algorithm, however within the options framework.

4.4.3 Multi-agent option-critic deep deterministic policy gradient (MA-OCDDPG)

Motivated by the advantages of MARL from one side and hierarchical RL from the other, we design a control strategy that merges both theories. Investigating the intersection between the two areas (i.e. hierarchical RL and MARL) is expected to yield performant control strategies that can deal effectively with extremely complex systems featuring multi-dimensional action spaces. Therefore, this work proposes a new control strategy that we call Multi-Agent Option-Critic Double Deterministic Policy Gradient (MA-OCDDPG). As the name suggests, it involves agents that learn based on the MADDPG approach however within the options framework. The aim is to obtain an improved performance, while harvesting the merits of both approaches. It is noted that the variants of the option-critic strategies available in the literature are single-agent and intended for discrete action spaces. Therefore, adjustments are introduced to the algorithm in order to extend it to multi-agent environments with continuous action spaces.

Figure 3 illustrates the control hierarchy of the proposed MA-OCDDPG approach. In particular, inspired by the principle of the option-critic approach, the MA-OCDDPG relies on creating and training in parallel a fixed number of actor networks. A "selector" (equivalent to the policy-over-options in the option-critic terminology) is also trained with the objective of choosing the "best" candidate of actor network at each time step. Each actor network relies on DDPG as a core algorithm for its learning. Then, we simply extend this framework based on CTDE to deal with multi-agent settings. It is noted that the selector receives only a local observation as input, and is trained based on the feedback provided by the critic network associated with each agent. The same is true for each actor network. In the primary version of the algorithm, we consider a particular case of the option-critic approach where the option terminates directly after taking an action, therefore a new option is selected at each decision time step. The addition of a termination function will be investigated in a future work.

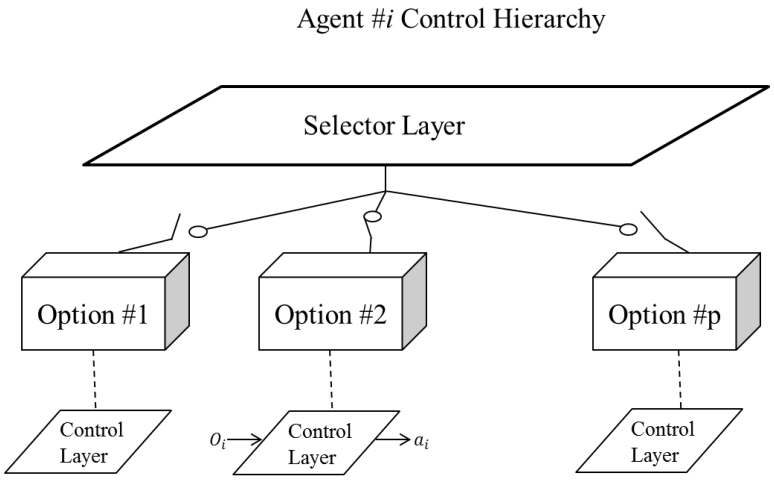

Figure 3: Control hierarchy (Execution phase) of one agent based on MA-OCDDPG (e.g. the selector layer chooses option #2 which is transmitted to the lower control layer to generate a primitive action).

5 Experimental results

5.1 Data

The problem scenarios are designed by the use of three-year data of load and PV generation with a one-hour time resolution [10]. To investigate the generalization capability of RL-based approaches, the first year data are used for training the different network models, while the second and third year data

(unseen data) are used for validation and testing, respectively. The aggregated load and PV data of the second year are shown in Figure 4.

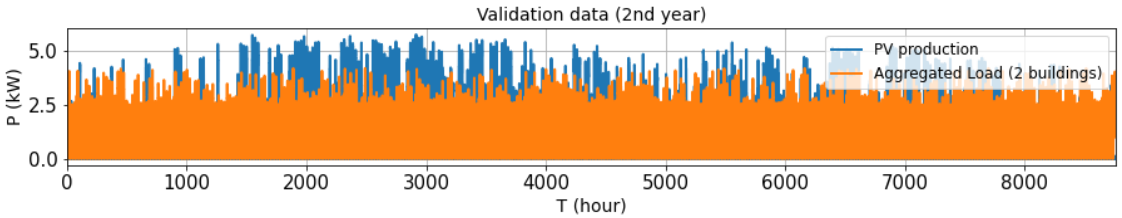

Figure 4: PV production and aggregated data of the second year (validation).

The parameters used for the experiments on Problem (A) and Problem (B) are illustrated in Table 1 and Table 2, respectively. The SOC for each battery is initialized with a value equal to 50% of its energy rating. It is noted that S_b^{max} and P_b^{max} denote the maximum energy capacity and maximum charging/discharging power of the ESS, respectively. The charging/discharging efficiencies are denoted by η_{ch} and η_{dch}. The hyperparameter k that specifies the time-window dimension of the history of observations is set to six time steps. It is noted that the choice of such particular value is a result of trying other different values, and then selecting the one that allows to reach the best compromise.

In all scenarios, the deterministic solution of the corresponding control optimization problem is calculated and serves as the benchmark for evaluating the performance of the implemented RL-based control strategies. We recall that such a *theoretical* solution, where the optimizer knows exactly how the future will unfold, is not realistic and cannot be obtained in real scenarios with multiple uncertainty sources.

Table 1: Experimental Parameters for the Environment of Problem (A).

c_0 (€)	c_1 (€/kWh)	c_2 (€/kW^{2}h)	c_{nsl} (€/kWh)	P_{dsl}^{max} (kW)	P_b^{max} (kW)	S_b^{max} (kWh)	η_{ch}	η_{dch}	k
0.0157	0.108	0.31	2	1	4	8	0.95	0.95	6

Table 2: Experimental Parameters for the Environment of Problem (B).

c (€/kWh)	P_{b1}^{max} (kW)	P_{b2}^{max} (kW)	S_{b1}^{max} (kWh)	S_{b2}^{max} (kWh)	η_{ch}	η_{dch}	k
0.132	1.1	1.1	4	4	0.95	0.95	6

5.2 Experiment 1

The objective of this experiment is to highlight the potential of hierarchical RL, and verify whether it is able to generate better actions, and enhance the overall performance of the control strategy. Therefore, we solve Problem (A) by A2C (primitive RL) and HA2OC (hierarchical RL).

Actor and critic models are obtained using the training data (i.e. first year data). For a fair comparison between A2C and HA2OC, the same actor and critic networks are used in both algorithms. In particular, the actor network is parametrized by four dense layers each with 512 units and exponential linear unit (ELU) activation function. The output layer is a dense layer with one unit and a softmax activation function. The termination network is composed of four dense layers each with 512 units and ELU activation function, and its output layer has a sigmoid activation function with a number of units equal to the number of options. The critic network has four dense layers with ELU activation, and an output layer with one unit and a linear activation. In this simulation, the number of options is set to three.

The use of training, validation and testing data aims to verify whether the proposed method is able to generalize well and perform acceptable results with unseen data.

The test and validation phases are performed with no exploration by calculating the argmax of the Q-function. However, as in [13], we keep an exploration rate of 10% for the options' selection.

The learning simulations are repeated 10 times with a total number of epochs equal to 200, and the model that generates the best validation and testing scores is saved. It is noted that the validation and testing are performed after each episode and that one epoch corresponds to one year, while each step corresponds to one hour [10].

Figure 5 shows the learning curves for primitive and hierarchical RL, while Table 3 compares the solution of hierarchical RL with the solutions of A2C and the deterministic optimization problem. We recall that A2C requires the discretization of the action space, while in the deterministic problem a continuous action space is used which will already make a difference between the RL-based solution and the deterministic solution. In particular, the discretized action space of the ESS is $\{-2.2, -1.4, -1.0, -0.4, 0.0, 0.4, 1.0, 1.4, 2.2\}$ all in kW. It is noted that the choice of the discretization step, and therefore the action-space size is delicate as a fine discretization usually results in a curse of dimensionality and also a slower training, while an unrefined discretization results in sub-optimal solutions.

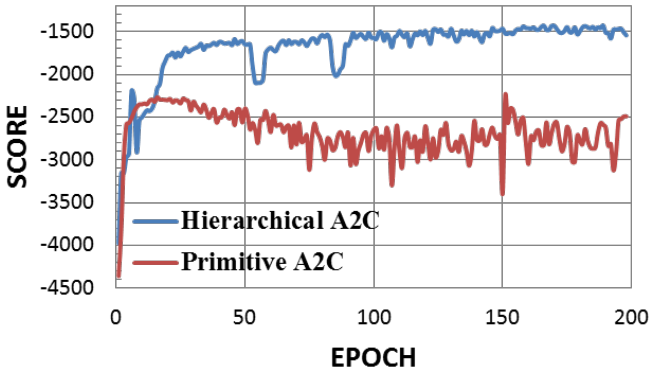

Figure 5: Problem (A): Learning curves for primitive A2C and HA2OC.

From Figure 5, it is clear that HA2OC allows to obtain higher score, faster convergence and more stable learning as compared to primitive A2C.

Table 3: Problem (A): Total annual cost of the test and validation years.

	Deterministic	A2C	HA2OC
Validation year (€)	719.33	1374.95	1150.05
Test year (€)	804.02	1458.51	1218.74

Furthermore, Table 3 shows that the solution of hierachical A2C is reasonable as it is close to the deterministic solution by 68% and 70% for the validation and test years, respectively), and it is able to generalize well on unseen data (validation and testing). It is noted that the action space is naturally continuous, therefore a discretization has been performed which always results in sub-optimal solutions. This encourages the use of policy-gradient based approaches as it allows to deal with such limitation.

5.3 Experiment 2

The objective of this experiment is to solve Problem (B) by different MARL algorithms in order to compare their performances and highlight the potential of hierarchical RL. In particular, three different control strategies are implemented 1) MADDPG, 2) MA-OCDDPG and 3) MA-A2C.

Actor and critic models for each approach are learned based on training data. In particular, the actor network of each agent in MADDPG and MA-OCDDPG is parametrized by 2 dense layers each with 64 units and Rectified Linear Unit (ReLu) activation function. The output layer is a dense layer with one unit and tanh activation function. In MA-A2C, the actor network is composed of two dense layers each with 64 units and ReLu activation function followed by an output dense layer with one unit and softmax activation function. In all approaches, the critic network has two dense layers with ReLu activation and an output layer with one unit and no activation function.

The first-year data are used for training, while the second-year and third-year data are used for validation and testing, respectively. The ϵ-greedy policy is used for random exploration of actions. The ϵ value starts with a value equal to 1 and gradually decreases over the epochs until it reaches 0.1 in order to reach a trade-off between exploration and exploitation. In MADDPG and MA-OCDDPG, the exploration is based on sampling from a uniform distribution, while in MA-A2C it is based on random selection of integers. The test and validation phases are performed with no exploration (i.e. $\epsilon = 0$). The simulations are repeated 15 times with a number of epochs equal to 100 and the model that generates the best validation and testing scores is saved.

Due to space limit, we present only some of the simulation results that correspond to the test year (Figure 8, Figure 9 and Figure 10). It can be seen that, given the same period (i.e. eight days from July), the behavior of the control decisions vary depending on the generating control algorithm. In particular, it can be noticed that the control decisions of hierarchical RL are more flexible and adaptive to load and PV variations than primitive RL.

Figure 6 plots the peak to average ratio (PAR) which is defined in [22], and is used as an indicator to evaluate the fluctuations and peaks reduction performance.

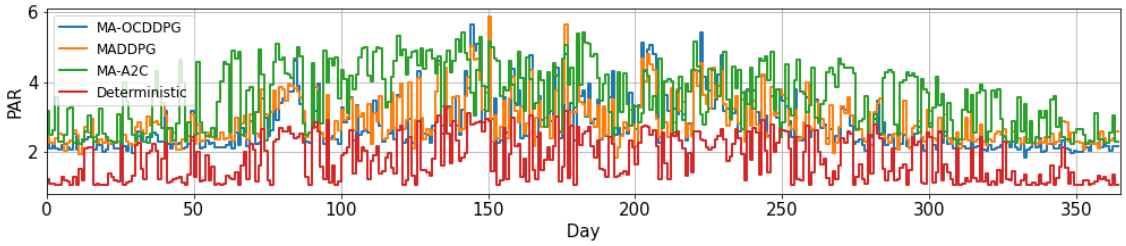

Figure 6: Problem (B): Variation over the test year of the daily PAR obtained with the deterministic solution, MADDPG, MA-OCDDPG, and MA-A2C.

From Figure 6, it is clear that the MADDPG and MA-OCDDPG approaches are reducing the PAR more effectively than the MA-A2C approach.

Table 4 presents the annual total cost for the test and validation years obtained with MADDPG, MA-OCDDPG, and MA-A2C approaches. In this scenario the annual total cost consists of two components: the annual energy cost (i.e. energy purchased by the two buildings) and the annual fluctuations' cost. The solution provided by the MADDPG and MA-OCDDPG approaches are reasonably close to the theoretical full-foresight deterministic solution, while further work and adjustments should be introduced to the MA-A2C approach to improve its performance. However, we highlight again that the "theoretical" solution is not realistic and cannot be obtained in real scenarios with high uncertainties. Overall, MA-OCDDPG, which combines hierarchical RL with MARL, has allowed to reach the best trade-off between the objective of minimizing the purchased energy cost and power fluctuations. This

trade-off is achieved by means of the options' integration into the framework which allows to take temporally extended actions. Figure 7 illustrates how each agent changes its choice of option in order to choose the best actor network (i.e. intra-option policy) whenever a decision is to be made.

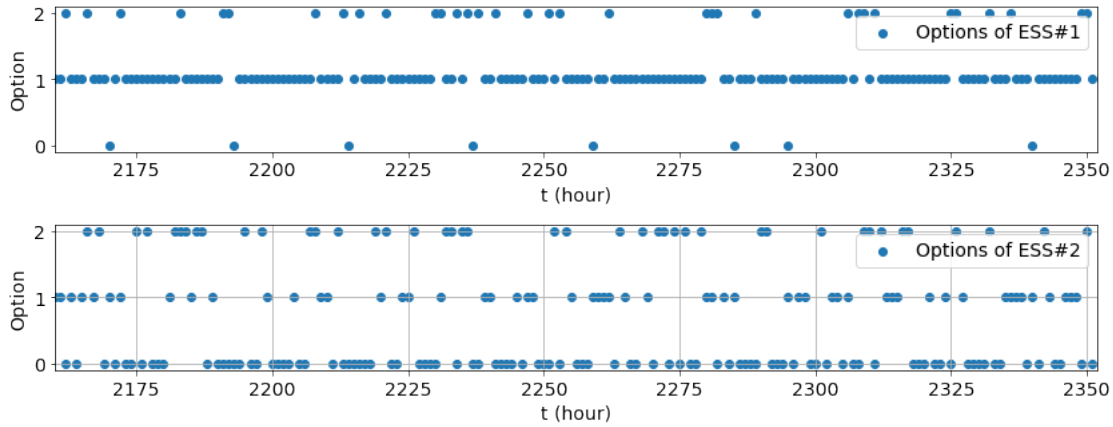

Figure 7: Problem (B): Options' choice (y-axis: 0 for option $\#1$, 1 for option $\#2$ and 2 for option $\#3$) of each agent obtained with MA-OCDDPG approach for eight days from January (time resolution of one hour).

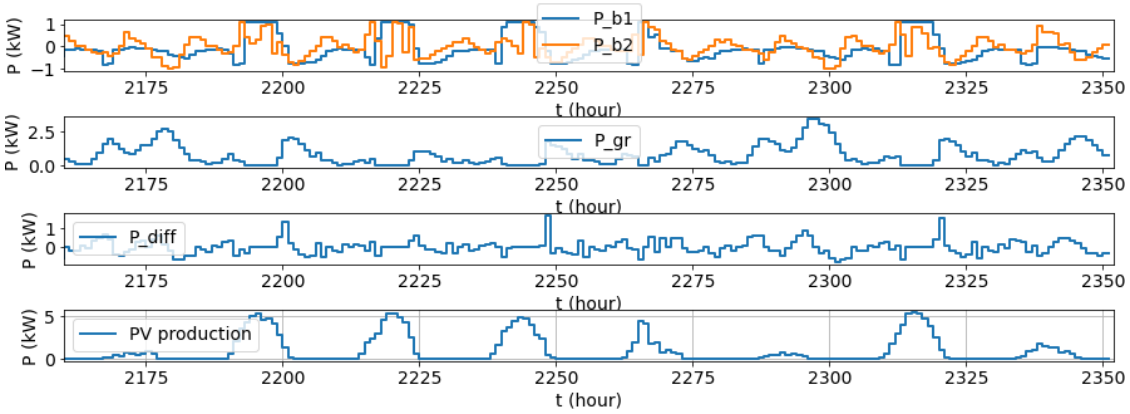

Figure 8: Problem (B): Control decisions obtained with MA-OCDDPG for one week from July of the second year (P_{b_1} and P_{b_2}: Amount of power to be charged/discharged from ESS $\#1$ and ESS $\#2$, respectively, P_{gr}: Amount of electricity delivered by the electricity provider, P_{diff}: Difference in P_g between two consecutive time steps).

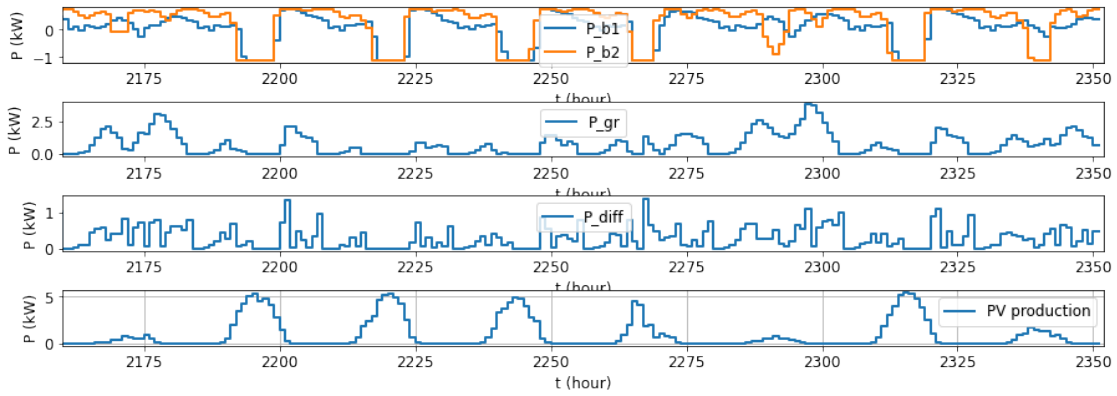

Figure 9: Problem (B): Control decisions obtained with MADDPG for one week from July of the second year.

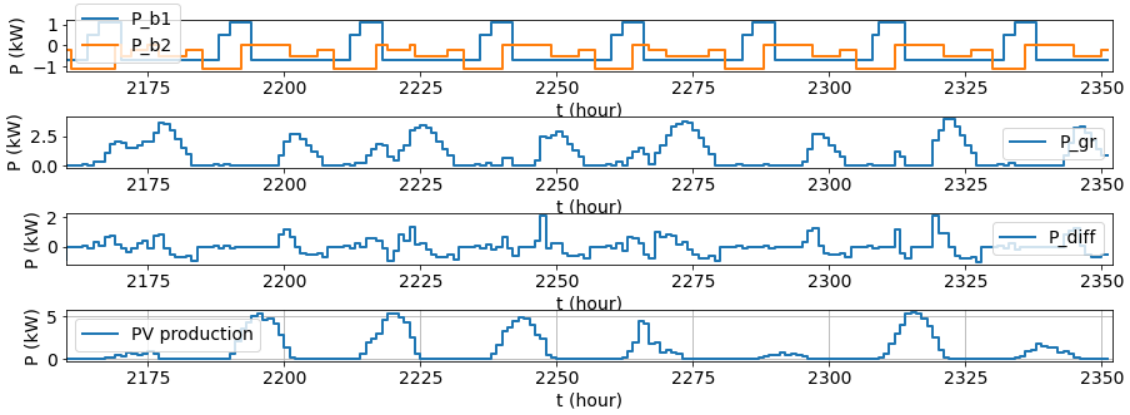

Figure 10: Problem (B): Control decisions obtained with MA-A2C for one week from July of the second year.

Table 4: Total annual cost of the test and validation years obtained by the deterministic, MADDPG, MA-OCDDPG, and MA-A2C approaches

	Deterministic	MA-A2C	MADDPG	MA-OCDDPG
Validation year (€)	1281.25	1928.32	1718.79	1687.77
Test year (€)	1363.36	2012.46	1818.30	1773.15

6 Conclusion

This work proposed control approaches for energy management through scheduling the operation of controllable DERs. We have presented control strategies that can efficiently handle complex and multi-objective power system control problems with uncertainties and high dimensional state and action spaces. It has been found that the implementation of hierarchical RL in general and the options framework in particular for solving a typical energy scheduling problem within a microgrid results in lower costs, better generalization, faster convergence and enhanced stability as compared with its primitive counterpart. Furthermore, three MARL-based control strategies have been implemented for solving a multi-agent energy management problem within an eco-neighborhood, namely MADDPG, MA-A2C and MA-OCDDPG. The proposed approaches are based on the principle of CTDE which allows to achieve coordination among different decision-making units with no communication burden. In particular, MA-A2C is only applicable with discretized action spaces, while MADDPG and MA-OCDDPG can be used with continuous action spaces. Simulation results have shown that MA-OCDDPG which merges hierarchical RL with MARL theory have resulted in the best performance when compared with the other control approaches, while at the same time comprising the merits of both sub-areas. Although this work has considered fully cooperative scenarios, all the proposed control approaches are also applicable for solving other scenarios with fully competitive or mixed cooperative competitive frameworks which is a future research direction.

References

[1] Mehdi Ahrarinouri, Mohammad Rastegar, and Ali Reza Seifi. Multiagent reinforcement learning for energy management in residential buildings. IEEE Transactions on Industrial Informatics, 17(1):659–666, 2020.

[2] Erick O Arwa and Komla A Folly. Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review. IEEE Access, 8:208992–209007, 2020.

[3] J. Bakakeu, D. Kisskalt, J. Franke, S. Baer, H. H. Klos, and J. Peschke. Multi-agent reinforcement learning for the energy optimization of cyber-physical production systems. In 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pages 1–8, 2020.

[4] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. Discrete event dynamic systems, 13(1):41–77, 2003.

[5] Di Cao, Weihao Hu, Junbo Zhao, Guozhou Zhang, Bin Zhang, Zhou Liu, Zhe Chen, and Frede Blaabjerg. Reinforcement learning and its applications in modern power and energy systems: A review. Journal of Modern Power Systems and Clean Energy, 8(6):1029–1042, 2020.

[6] T. Chu, J. Wang, L. Codecà, and Z. Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 21(3):1086–1095, 2020.

[7] Zihan Ding and Hao Dong. Challenges of Reinforcement Learning, pages 249–272. Springer Singapore, Singapore, 2020.

[8] Elham Foruzan, Leen-Kiat Soh, and Sohrab Asgarpoor. Reinforcement learning approach for optimal distributed energy management in a microgrid. IEEE Transactions on Power Systems, 33(5):5749–5758, 2018.

[9] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. Found. Trends Mach. Learn., 11(3–4):219–354, December 2018.

[10] Vincent François-Lavet, David Taralla, Damien Ernst, and Raphaël Fonteneau. Deep reinforcement learning solutions for energy microgrids management. In European Workshop on Reinforcement Learning (EWRL 2016), 2016.

[11] Mevludin Glavic. (deep) reinforcement learning for electric power system control and related problems: A short review and perspectives. Annual Reviews in Control, 48:22–35, 2019.

[12] Mevludin Glavic, Raphaël Fonteneau, and Damien Ernst. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. IFAC-PapersOnLine, 50(1):6918–6927, 2017.

[13] Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. 2018.

[14] Yanhua Huang. Hierarchical Reinforcement Learning. Springer Singapore, Singapore, 2020.

[15] Ying Ji, Jianhui Wang, Jiacan Xu, Xiaoke Fang, and Huaguang Zhang. Real-time energy management of a microgrid using deep reinforcement learning. Energies, 12(12):2291, 2019.

[16] Bingnan Jiang and Yunsi Fei. Smart home in smart microgrid: A cost-effective energy ecosystem with intelligent hierarchical agents. IEEE Transactions on Smart Grid, 6(1):3–13, 2014.

[17] Isaac Kargar. Centralised training and decentralised execution in multi-agent reinforcement learning.

[18] J. Lee, W. Wang, and D. Niyato. Demand-side scheduling based on multi-agent deep actor-critic learning for smart grids. In 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pages 1–6, 2020.

[19] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[20] Renzhi Lu, Seung Ho Hong, and Mengmeng Yu. Demand response for home energy management using reinforcement learning and artificial neural network. IEEE Transactions on Smart Grid, 10(6):6629–6639, 2019.

[21] Ofir Nachum, Haoran Tang, Xingyu Lu, Shixiang Gu, Honglak Lee, and Sergey Levine. Why does hierarchy (sometimes) work so well in reinforcement learning? CoRR, abs/1909.10618, 2019.

[22] Hung Khanh Nguyen, Ju Bin Song, and Zhu Han. Demand side management to reduce peak-to-average ratio using game theory in smart grid. In 2012 Proceedings IEEE INFOCOM Workshops, pages 91–96, 2012.

[23] ATD Perera and Parameswaran Kamalaruban. Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews, 137:110618.

[24] A. Prasad and I. Dusparic. Multi-agent deep reinforcement learning for zero energy communities. In 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), pages 1–5, 2019.

[25] Xu Xu, Youwei Jia, Yan Xu, Zhao Xu, Songjian Chai, and Chun Sing Lai. A multi-agent reinforcement learning-based data-driven method for home energy management. IEEE Transactions on Smart Grid, 11(4):3201–3211, 2020.

[26] Ting Yang, Liyuan Zhao, Wei Li, and Albert Y Zomaya. Reinforcement learning in sustainable energy and electric systems: A survey. Annual Reviews in Control, 2020.

[27] Liang Yu, Shuqi Qin, Meng Zhang, Chao Shen, Tao Jiang, and Xiaohong Guan. Deep reinforcement learning for smart building energy management: A survey. arXiv preprint arXiv:2008.05074, 2020.

[28] Liang Yu, Yi Sun, Zhanbo Xu, Chao Shen, Dong Yue, Tao Jiang, and Xiaohong Guan. Multi-agent deep reinforcement learning for hvac control in commercial buildings. IEEE Transactions on Smart Grid, 12(1):407–419, 2020.