Deep Reinforcement Learning for Optimal Energy Management of Multi-energy Smart Grids

Dhekra Bousnina, Gilles Guerassimoff

Deep Reinforcement Learning for optimal energy management of multi-energy smart grids*

Dhekra Bousnina[1] and Gilles Guerassimoff[1]

MINES ParisTech, PSL Research University, CMA - Centre de Mathématiques Appliquées, Sophia Antipolis, France.

Abstract. This paper proposes a Deep Reinforcement Learning approach for optimally managing multi-energy systems in smart grids. The optimal control problem of the production and storage units within the smart grid is formulated as a Partially Observable Markov Decision Process (POMDP), and is solved using an actor-critic Deep Reinforcement Learning algorithm. The framework is tested on a novel multi-energy residential microgrid model that encompasses electrical, heating and cooling storage as well as thermal production systems and renewable energy generation. One of the main challenges faced when dealing with real-time optimal control of such multi-energy systems is the need to take multiple continuous actions simultaneously. The proposed Deep Deterministic Policy Gradient (DDPG) agent has shown to handle well the continuous state and action spaces and learned to simultaneously take multiple actions on the production and storage systems that allow to jointly optimize the electrical, heating and cooling usages within the smart grid. This allows the approach to be applied for the real-time optimal energy management of larger scale multi-energy Smart Grids like eco-distrits and smart cities where multiple continuous actions need to be taken simultaneously.

Keywords: Deep Reinforcement Learning · Actor-Critic · energy management · smart grids · multi-energy.

1 Introduction

1.1 Context of the problem

Within the radical changes that the energy landscape is currently undergoing, Smart Grids are playing a major role in the modernization of the electric grid [5]. These smart electricity networks have the great advantage of integrating in a cost-effective way the behavior and actions of all the users connected to it, including consumers, producers and prosumers, to ensure a cost-efficient and sustainable operation of the power system while guaranteeing quality and security of supply [36]. Besides electrical networks, district heating and cooling

* Supported by the program Investissement d'Avenir, operated by l'Agence de l'Environnement et de la Maitrise de l'Energie ADEME, France

systems also play a paramount role in the implementation of the new smart energy systems [39]. In fact, the concept of smart thermal grids also comes up with numerous advantages including flexibility potentials and ability to adapt to the changes that affect the thermal demand and supply in short, medium and long terms. Thus, Smart Thermal Grids, as well, are expected to be an integrated part of the future energy system [4, 33]. However, research works on the optimal control and energy management within the smart grid context traditionally focus solely on the electrical usages. Though, jointly optimizing the electrical networks together with other energy vectors interacting with them like heating and cooling networks has a great potential to increase the overall economic and environmental efficiency and flexibility of the energy systems. This idea brings about a generalization of the Smart Grid concept to Smart Multi Energy Grids [22] that lies on the interaction between electricity and other energy sectors (like heating, cooling, gas and hydrogen) as well as other sectors that electricity might interact with like the transport sector. Considering all these interactions in the optimal management of energy systems allows to unlock considerable efficiency and flexibility potentials and represents one of the main advantages of Smart Multi Energy Grids.

Optimal control of smart (multi-energy) grids is essential to guarantee a reliable operation for the smart grid components and ensure an optimal management of controllable loads, production units and storage systems while minimizing energy and operational costs [21]. One of the most popular and widely used optimal control techniques is Model Predictive Control (MPC), also referred to as Receding Horizon Control [10, 25]. MPC is a feedback control method where the optimal control problem is solved at each time step to determine a sequence of control actions over a fixed time horizon. Only the first control actions of this sequence are then applied on the system and the resulting system state is measured. At the next time step, the time horizon is moved one step forward and a new optimization problem is then solved, taking into account the new system state and updated forecasts of future quantities. This receding time horizon and periodic adjustment of the control actions make the MPC robust against the uncertainties inherent to the model and forecasts [11]. MPC has been used in many successful applications in the field of Microgrid/ Smart Grid energy management including [1,26,27,38]. Nevertheless, MPC and model-based approaches in general, rely on the development of accurate models and predictors and on the usage of appropriate solvers. This does not only require domain expertise but also needs to re-design these components each time that a change occurs on the architecture or scale of the Smart Grid [12]. Furthermore, classical optimization approaches based on Mixed Integer Linear Programming (MILP), Dynamic Programming (DP) or heuristic methods like Particle Swarm Optimization (PSO) generally suffer from time-consuming procedures. In fact, they have to compute all or part of possible solutions in order to choose the optimal one, and have to re-run a generally time-consuming optimization procedure each time that an optimal decision needs be taken. Therefore, such methods, despite their ability to provide quite accurate results, generally fail to consider on-line solutions for

large-scale real data-bases [32].
Learning-based techniques, on the other hand, do not need accurate system models and uncertainty predictors and can, thus, be an alternative to model-based approaches. Reinforcement Learning (RL) [34] has been gaining popularity over the past few years when it comes to dealing with challenging sequential decision making tasks [6]. Nevertheless, RL-based approaches fail to handle large state and actions spaces owing to the curse of dimensionality [41]. This major limitation of RL can be overcome by Deep Reinforcement Learning (DRL) which is a state-of-the art Machine Learning (ML) technique evolving through the interface between RL and Deep Learning (DL) [23]. In other words, it combines the strong nonlinear perceptual capability of deep neural networks (DNNs) with the robust decision making ability of RL [7]. Unlike RL, it therefore exhibits strong generalization capabilities in problems with complex state spaces. One of the main advantages of DRL compared to other classical optimization approaches is that, once it learned an optimal strategy, it can take optimal decisions in a few milliseconds without having to re-compute any costly optimization procedure. This makes DRL algorithms less time-consuming than classical optimization approaches and makes them, as a consequence more suitable for real-time optimization problems. DRL has, this way, shown successful applications in various real-life problems with large state spaces like Atari and Go games [30], robotics [2, 37], autonomous driving [16, 29] and other complex control tasks [23]. More recently, [3] proposed a novel assembling methodology of Q-learning agents trained several times with the same training data for stock market forecasting. The use of DQN aimed at avoiding problems that may occur when using supervised learning-based classifiers like over-fitting. Other recent successful applications of DRL include intrusion detection systems as presented in [20]. Furthermore, [19] proposed a new ensemble DRL model for predicting wind speed and the comparison of the proposed model with nineteen alternative mainstream forecasting models showed that the DRL-based approach provided the best accuracy. Moreover, Google has announced in 2018 that it gave control over the cooling of several of its data centers to a DRL algorithm [13].

1.2 Deep reinforcement Learning in Smart Grids: related work

When it comes to the energy field, there have recently been several successful applications of DRL for instance in the context of microgrids, smart homes and Smart Grids, mainly for the development of cost optimization and energy management strategies. For example, [8] considers an electricity microgrid featuring PV generation, a Battery Energy Storage System (BESS) and a hydrogen storage, and adresses the problem of optimally operating these storage systems using a Deep Q-Learning (DQL) architecture. The developed Deep Q-Network (DQN) agent was tested on the case of a residential customer microgrid located in Belgium and showed to successfully extract knowledge from the past PV production and electricity consumption time series. However, it only takes discrete actions for the hydrogen storage (whether to charge at maximum rate, discharge at maximum rate or stay idle). The operation of the BESS, on the other hand, is not a

direct action of the DRL agent but is rather dynamically adapted based on the balance equation of the microgrid. Similarly, [12] proposed a DQN approach to develop real-time generation schedules for a microgrid while optimizing its daily operational costs. DQL algorithms have also been applied in [28] for the coordinated operation of wind farms and energy storage and in [18] for the on-line optimization of a microgrid featuring PV and wind generation, diesel generators, fuel cells, electric load and a BESS. Among the various DRL algorithms, the conventional DQL remains the most widely used approach and algorithms such as Policy Gradient (PG) and Actor-Critic (AC) are rarely investigated. This is primarily due to the simplicity of the DQL and to the fact that it handles well discrete action spaces. Meanwhile, DQL can not be directly applied to problems with continuous action spaces since they need to discretize the action space which leads to an explosion of the number of actions and, as a consequence, to a decreased performance [9, 17]. Indeed, considering only discrete actions for the planning and control of the Smart Grid components significantly restrains their flexibility potentials and prevents from obtaining the best optimal scheduling and control strategies. Unlike DQL, Deep Policy Gradient (DPG) algorithms are capable of dealing with environments with continuous actions spaces. In this respect, [24] proposed the use of DQL and DPG for online building energy optimization through the scheduling of electricity consuming devices. The results showed that DPG algorithms are more suitable than DQN to perform online energy resources scheduling. Even though this work pioneered the use of DRL for online building energy optimization, the actions it considers are restricted to the on/off status of flexible load devices in a smart building. Besides, the DPG algorithms are also often criticized for their low sampling efficiency as well as the fact that their gradient estimator may have a large variance, which is likely to lead to slow convergence [14]. In order to overcome this limitation, Actor-Critic (AC) algorithms were proposed to combine the strong points of DPG and DQL approaches by estimating both the policy and the Q-value function during the training. In this respect, two DRL algorithms were designed for Smart Grid optimization in [32]: on the one hand, DQL was applied for the discrete action control tasks like charging/discharging the BESS or switching the buy/sell modes of the grid. On the other hand, an AC algorithm named H-DDPG (Hybrid-Deep Deterministic Policy Gradient) was developed to deal with continuous state and action spaces. Yet, only the results of the DQN approach were presented in the paper and benchmarked with the results of a Mixed Integer Linear Programming (MILP) optimization Matlab tool. Even though DDPG algorithms were proposed for some applications in the energy systems context namely for dealing with cost optimization problems in Smart Home energy systems in [40], for flow rate control in Smart District Heating Systems (SDHS) in [42], and for solving the Nash Equilibrium in energy sharing mechanisms in [15], most of these applications consider mono-action and/or mono-fluid use-cases. In other words, they consider solely electrical or thermal Smart grids and do not consider jointly optimizing the uses of several energy vectors within a multi-energy Smart Grid. Besides, most of the previous works consider applications on the

Smart Home or building level and do not consider testing these approaches on a larger smart district-level. Finally, thourough comparisons of the performance of DDPG-based approaches with other widely used techniques like MPC for dealing with energy management systems in Smart Grids have rarely been reported in the literature.
In the present work, we propose a DDPG-based approach to deal with the real-time energy management of multi-energy Smart Grids. More specifically, we formulated the optimal control problem as a POMDP and developed a DDPG agent to perform real-time scheduling of the multi-energy systems within a Smart Grid. The main contributions of the present work are the following:

- Unlike most of previous works where mono-fluid (electrical or thermal) Smart Grids are considered, we focus on multi-energy (electrical, heating, cooling, hydrogen) smart grids that interact with the main utility grid. A variable electricity price signal is considered and a DRL-based energy management system is developed to take price-responsive control actions.
- The DDPG algorithm is proposed instead of the mostly used DQN to deal with the continuous action and state spaces inherent to the multi-energy smart grid model. At each time step of the control horizon, multiple continuous actions are simultaneously taken by the DDPG agent to optimally schedule the various storage systems as well as the thermal production units.
- The proposed approach is tested on a residential multi-energy smart grid model and will be applied on a real-life district-level multi-energy smart grid which is being currently under construction in France. More specifically, the developed DDPG agent is aimed at operating real-time energy management of the various energy systems within an eco-district: BESS, heating and cooling storage systems, controllable loads of the buildings, heated water storage tanks, as well as District Heating and Cooling production units, Electric Vehicle Charging Stations and the public lighting of the district.
- The proposed approach is benchmarked with an MPC-based approach.

The remainder of this paper is organized as follows: in section 2, the considered multi-energy smart grid model is described together with the optimal energy management problem addressed in this work. In section 3, the problem is formulated as an MDP and a DRL-based approach is proposed to solve it. Section 4 presents the simulations and results and finally conclusions and future work are asserted in section 5.

2 The multi-energy smart grid model and optimal control mechanism

The Smart multi-energy grid model considered in this paper is shown in figure 1. It is composed of residential electric, heating and cooling loads, distributed energy generators (PV panels), heating and cooling production units consisting of geothermal Thermo-Refrigerating Heat Pumps (TRHPs), a BESS, a heat storage system (by phase-change materials) and a cold storage system (by ice storage

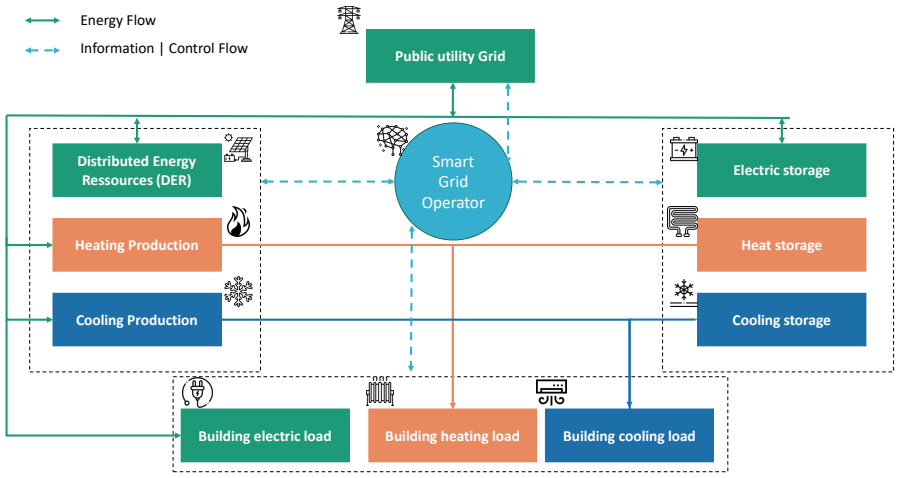

Fig. 1: Architecture of the multi-energy Smart Grid

tanks). The Smart Grid components are related to the main utility grid. In fact, besides the residential electrical usages, the TRHPs also consume electric power to produce heat and cold for the thermal needs of the buildings. At each time step, the electric loads of the buildings are met by the local PV generation, by discharging the BESS or by withdrawing electricity from the public utility grid. Thermal needs in terms of heating, on the other hand, are met whether by directly producing heat via thermo-refrigerating heat pumps or by discharging the heat storage system. Similarly, cooling loads are ensured by directly producing cold via TRHPs or by discharging the cooling storage system.
In order to jointly optimize the operation of the multi-energy systems of the Smart Grid, an energy management system is needed to schedule the different controllable units while minimizing the daily operational costs. To solve this sequential decision making problem, we formulate it as a Markov Decision Process (MDP). In fact, the energy level of each energy storage system, at each time step, depends only on the current energy level, together with the current charge/discharge power, and as a consequence, the scheduling of the different energy storage systems and production units can be formulated as an MDP $M = (S, A, T, R, \gamma)$ where its key components, the state space S, the action space A, the reward R and the transition function T are designed as follows:

- State: the environment state at each time step $t \in H$ is denoted by s_t and is composed of six types of information:
  $s_t = (s_t^{Storage}, s_t^{Load}, s_t^{DER}, s_t^{Grid}, s_t^{Prod}, s_t^{Temp})$ where $s_t^{Storage} \in S^{Storage}$ denotes the storage operation of the Smart Grid and describes the amount of energy stored in each of the battery, hydrogen, heating and cooling storage systems $s_t^{Storage} = (s_t^{Bat}, s_t^{H2}, s_t^{HS}, s_t^{CS})$, $s_t^{Load} \in S^{Load}$ contains the

h past realizations of the electric, heating and cooling loads, where h, the history length is set as 24, so that the history length covers one day of past realizations with time steps $\Delta t = 1$ hour. Similarly, $s_t^{DER} \in S^{DER}$ contains the h past realizations of PV generation, $s_t^{Grid} \in S^{Grid}$ contains the h past realizations of the electricity prices as well as the amount of power withdrawn from the main utility grid at time step t, $s_t^{Prod} \in S^{Prod}$ contains the quantities of heat and cold produced by the TRHPs at time step t. Finally, $s_t^{Temp} \in S^{Temp}$ contains both the indoor and outdoor temperatures.

- Action: the aim of the energy management system is to decide the charging/discharging power of each energy storage system P^{SS}, the amount of energy to be purchased from the public utility grid P^{Grid} and the thermal energy (heat or cold) produced by the TRHPs Q^{TRHP}.
- Reward: when an action $a_t \in A_t$ is applied on the system, this triggers the environment to move from state s_{t-1} to state s_t and hence a reward r_t is obtained. Since the aim of the agent is to minimize the total energy costs within the Smart Grid, the reward signal r_t corresponds to the negative of rescaled instantaneous operational revenues at time step t:

$$r_t = -\alpha.[C_{gen}.P_{gen}(t) + C_{grid}(t).P_{grid}(t)] \tag{1}$$

Where C_{gen} is the cost of distributed power generation and $C_{grid}(t)$ is the cost of power purchase from the public utility grid ie the variable energy price, and α is a factor by which we rescale the cost function, such that

$$0 < \alpha \leq 1 \tag{2}$$

3 The proposed Deep Reinforcement Learning-based approach

RL is an Artificial Intelligence (AI) paradigm where the AI agent interacts with its environment by taking actions over a sequence of time steps in order to maximize a cumulative reward signal [34]. At each time step t, the agent performs a control action a_t based on the measure of the current state of the environment s_t and receives, in return, a reward r_t and information on the new state of the environment s_{t+1} for the next time step $t+1$. This way, the RL agent learns an optimal control policy through the interaction with the environment as shown in figure 2.a.

DRL [23] is a family of methods which evolve through the interface between RL and DL. Such a combination of RL and DL has recently shown its ability to learn complex tasks directly from high-dimensional inputs. DRL methods are divided into two main types, namely value-based and policy-based methods. In value-based methods, the neural network learns the optimal Q-function $Q^*(s, a)$ of each action a given a state s, which is the maximum sum of rewards r_t discounted by a factor γ at each time step t achievable by a policy $\pi = P(a_t|s_t)$

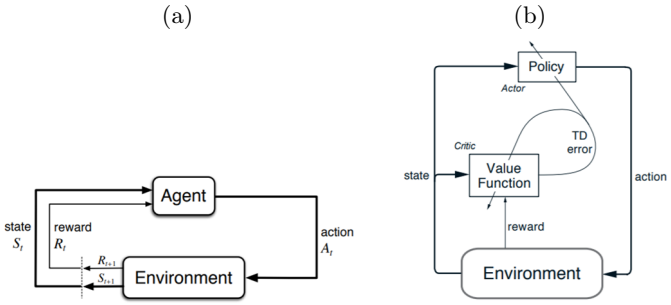

Fig. 2: (a) : The agent-environment interaction in Reinforcement Learning; (b) : The actor-critic architecture, from [34].

after taking and action a_t given a state s_t:

$$Q^*(s,a) = \max_{\pi} E[r_t + \gamma.r_{t+1} + \gamma^2.r_{t+2} + ...|s_t = s, a_t = a, \pi] \tag{3}$$

Meanwhile, for policy-gradient methods, the artificial neural network learns a probability distribution of the action a at a given state s instead of computing the Q-function. Value-based methods are known to be suitable for discrete action spaces whereas policy-based algorithms handle well continuous actions spaces.
This work proposes an application of the DDPG (Deep Deterministic Policy Gradient) algorithm which is a policy-based algorithm belonging to the actor-critic (AC) family [31]. AC methods rely on the idea of combining DPG and DQN: the policy function $\mu(s, \theta^\mu)$ is referred to as the actor where θ^μ represent the weights of the actor network. It specifies the current policy by deterministically mapping states to a specific action. The value-function $Q(s, a)$ is known as the critic and produces an error signal given the state, the output of the actor and the resultant reward signal as shown in figure 2.b [17].
When the agent takes an action a_t, under a given state s_t, according to a policy $\mu(s, \theta^\mu)$, the value of reward is given by the Bellman equation [35]:

$$Q^\mu(s_t, a_t) = E^\mu[r_t + \gamma.Q^\mu(s_{t+1}, \mu(s_{t+1}, \theta^\mu))] \tag{4}$$

The Q-network loss function is then given by:

$$L(\theta^Q) = E^\mu[(Q(s_t, a_t|\theta^Q) - y_t)^2] \tag{5}$$

where

$$y_t = r_t + \gamma.Q(s_{t+1}, \mu(s_{t+1}|\theta^Q)) \tag{6}$$

The performance objective which measures the performance of the policy μ is given by:

$$J_\beta(\mu) = \int_S \rho^\beta(s) Q^\mu(s, \mu(s)) ds \tag{7}$$

Where ρ^β is the probability-distribution function of s_t. The aim of the training process is to maximize performance objective $J_\beta(\mu)$ while minimizing the loss

function $L(\theta^Q)$. The training process of the used DDPG algorithm implemented in this work is given by Algorithm1, also described in [15].

Algorithm 1: DDPG algorithm

Initialize the actor network μ and the critic network Q with random weights θ^μ and θ^Q ;
Initialize target network μ' and Q' with the weights $\theta^{\mu'} \leftarrow \theta^\mu$ and $\theta^{Q'} \leftarrow \theta^Q$;
Initialize the experience replay Buffer B ;
for $episode \leftarrow 0$ **to** $N_{episodes}$ **do**
  Initialize a random process R for action exploration;
  Get initial observation of state S_1 at time step $t = 1$;
  for $T \leftarrow 1$ **to** N_{steps} **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + R_t$ according to the current policy and exploration noise ;
    Execute action a_t in the environment and observe the resulting reward r_t and the new state s_{t+1} ;
    Store the transition (s_t, a_t, r_t, s_{t+1}) in experience replay buffer B;
    Sample a random mini-batch of N transitions (s_i, a_i, r_i, s_{i+1}) from B ;
    Set $y_i(r_i, s_{i+1}) = r_i + \gamma.Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$;
    Update the critic by minimising the loss $L = 1/N \sum_i Q(s_i, a_i|\theta^Q) - y_i)^2$;
    Update the actor policy using the policy gradient $\nabla_{\theta^\mu} 1/N \sum_{s \in B} Q(s, \mu(s|\theta^\mu)|\theta^Q)$;
    Update the target networks: $\theta^{Q'} \leftarrow (1 - \rho).\theta^Q + \rho.\theta^{Q'}$ and $\theta^{\mu'} \leftarrow (1 - \rho).\theta^\mu + \rho.\theta^{\mu'}$
  end
end

This algorithm was integrated in a specifically-designed multi-energy Smart Grid energy management framework where the DDPG agent interacts with the Smart Grid environment to generate an optimal schedule of its various energy systems. The Smart Grid environment describes the dynamics of the energy

systems within the Smart Grid and is modeled as follows:

$$\min \quad \sum_{t=1}^{H} C_{gen}.P_{gen}(t) + C_{grid}(t).P_{grid}(t) \tag{8a}$$

$$\text{s.t. } P_{\text{Grid},t} = P_{\text{Load},t} + P_{\text{Bat},t} + P_{\text{H2},t} + P_{\text{pv},t} + P_{\text{TRHP},t} \qquad \forall t \tag{8b}$$

$$Q^{H-prod}_{\text{TRHP},t} + Q^{C-prod}_{\text{TRHP},t} = COP_{TRHP}.P_{\text{TRHP},t} \forall t \tag{8c}$$

$$Q^{H-prod}_{\text{TRHP},t} = Q_{H-load,t} + Q_{\text{HS},t} \forall t \tag{8d}$$

$$Q^{C-prod}_{\text{TRHP},t} = Q_{C-load,t} + Q_{\text{CS},t} \tag{8e}$$

$$P^{(i)}_{t} = P^{(i)}_{\text{ch,t}} + P^{(i)}_{\text{disch, t}} \qquad \forall i \in SS, \forall t \tag{8f}$$

$$E^{(i)}_{1} = E^{(i)}_{\text{init}}.(1 - k^{(i)}_{\text{sd}}) + \Delta_t \left(P^{(i)}_{\text{Ch},0} \eta_{\text{Ch}} - P^{(i)}_{\text{Disch},0} \frac{1}{\eta_{\text{Disch}}} \right) \forall i \in SS \tag{8g}$$

$$E^{(i)}_{\text{t+1}} = E^{(i)}_{\text{t}}.(1 - k^{(i)}_{\text{sd}}) + \Delta_t (P^{(i)}_{\text{ch,t}}.\eta_{\text{ch}} - \frac{1}{\eta_{\text{disch}}}.P^{(i)}_{\text{disch,t}} \qquad \forall i \in SS, \forall t \tag{8h}$$

$$E^{(i)}_{\text{min}} \leq E^{(i)}_{t} \leq E^{(i)}_{\text{max}} \qquad \forall i \in SS, \forall t \tag{8i}$$

$$P^{(i)}_{\text{min}} \leq P^{(i)}_{t} \leq P^{(i)}_{\text{max}} \qquad \forall i \in SS, \forall t \tag{8j}$$

Where Δt is the time slot (set to 1hour) and H is the optimization time horizon. Equation (8a) represents the cost function to be minimized, equations (8b) to (8e) express the electrical and thermal power balance within the Smart Grid, equations (8f) to (8h) express the dynamics of the multi-energy storage systems within the Smart Grid and equations (8i) to (8j) express the limitations on energy and charge and discharge power of each storage system, while SS represents the set of energy storage systems within the Smart-Grid.

4 Implementation details, simulations and results

A framework was developed based on the previously described DDPG algorithm and tested on the designed environment of a residential consumer multi-energy smart grid which parameters are given in table 1. During the training process, the DDPG agent was provided with three years of actual past realizations of PV generation, electric loads and electricity prices, as well as simulated data of heating and cooling loads and indoor and outdoor temperatures, for a residential consumer located in France. The historical data of a typical day in winter and in summer can be visualized in figures 3.a and 3.b. As in [8], we split the time series into a training set and a validation set that correspond to a different one year of historical data each. The Deep Neural Network (DNN) obtained at the end of the training process is then used in a test environment to provide an independent estimation of the final policy. Finally, to evaluate the performance of the proposed DRL approach, we use a benchmark solution that we refer to as "theoretical MPC". In this solution, we use an MPC controller that is supposed to have, at each day, a "perfect knowledge" of the stochastic system variables for the next 24 hours. Unlike the DDPG, the MPC was given the actual future realizations of the unknown quantities in the predictor. The MPC

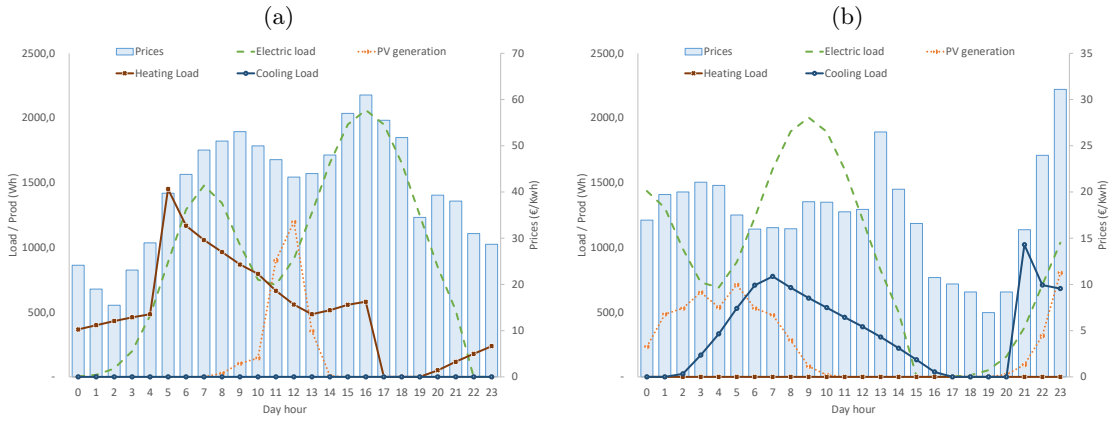

Fig. 3: Historical data used for (a): a typical winter day ; (b) : a typical summer day.

Table 1: Implementation details.

Parameters of the Smart Grid	Optimization parameters
size of the battery ξ_{bat}: 15 kWh	DDPG number of training episodes: 5000
Battery charge/discharge efficiency η_{bat}: 90%	DDPG number of training steps 438.10^5
size of the hydrogen ξ_{H2}: 1,1 kWh	DDPG learning rate of the actor:
Hydrogen charge/discharge efficiency η_{H2}: 65%	DDPG learning rate of the critic: 0,0001
size of the heat storage ξ_{HS}: 1,2 kWh	DDPG learning rate of the critic: 0.0002
HS charge/discharge efficiency η_{HS}: 75%	DDPG discount factor γ: $0,99$
size of the cooling storage ξ_{CS}: 0,8 kWh	DDPG and MPC time step: 1hour
CS charge/discharge efficiency η_{CS}: 75%	DDPG reward rescale factor α : $0,001$
Average electric consumption/day : 18 kWh/day	MPC time-horizon: $H = 24$ hours
Peak power generation of PV: 15 kWp	Solver used in MPC optimization: GLPK
Maximal heat/cold generated by TRHP :50 kWh	

with a time step $t = 1$ hour and a time horizon $H = 24$ hours was run for a one-year simulation, with the objective of minimizing the total operational costs. As shown in figure 4, the performance of the proposed DDPG-based approach is close to "theoretical MPC" optimum. These results demonstrate the effectiveness of the proposed DRL approach for multi-energy management of Smart Grids under uncertainty. The DDPG was able to take multiple scheduling continuous actions simultaneously and succeeded to extract knowledge from the past realizations of the stochastic variables. The DRL agent learnt a strategy similar to the optimal strategy given by the MPC-based approach. We notice for instance that the DRL agent successfully learnt to purchase electricity from the main utility grid at low price periods and to rather discharge the storage systems during peak price periods. It also successfully learnt to maintain the power balance within the multi-energy Smart Grid.

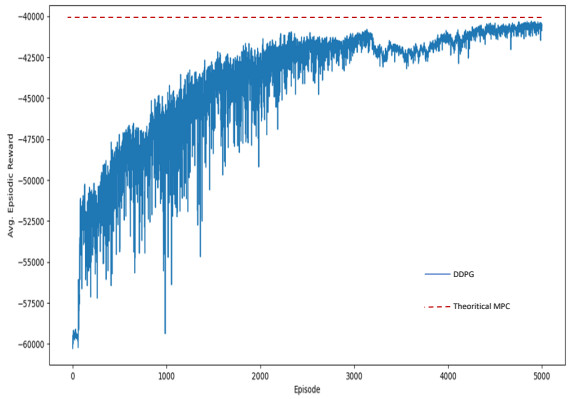

Fig. 4: Learning curve of the DDPG approach for a 5000-episode training process

5 Conclusion

This paper presented a DRL-based approach to deal with optimal energy management of multi-energy Smart Grids. The considered sequential decision making problem was formulated as an MDP and is addressed using a Deep Deterministic Policy Gradient (DDPG) algorithm. The developed framework was tested on the model of a multi-energy smart grid where the DDPG agent was designed to optimally schedule the various energy storage and thermal production units. The simulations showed that the agent handles well the continuous state and action spaces and learns to take multiple control actions simultaneously. Benchmark tests were conducted using a "theoretical MPC" solution to evaluate the performance of the proposed approach. Results showed that the total rewards obtained by the DDPG algorithm were close to the theoretical optimum and thus showed the effectiveness of the proposed DRL-based approach for dealing with optimal energy management of multi-enenrgy Smart Grid. More detailed results regarding the behavior of the policy will be given at the conference and will be the subject of upcoming papers. Future works also include the extension of the smart grid model to a district level smart grid where further devices are to be controlled including Heated water Storage tanks and other buildings controllable loads, Electric Vehicle Charging Stations and public lighting of the district. The proposed framework will also be applied on a real-life project of a multi-energy smart grid currently under construction in France.

Nomenclature

P_{Grid}	Grid power consumption
P_{gen}	Distributed power generation
C_{Grid}	Cost of power purchase from the grid
C_{gen}	Cost of distributed power generation
P_{Load}	Load power
P_{pv}	PV power generation
P_{Bat}	Battery power
P_{H2}	Hydrogen storage power
P_{TRHP}	electric power consumed by TRHP
$Q^{H-prod}_{\text{TRHP},t}$	heat produced by TRHP
$Q^{C-prod}_{\text{TRHP},t}$	cold produced by TRHP
COP_{TRHP}	Coefficient of performance of TRHP
Q_{H-load}	heating load
Q_{C-load}	cooling load
t	Time step
$P_{(\text{i})}$	power of a storage system i
$P^{(i)}_{\text{Ch}}$	Charging power of a storage system i
$P^{(i)}_{\text{Disch}}$	Discharging power of a storage system i
$P^{(i)}_{\text{min}}$	minimum power of storage system i
$P^{(i)}_{\text{max}}$	maximum power of storage system i
$\eta^{(i)}_{\text{Ch}}$	Charging efficiency of a storage system i
$\eta^{(i)}_{\text{Disch}}$	Discharging efficiency of a storage system i
$k^{(i)}_{\text{sd}}$	self-discharge rate of a storage system i
$E^{(i)}_{init}$	energy initially stored in storage system i
$E^{(i)}$	energy stored in storage system i

Acronyms

PV	Photo-voltaic
SoC	State of Charge
MG	Microgrid
SG	Smart Grid
$TRHP$	Thermo-Refrigerating Heat Pump
$SDHS$	Smart District Heating System
MPC	Model Predictive Control
MDP	Markov Decison Process
ML	Machine Learning
DL	Deep Learning
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
DQN	Deep Q-Networks
DQL	Deep Q-Learning
DPG	Deep Policy Gradient
$DDPG$	Deep Deterministic Policy Gradient

References

1. Bousnina, D., de Oliveira, W., Pflaum, P.: A stochastic optimization model for frequency control and energy management in a microgrid. In: International Conference on Machine Learning, Optimization, and Data Science. pp. 177–189. Springer (2020)
2. de Bruin, T., Kober, J., Tuyls, K., Babuška, R.: Improved deep reinforcement learning for robotics through distribution-based experience retention. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3947–3952. IEEE (2016)
3. Carta, S., Ferreira, A., Podda, A.S., Recupero, D.R., Sanna, A.: Multi-dqn: An ensemble of deep q-learning agents for stock market forecasting. Expert systems with applications **164**, 113820 (2021)
4. van den Ende, M., Lukszo, Z., Herder, P.M.: Smart thermal grid. In: 2015 IEEE 12th International Conference on Networking, Sensing and Control. pp. 432–437. IEEE (2015)
5. Fang, X., Misra, S., Xue, G., Yang, D.: Smart grid—the new and improved power grid: A survey. IEEE communications surveys & tutorials **14**(4), 944–980 (2011)
6. François-Lavet, V.: Contributions to deep reinforcement learning and its applications in smartgrids. Ph.D. thesis, Université de Liège, Liège, Belgique (2017)

7. François-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., Pineau, J.: An introduction to deep reinforcement learning. arXiv preprint arXiv:1811.12560 (2018)
8. François-Lavet, V., Taralla, D., Ernst, D., Fonteneau, R.: Deep reinforcement learning solutions for energy microgrids management. In: European Workshop on Reinforcement Learning (EWRL 2016) (2016)
9. Gao, G., Li, J., Wen, Y.: Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning. arXiv preprint arXiv:1901.04693 (2019)
10. Garcia, C.E., Prett, D.M., Morari, M.: Model predictive control: theory and practice—a survey. Automatica **25**(3), 335–348 (1989)
11. Gelleschus, R., Böttiger, M., Stange, P., Bocklisch, T.: Comparison of optimization solvers in the model predictive control of a pv-battery-heat pump system. Energy Procedia **155**, 524–535 (2018)
12. Ji, Y., Wang, J., Xu, J., Fang, X., Zhang, H.: Real-time energy management of a microgrid using deep reinforcement learning. Energies **12**(12), 2291 (2019)
13. Knight, W.: Google just gave control over data center cooling to an ai (2018)
14. Konda, V.R., Tsitsiklis, J.N.: Actor-critic algorithms. In: Advances in neural information processing systems. pp. 1008–1014 (2000)
15. Kuang, Y., Wang, X., Zhao, H., Huang, Y., Chen, X., Wang, X.: Agent-based energy sharing mechanism using deep deterministic policy gradient algorithm. Energies **13**(19), 5027 (2020)
16. Liaw, R., Krishnan, S., Garg, A., Crankshaw, D., Gonzalez, J.E., Goldberg, K.: Composing meta-policies for autonomous driving using hierarchical deep reinforcement learning. arXiv preprint arXiv:1711.01503 (2017)
17. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)
18. Lin, S., Yu, H., Chen, H.: On-line optimization of microgrid operating cost based on deep reinforcement learning. In: IOP Conference Series: Earth and Environmental Science. vol. 701, p. 012084. IOP Publishing (2021)
19. Liu, H., Yu, C., Wu, H., Duan, Z., Yan, G.: A new hybrid ensemble deep reinforcement learning model for wind speed short term forecasting. Energy **202**, 117794 (2020)
20. Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A.: Application of deep reinforcement learning to intrusion detection for supervised problems. Expert Systems with Applications **141**, 112963 (2020)
21. Ma, T., Wu, J., Hao, L., Lee, W.J., Yan, H., Li, D.: The optimal structure planning and energy management strategies of smart multi energy systems. Energy **160**, 122–141 (2018)
22. Mancarella, P.: Smart multi-energy grids: concepts, benefits and challenges. In: 2012 IEEE Power and Energy Society General Meeting. pp. 1–2. IEEE (2012)
23. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)
24. Mocanu, E., Mocanu, D.C., Nguyen, P.H., Liotta, A., Webber, M.E., Gibescu, M., Slootweg, J.G.: On-line building energy optimization using deep reinforcement learning. IEEE transactions on smart grid **10**(4), 3698–3708 (2018)
25. Morari, M., Lee, J.H.: Model predictive control: past, present and future. Computers & Chemical Engineering **23**(4-5), 667–682 (1999)
26. Parisio, A., Rikos, E., Glielmo, L.: A model predictive control approach to microgrid operation optimization. IEEE Transactions on Control Systems Technology **22**(5), 1813–1827 (2014)

27. Pflaum, P., Alamir, M., Lamoudi, M.Y.: Comparison of a primal and a dual decomposition for distributed mpc in smart districts. In: 2014 IEEE international conference on smart grid communications (SmartGridComm). pp. 55–60. IEEE (2014)
28. Qin, J., Han, X., Liu, G., Wang, S., Li, W., Jiang, Z.: Wind and storage cooperative scheduling strategy based on deep reinforcement learning algorithm. In: Journal of Physics: Conference Series. vol. 1213, p. 032002. IOP Publishing (2019)
29. Sallab, A.E., Abdou, M., Perot, E., Yogamani, S.: Deep reinforcement learning framework for autonomous driving. Electronic Imaging **2017**(19), 70–76 (2017)
30. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. nature **529**(7587), 484–489 (2016)
31. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: International conference on machine learning. pp. 387–395. PMLR (2014)
32. Sogabe, T., Malla, D.B., Takayama, S., Shin, S., Sakamoto, K., Yamaguchi, K., Singh, T.P., Sogabe, M., Hirata, T., Okada, Y.: Smart grid optimization by deep reinforcement learning over discrete and continuous action space. In: 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC)(A Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC). pp. 3794–3796. IEEE (2018)
33. Stănişteanu, C.: Smart thermal grids–a review. The Scientific Bulletin of Electrical Engineering Faculty **1**(ahead-of-print) (2017)
34. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
35. Sutton, R.S., Barto, A.G., et al.: Introduction to reinforcement learning, vol. 135. MIT press Cambridge (1998)
36. Tuballa, M.L., Abundo, M.L.: A review of the development of smart grid technologies. Renewable and Sustainable Energy Reviews **59**, 710–725 (2016)
37. Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Rothörl, T., Lampe, T., Riedmiller, M.: Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. arXiv preprint arXiv:1707.08817 (2017)
38. Wang, T., Kamath, H., Willard, S.: Control and optimization of grid-tied photovoltaic storage systems using model predictive control. IEEE Transactions on Smart Grid **5**(2), 1010–1017 (2014)
39. Yang, L., Entchev, E., Rosato, A., Sibilio, S.: Smart thermal grid with integration of distributed and centralized solar energy systems. Energy **122**, 471–481 (2017)
40. Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y., Jiang, T.: Deep reinforcement learning for smart home energy management. IEEE Internet of Things Journal **7**(4), 2751–2762 (2019)
41. Zhang, B., Hu, W., Cao, D., Huang, Q., Chen, Z., Blaabjerg, F.: Deep reinforcement learning–based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy. Energy Conversion and Management **202**, 112199 (2019)
42. Zhang, T., Luo, J., Chen, P., Liu, J.: Flow rate control in smart district heating systems using deep reinforcement learning. arXiv preprint arXiv:1912.05313 (2019)