Managing Power Flows in Microgrids Using Multi-Agent Reinforcement Learning

Fabrice Lauri, Gillian Basso, Jiawei Zhu, Robin Roche, Vincent Hilaire,
Abderrafiâa Koukam

Institut de Recherche sur les Transports, l'Energie et la Société, laboratoire Systèmes et Transport (IRTES-SeT)

Université de Technologie de Belfort-Montbéliard, Belfort, France

{fabrice.lauri,gillian.basso,jiawei.zhu,robin.roche,vincent.hilaire,abder.koukam}@utbm.fr

ABSTRACT

Smart Microgrids bring numerous challenges, including how to leverage the potential benefits of renewable sources while maintaining acceptable levels of reliability in the power infrastructure. One way to tackle this challenging problem is to use intelligent storage systems (batteries and supercapacitors). Charging and discharging them at the proper time by exploiting the variablity of the renewable sources guarantees to balance supply and demand at any time. Reinforcement learning (RL) is a branch of artificial intelligence encompassing techniques that allow agents (in our case electrical devices) to learn to behave rationnally, that is to perform sequences of decisions in order to optimize a given performance criteria. The theoretically sound framework of Reinforcement Learning makes these techniques to be increasingly used for solving difficult control problems. In this paper, a multi-agent reinforcement learning technique is proposed as an exploratory approach for controling a grid-tied microgrid in a fully distributed manner, using multiple energy storage units and the grid. Preliminary simulation results using different scenarios show the feasibility and validity of the approach on a test microgrid, and open the way for future work in the field of agent-based learning control strategies in Smart Microgrids.

Categories and Subject Descriptors

I.2.6 [**Learning**] ; I.2.11 [**Distributed Artificial Intelligence**]: Multi-Agent Systems.

General Terms

Algorithms, Management.

Keywords

Distributed control, microgrid, multi-agent systems, reinforcement learning, Q-learning.

1. INTRODUCTION

Most of today's power grid evolved after 1893, based on Nikola Tesla's design [15]. However with the dramatic development of human society, more and more disadvantages of traditional power grid appear, such as low efficiency, limited extensibility, centralization, etc. Moreover, growing concerns over the environmental damage from conventional power stations call for increasing the use of renewable energy like wind power and solar power. Therefore, the requirement of Smart Grids becomes apparent.

The key benefits enabled by Smart Grids are bidirectional communication of information and power flow, efficient management of power consumption and distribution, decentralized power transmission. By integrating information and communication technology with advanced monitoring and control at all levels in the system, future infrastructures will become smarter, more efficient, more reliable and more flexible.

While the technologies of Smart Grid and renewable energy bring numerous opportunities for future power system, they also require to solve important challenges for the design and management of power infrastructures and power distribution. One of them comes from the usage of renewable energy sources. Although this sort of energy helps reduce carbon emissions, it is difficult to provide reliable and stable power as a result of uncertainties of the weather and other factors. Therefore, how to leverage the potential benefits of renewable energy sources while maintaining acceptable levels of reliability in the power infrastructure becomes a key point. One way to mitigate this problem is to take advantage of intelligent storage systems (i.e. battery and super capacitor) to conveniently coordinate the electrical supply and demand in the Smart Grid: the storage systems charge when there exists excess energy, and they discharge when renewable energy sources solely provide insufficient energy.

Multi-Agent Systems (MAS) constitute a useful paradigm for modelling complex systems or for providing alternative perspectives on systems that are originally considered as centralized [13]. A MAS consists of a collection of autonomous entities called agents that interact within a common environment: they perceive states of their environment, and act upon it based on their perceptions [19, 18].

A microgrid may be viewed as a MAS, where the loads, the renewable energy sources (RES), the storage systems and the distribution grid access point can be considered as agents embedded in power electronics devices. For instance, while the electrical devices in a microgrid are generally controlled by a central authority, identifying each device as an agent with its own intelligence may provide a meaningful, distributed perspective on the system. This paradigm is increasingly applied to solve problems related to power systems control and monitoring [4, 5, 9, 10, 11, 12, 16, 17].

The problem of optimally controlling power flows in a microgrid by minimizing the cost of energy drawn from the grid can be formulated as a multi-agent sequential decision

problem. Its resolution can be efficiently handled by some artificial intelligence techniques, and more precisely by reinforcement learning (RL) techniques. Such techniques can be used to allow agents to find the rational behaviors leading to the achievement of a given common task, e.g., maintaining the balance between supply and demand.

A reinforcement learning agent learns to achieve a given task by interacting with its environment. At each time step, the agent perceives the environment state and performs an action, which causes the environment to transit to a new state. A scalar reward evaluates the quality of each transition, allowing the agent to observe the cumulative reward along sequences of interactions. By trials and errors, the agent can manage to find a policy, that is a mapping from states to actions, which maximizes the cumulative reward. Provably convergent algorithms are available for solving single-agent learning tasks, whether the environment state space is small [14], large or continuous [2].

Under some assumptions, classical RL techniques can be extended to cooperative and/or competitive multi-agent RL and they have already shown impressive results in various fields, like in robotics, games, health and transportation networks management.

In this paper, we advocate the use of a multi-agent reinforcement learning (MARL) technique to manage power flows in a grid-tied microgrid, in a fully distributed manner. The remainder of this paper is organized as follows. Section 2 presents related work and discusses their advantages and drawbacks. Section 3 describes the problem related to distributed microgrid control. Section 4 introduces the classical reinforcement learning framework as well as the proposed approach for solving the studied problem. Section 5 relates the simulation results obtained using this approach, and section 6 concludes this paper.

2. RELATED WORK

2.1 Reinforcement Learning in MAS Design

Reinforcement Learning techniques applied to the design of Multi-Agent Systems have been used to solve problems in a variety of domains, including robotics, telecommunications, scheduling and traffic control.

Mataric described a multi-agent RL method which enables agents to learn collective tasks in the complex concurrent multi-robot domain ([8]). In this method, the author reduces the learning space with the help of behaviors (macro-actions), which abstract away the details of the low-level control of the robot and the details of the agent's state space separately. Chang et al. [3] treat mobilized ad-hoc networks as a multi-agent learning system and apply reinforcement learning techniques with global rewards as a coordination signal to solve packet routing and node movement problems. Wu et al. [20] solve the job scheduling problems in grid computing by a novel multi-agent RL method, called ordinal sharing learning (OSL). According to this method, load balancing in grids can be realized. Moreover, the approach fulfils the scalability by an ordinal distributed learning strategy and the coordination among agents with limited communication based on an information sharing mechanism. Kuyer et al. [7] figure out the traffic control problem by considering traffic lights as RL agents, the coordination among which is achieved utilizing the max-plus algorithm. But the coordination graphs used is fixed, with dependencies only

between intersections connected by a road. Besides, max-plus is centralized.

2.2 MAS in problems related to Power Systems

Multi-agent systems have been advocated recently as powerful tools for managing power systems [9, 10] or for simulating Smart Grids [6]. Multi-agent based coordination strategies are increasingly used to solve problems similar to the one addressed in this article.

For example, Roche et al. [12] review several solutions based on Multi-Agent Systems for managing grid energy. In [17], a game-theoretic framework is adopted to study micro-storage strategies. The purpose of the framework is to reduce the peak of the demand thanks to the micro-storage. In this system, the agents are selfish, that is they always aim at minimizing their individual costs. The results show the average storage profiles of micro-storages converge towards a Nash equilibrium. In [16], authors focus on the problem of integrating renewable energy generation into the smart grid. As renewable sources are stochastic suppliers, consumer agents, which manage a home consumption and storage system, have to choose the amount of energy given by the supplier to reduce the cost of the energy by optimizing the life of their own storage system. The equilibrium between the suppliers production and the agents consumption is reached using a Homeostatic Control that minimizes the difference between green energy and total demand. Coordination in large-scale dynamic multi-agent environments usually suffer from convergence problems. Chongjie et al. [4] propose a decentralized supervision framework that enables lower-level agents to communicate abstracted states to higher-level supervision agents so that the latter ones can efficiently guide the learning of the former ones on the basis of a broader view of the system.

3. PROBLEM FORMULATION

We assume a DC microgrid, in which RES (photovoltaic arrays and wind generators), storage systems (super-capacitors and batteries), a distribution grid access point and loads are employed, as shown in Fig. 1. In order to maintain the microgrid stable, the grid is assumed to be able to deliver or receive unlimited amounts of power and energy. For example, when the RES and the storage systems cannot satisfy the demand, the grid provides the remaining power, while when there exists excess power, it can be sold to the grid.

The DC bus is treated as a capacitance satisfying at any time t:

$$I_{dc}(t) = -C \frac{dV_{dc}(t)}{dt} \qquad (1)$$

where $I_{dc}(t)$ is the current on the DC bus, C is its capacity, and $V_{dc}(t)$ is its voltage. For $V_{dc}(t)$ to be considered stable, it must verify the following constraint at all time:

$$V_{dc}^{min} \leq V_{dc}(t) \leq V_{dc}^{max} \qquad (2)$$

Current $I_{dc}(t)$ is computed from the currents delivered by the devices (loads, RES, storage systems, and grid) in the
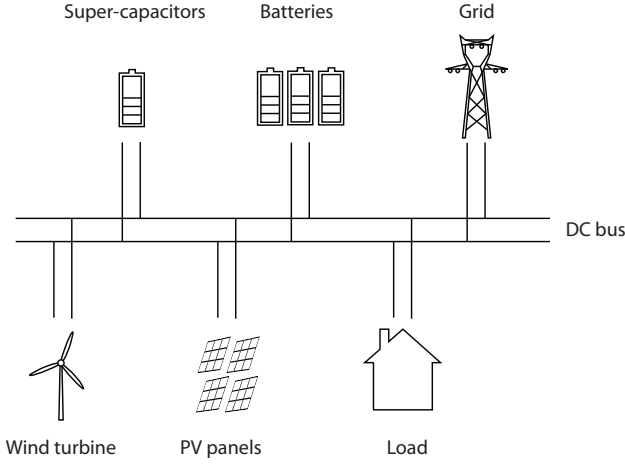
Figure 1: Basic microgrid components.

microgrid, according to Kirchhoff's law:

$$\sum_k I_k(t) = \sum_{a=1}^{n_{sto}} I_{sto,a}(t) + \sum_{b=1}^{n_{lo}} I_{lo,b}(t)$$
$$+ \sum_{c=1}^{n_{res}} I_{res,c}(t) + I_{grid}(t) + I_{dc}(t)$$
$$= 0 \qquad (3)$$

where:

- n_{sto}, n_{lo}, and n_{res} are the number of storage systems, loads, and RES, respectively.

- $I_{sto,a}(t)$, $I_{lo,b}(t)$, $I_{res,c}(t)$, $I_{grid}(t)$ are the currents delivered by storage system a, load b, RES c, and the grid, respectively.

The DC Bus voltage is computed using:

$$V_{dc}(t) = V_{dc}^{init} - \frac{1}{C}\int I_{dc}(t)\, dt$$
$$= V_{dc}^{init} + \frac{1}{C}\int \left(\sum_{a=1}^{n_{sto}} I_{sto,a}(t) + \sum_{b=1}^{n_{lo}} I_{lo,b}(t) \right.$$
$$\left. + \sum_{c=1}^{n_{res}} I_{res,c}(t) + I_{grid}(t) \right) dt \qquad (4)$$

Every storage system k is defined by the following dynamic variables:

- Its power $P_{sto,k}(t)$, such that:

$$P_{storage,k}(t) = \begin{cases} \frac{I_{sto,k} \cdot V_{dc}}{\eta_k} & \text{if } I_{sto,k} \geq 0 \\ I_{sto,k} \cdot V_{dc} \cdot \eta_k & \text{if } I_{sto,k} < 0 \end{cases} \qquad (5)$$

where η_k is the conversion efficiency.

- Its state-of-charge $SOC_k(t)$, calculated as follows:

$$SOC_k(t) = SOC_k^{init} - \frac{\int_0^t P_{sto,k}(t)\, dt}{E_k}$$

where SOC_k^{init} is the initial state-of-charge and E_k is its capacity.

Every storage system k is also subject to the following constraints:

- Constraint $\mathcal{C}1$ on power:

$$P_{sto,k}^{min} \leq P_{sto,k}(t) \leq P_{sto,k}^{max}$$

with $P_{sto,k}^{min} < 0$ and $P_{sto,k}^{max} > 0$

- Constraint $\mathcal{C}2$ on power dynamics:

$$\dot{P}_{sto,k}^{min} \leq \dot{P}_{sto,k}(t) \leq \dot{P}_{sto,k}^{max}$$

with $\dot{P}_{sto,k}^{min} < 0$ and $\dot{P}_{sto,k}^{max} > 0$

- Constraint $\mathcal{C}3$ on state-of-charge limits:

$$0 \leq SOC_k^{min} \leq SOC_k(t) \leq SOC_k^{max} \leq 1$$

To simplify the resolution of the problem, constraint $\mathcal{C}3$ is relaxed. In this case, it is assumed that the life-time of any storage system is maximized as long as $\mathcal{C}3$ is satisfied.

In the previous equations, current pointing to the DC bus is considered as positive, i.e., power generated by a source is positive and the power consumed by a load is negative. In particular, $I_{lo,b} \leq 0$ and $I_{res,c} \geq 0$ for any load b and RES c. It is also worth noting that batteries and super-capacitors are distinguished solely by their different parameter settings, i.e., super-capacitors are capable of higher powers while containing less energy than batteries.

Optimally managing power flows in the microgrid consists in deciding at every time t, for each storage system k, and for the main grid, which currents $I_{sto,k}(t)$ and $I_{grid}(t)$ are to be delivered, so that the voltage V_{dc} remains stable, and such that the main grid is solicited as little as possible. Mathematically speaking, the problem consists in minimizing the following objective function f:

$$f = \lim_{T \to \infty} \int_0^T \gamma^t (I_{grid}(t))^2\, dt \qquad (6)$$

where $\gamma \in [0;1)$ is a discount factor used to bound the integral.

4. PROPOSED APPROACH

4.1 Single-Agent Reinforcement Learning

The classical reinforcement learning framework involves one learning agent whose task can be generally modelled by a discrete-time Markov decision process (MDP) $M = (X, U, f, \rho, T, \gamma)$. X is the set of states it can perceive, U is the set of the possible actions it can perform in these states, f is the state transition function, ρ is the reward function that evaluates the immediate effect of an action, $T \subseteq \mathbb{N}$ is the discrete time space, and γ is the discount factor.

In the infinite-horizon case, that is when T is infinite and $\gamma \in [0;1)$, the optimal behavior of the agent can be fully described by a policy $h^* : S \to A$, that is a mapping from states to actions that maximizes the expected total discounted return $R^h(x_0)$ over a set of initial states $x_0 \in X_0 \subseteq X$, such that:

$$R^h(x_0) = \left\langle \sum_{t \in T} \gamma^t r_{t+1} \right\rangle_{x_0, h} \qquad (7)$$

r_t is the reward received at time $t \in T$, $\langle \ \rangle_{x_0,h}$ is the expected return over all stochastic state transitions from state x_0 when following policy h, and $\gamma \in [0,1)$ is the discount factor that gives more weight to near-term reward, and guarantees that the sum of rewards is bounded.

The MDP encodes a decision sequential problem that can be solved using dynamic programming algorithms when the state transition function f is known, or RL algorithms like Q-Learning when f is unknown (which is the case here). For complex tasks, when the state-space and/or the action space are very large, RL algorithms need to be used as they do not require global update sweeps on every state. Updates represent the learning that occurs after observing a transition from state x to state x' with immediate reward r when applying action u in x. In Q-learning, an update of the Q-function (a value function that estimates the expected discounted total return for each pair of states and actions) is defined as:

$$Q(x,u) \quad \overset{\alpha}{\leftarrow} \quad r + \gamma \max_{u' \in U} Q(x',u') \qquad (8)$$

where $A \overset{\alpha}{\leftarrow} B$ means $A \leftarrow (1-\alpha)A + \alpha B$. The learning factor $0 \le \alpha \le 1$ influences the learning speed.

Q is guaranteed to converge to the optimal Q-function Q^* under some statistical conditions that can be satisfied if every state is visited infinitely often and if action u is selected using an action selection method like ϵ-greedy or *Softmax* [14].

Once Q has converged to the optimal Q-function Q^*, the optimal action at each state x can be obtained as follows:

$$h^*(x) \quad = \quad \text{argmax}_{u \in U} Q^*(x,u) \qquad (9)$$

Equation (8) implicitly assumes that a Q-value estimate can be stored for each state-action pair. When the state-action space is too large or continuous, function approximators need to be used [2], as all the Q-values could no longer either be stored or reliably estimated.

4.2 MARL Algorithm

Several approximations have to be provided in order to solve this problem in the framework of RL, since the state space, the action space and the time space in this problem are all continuous and multiple agents are involved. The MARL algorithm presented thereafter makes the following assumptions:

- *Agent groups:* It is assumed that agents can be grouped into some equivalence classes. All the agents belonging to the same class will exhibit the same behavior. Let $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_N\}$ be a set of N classes of agents, or *taxons*. $\mathcal{G}_i$ represents the indices of the agents belonging to taxon i. Every agent $k \in \mathcal{G}_i$ learns a policy and select their actions by using the Q-values associated with taxon $\mathcal{G}_i$. As they exploit the same Q-values, agents belonging to the same taxon will exhibit the same behavior whenever they perceive the same values of the state variables. Using taxons both simplifies the problem, by suggesting that agents of the same type are likely to behave similarly, and it makes the multi-agent system open and cooperation between similar agents efficient, as new agents added to an existing taxon will immediately benefit from the experiences of the other same agents and will be able to indirectly

communicate its experience to them through the associated Q-values.

- *Time space:* The time space is discretized, in such a way that the devices (storage systems and grid) deliver current at a (time) frequency sufficiently high for enabling the DC bus voltage to remain stable. In particular, only storage systems are controllable, and act successively, one after the other. The grid is solicited every time after a controllable device has performed an action, and guarantees the stability of the DC bus voltage by providing the required power. This agent scheduling is depicted in Fig. 2.
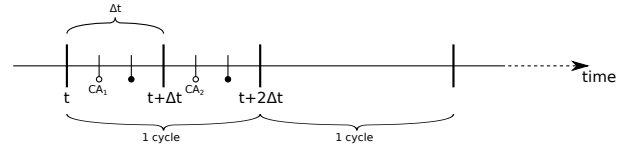

Figure 2: Action scheduling for 2 agents. Filled circles represent the actions of the main grid, hollow circles the actions of the controllable devices (batteries and super-capacitors). A cycle consists of a sequence of actions that ends when any controllable device has performed an action.

- *State space:* Any controllable device k (batteries and super-capacitors) observes the following information at any time t:

  - Relative demand: $D_k(t) = I_{lo}(t) + I_{res}(t) + I_{grid}(t)$
  - Current: $I_{sto,k}(t)$
  - State-of-charge: $SOC_k(t)$

- *Action space:* Each controllable device $k \in \mathcal{G}_i$ has the ability to choose a command $C_k(t) \in \mathcal{U}_i \subset [-1,1]$ that influences the variation in current as follows:

$$\dot{I}_{sto,k}(t) = \begin{cases} C_k(t)\ \dot{I}_{sto,k}^{max} & \text{if } C_k(t) \ge 0, \\ -C_k(t)\ \dot{I}_{sto,k}^{min} & \text{otherwise} \end{cases} \qquad (10)$$

  where $\dot{I}_{sto,k}^{min}$ and $\dot{I}_{sto,k}^{max}$ are respectively the minimum and the maximum allowed variation in current for storage system k, satisfying constraints $\mathcal{C}1$ and $\mathcal{C}2$.

- *Reward function:* Each controllable device receives a reward defined by:

$$r_k(t) \quad = \quad -|I_{grid}(t)| \qquad (11)$$

- *Approximation of the Q-functions:* As the state space of this problem is continuous, the Q-function learnt by each agent must be approximated. For theoretical reasons of guaranteed convergence, we chose to linearly approximate the Q-function learnt by each agent $k \in \mathcal{G}_i$ as follows:

$$\widehat{Q}_i(x,u) \quad = \quad \phi_i^T(x)\,\theta_{i,u}$$

  where:

  - $u \in \mathcal{U}_i = \{-1, \cdots, 0, \cdots, 1\}$ is the command chosen from the *a priori* discrete set of commands $\mathcal{U}_i$ associated with taxon i.

- $\phi_i(x) \in \mathbb{R}^n$ is the vector of features extracted from the state x using information defined for the taxon i.
- $\theta_{i,u} \in \mathbb{R}^n$ is the vector of weights associated with the taxon i and the discrete action u.

Each agent $k \in \mathcal{G}_i$ uses $Q(\lambda_i)$ with a linear parametrization and ϵ-greedy for learning the weights $\theta_{i,u}$. $Q(\lambda)$ is a version of Q-Learning with eligibility traces (see [2, p. 61] for $Q(0)$, the version of Q-Learning without eligibility traces and [14] for the computation of the eligibility traces in $Q(\lambda)$). Features are extracted using CMAC [1].

With a linear function approximator, $Q(\lambda_i)$ updates the weigths $\theta_{i,u}$ associated with the taxon i and the action u performed when the state x is perceived as follows:

$$\delta_i(x,u) \quad \leftarrow \quad r + \gamma_i \max_{u'} \hat{Q}_i(x',u') - \hat{Q}_i(x,u) \quad (12)$$

$$\Delta\theta_{i,u} \quad \leftarrow \quad \alpha_i \, \delta_i(x,u) \, \phi_i(x) \, e_i(x,u) \quad (13)$$

where x' and r are respectively the state and the reward observed after performing the action u in state x. $e_i(x,u)$ is the eligibility trace associated with taxon i and used by any agent $k \in \mathcal{G}_i$. An eligibility trace is a record of the occurrence of an event, such as the execution of an action in a state. When eligibility traces are used in reinforcement learning techniques like Q-Learning, these techniques may learn more efficiently. Only the eligibility traces that are greater than zero allow the associated state-action pairs to undergo learning changes.

Algorithm 1 was used during the learning stage. Once the Q-functions have been learnt, every agent $k \in \mathcal{G}_i$ selects the optimal actions they will perform when perceiving the state x according to the following equation:

$$h_k(x) \quad = \quad \mathrm{argmax}_{u \in U} \hat{Q}_i(x,u) \quad (14)$$

Algorithm 1 Shared Q-Learning used by every agent k

Require: Taxon i such that agent $k \in \mathcal{G}_i$.
Require: Set of weights $\theta_{i,u}$ for all the actions $u \in \mathcal{U}_i$.
Require: Learning factor α_i.
Ensure: Function $\hat{Q}_i$.
    Observe state x.
    repeat
        Select action $u \in \mathcal{U}_i$ in x by using a selection action strategy like $\epsilon - Greedy$.
        Execute u and observe the new state x' and the reward r.
        Update the weights $\theta_{i,u}$.
        $x \leftarrow x'$.
    until Convergence is reached.

4.3 Relevance of the Approach

The main advantages of the proposed MARL approach is that the control of the microgrid is fully distributed. Each agent performs a Q-Learning algorithm that exploits state variables based solely on the information perceived directly from the environment, and not built from the information

transmitted by other agents. In this algorithm, the communication between agents is stigmergic, that is it emerges from the interactions of all the agents. Since the agents successively act upon the environment, each agent's action updates the environmental state from which reliable information can be extracted and exploited immediately by other agents. Besides, the proposed algorithm allows new devices to immediately operate into the environment, as their decisions are solely based upon the resolution of equation (14).

5. SIMULATION RESULTS

Several simulations were conducted to validate the robustness of the proposed approach, using 16 different scenarios. In each scenario, the main grid as well as n_1 batteries and n_2 super-capacitors were involved, where n_1 and n_2 could take on values ranging from 0 to 4.

One learning stage and one evaluation stage were run successively. A learning stage consists of 200 episodes for each scenario, thus resulting in $16 \times 200 = 3,200$ learning episodes. During a learning episode, every agent takes a sequence of decisions and updates the associated short memory of Q-values using equations (12) and (13) every time it performs an action. During an evaluation episode, learning is disabled for all agents and equation (14) is used by all the agents to select optimal actions. The total grid contribution (equation (15)), the average grid contribution and the maximum grid contribution is computed at the end of an evaluation episode (see Table 3).

$$\int_{t \in T} |I_{grid}(t)| \, V_{dc}(t) \, dt \quad (15)$$

Typical load and RES profiles over two days are used for both the learning stage and the evaluation stage. The cycle duration is set to one second, during which a controllable device and the grid perform an action, i.e., $\Delta t = 1s$ in Fig. 2, for a total of $172,800$ decisions per episode. The algorithm settings and the microgrid parameters shown in Tables 1 and 2, respectively, were used in all experiments.

Taxon $\mathcal{G}_1$ (Batteries)	Taxon $\mathcal{G}_2$ (Super-capacitors)
$\mathcal{U}_1 = \{-1, 0, 1\}$	$\mathcal{U}_2 = \{-1, -0.5, 0, 0.5, 1\}$
$\epsilon_1 = 0.2$	$\epsilon_2 = 0.2$
$\gamma_1 = 0.995$	$\gamma_2 = 0.995$
$\lambda_1 = 0.9$	$\lambda_2 = 0.9$
$\alpha_1 = 0.5$	$\alpha_2 = 0.5$
Regular CMAC grids 60 tilings	Regular CMAC grids 60 tilings
$12 \times 8 \times 4$ tiles per tiling	$12 \times 8 \times 6$ tiles per tiling

Table 1: Algorithm settings

Results in Table 3 show that the total grid contribution decreases as more similar agents (batteries or super-capacitors) are used. Contrary to batteries, super-capacitors can deliver a larger amount of current during a shorter period. This explains why the grid contribution is larger when super-capacitors are used together than when batteries are.

When batteries are used in conjunction with super-capacitors, the results are still quite satisfying: the main grid contribution significantly decreases as more agents (especially batteries) are involved.

DC bus parameters	
$V_{\mathrm{dc}}^{\min} = 680$ V	$V_{\mathrm{dc}}^{\max} = 720$ V
$V_{\mathrm{dc}}(0) = 700$	$C = 20$ mF
Battery parameters	
$V_{\mathrm{bat}} = 400$ V	
$P_{\mathrm{bat}}^{\min} = -300$ kW	$P_{\mathrm{bat}}^{\max} = 300$ kW
$\dot{P}_{\mathrm{bat}}^{\min} = -2$ kW/s	$\dot{P}_{\mathrm{bat}}^{\max} = 2$ kW/s
$E_{\mathrm{bat}}^{\mathrm{tot}} = 10$ MWh $= 25$ kAh	$\eta_{\mathrm{bat}} = 1$
Super-capacitor parameters	
$V_{\mathrm{sc}} = 400$ V	
$P_{\mathrm{sc}}^{\min} = -600$ kW	$P_{\mathrm{sc}}^{\max} = 600$ kW
$\dot{P}_{\mathrm{sc}}^{\min} = -10$ kW/s	$\dot{P}_{\mathrm{sc}}^{\max} = 10$ kW/s
$E_{\mathrm{sc}}^{\mathrm{tot}} = 600$ kWh $= 1500$ Ah	$\eta_{\mathrm{sc}} = 1$

Table 2: Microgrid parameters

Scenario	(n_1, n_2)	Grid Contribution		
		Total (MWh)	Average (kW)	Max (kW)
$\mathcal{S}_1$	$(1, 0)$	1.69 (0)	35.26 (0.03)	325.3 (0)
$\mathcal{S}_2$	$(2, 0)$	0.1 (0.01)	1.99 (0.25)	325.3 (0)
$\mathcal{S}_3$	$(0, 1)$	5.25 (0)	109.45 (0.01)	322.3 (0)
$\mathcal{S}_4$	$(0, 2)$	4.87 (0)	101.54 (0.08)	365.3 (84.1)
$\mathcal{S}_5$	$(1, 1)$	1.38 (0.01)	28.9 (0.29)	325.33 (0)
$\mathcal{S}_6$	$(2, 1)$	0.17 (0.09)	3.51 (1.96)	325.33 (0)
$\mathcal{S}_7$	$(1, 2)$	1.02 (0.12)	21.26 (2.49)	325.33 (0)
$\mathcal{S}_8$	$(2, 2)$	0.1 (0)	2.03 (0.1)	325.33 (0)

Table 3: Results obtained in the evaluation stage. 95% confidence intervals are listed in parentheses.

The learnt strategies exhibited by the storage systems as well as the variations of their states-of-charge are shown in Figs. 3 to 8. When the microgrid consists of only one battery, in Fig. 3, one can see that the load is supplied by the battery as long as its state-of-charge allows it. The same conclusion can be drawn when one super-capacitor (Fig. 4) has to supply the load, but for a shorter period for obvious reasons. When two super-capacitors are used, the time during which the demand is at least partly supplied by the storage systems is doubled. When two batteries are used (Fig. 5), the main grid is very little sollicited because both batteries have the required cumulated capacities to ensure supplying the demand. The joint strategies is not optimal from the point of view of an electrical engineer. Indeed, it is very unlikely in real applications that a battery is used for charging another one. Yet the learnt strategies is optimal in the context of the formulated problem. Further work will be dedicated among others for avoiding this kind of situations. When the microgrid consists of one battery and one super-capacitor, in Fig. 7, the contribution of the main grid is maintained around a low value most of the time, as long as the storage units have enough energy left. Power is mainly delivered by the super-capacitor during the first hour. It is then delivered mainly by the battery until the 33rd hour. From the 33rd hour for the battery (respectively, from the 2nd hour for the super-capacitor), the state-of-charge of the battery (respectively of the super-capacitor) reaches its lowest value. For this reason, power is mostly delivered by the grid from the 33rd hour until the end of the simulation. When two batteries and two super-capacitors are used, the main grid

is very little sollicited as all the storage systems participate in supplying the demand. As in Fig. 5, including some electrical engineering knowledge will surely improve the joint learnt strategies. In all the figures, the peaks are due to the loss of one or more controllable devices.

In the light of the above analysis, one can assert that two types of devices, batteries and super-capacitors, whose power capacities and dynamics are different, have learnt to jointly supply the demand in a microgrid in a rather efficient way.
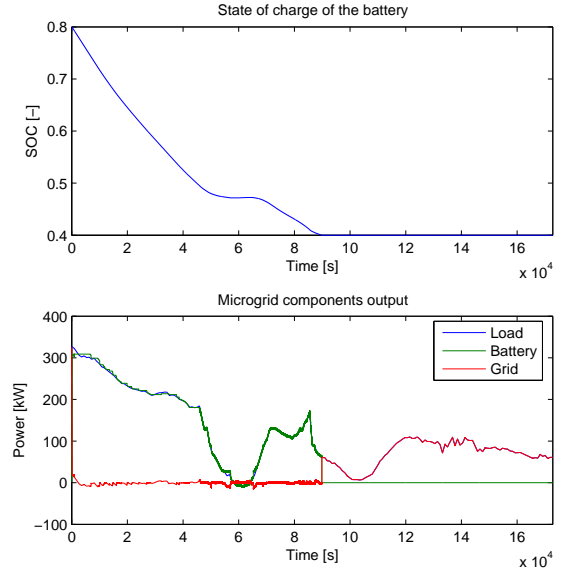

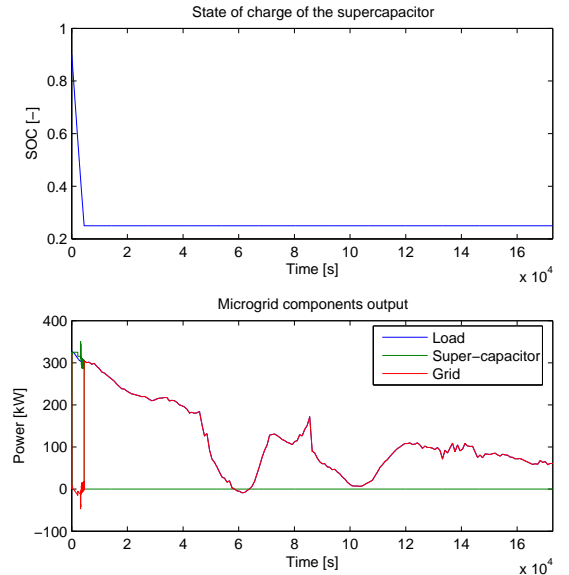**Figure 3: Learnt strategies exhibited by one battery.**

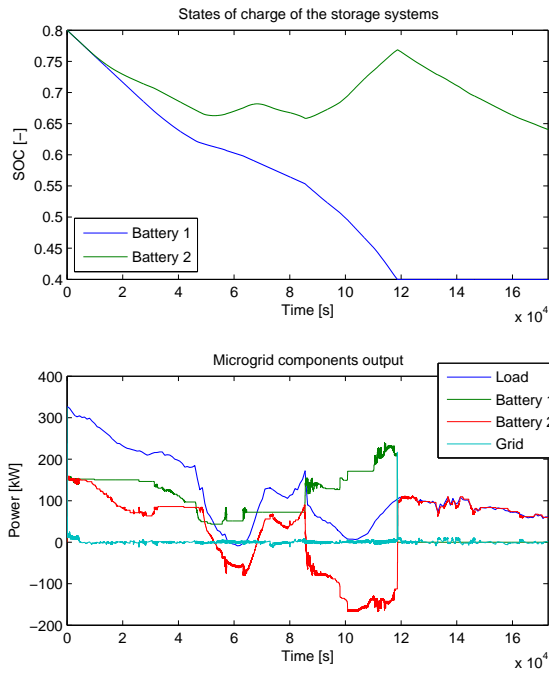**Figure 4: Learnt strategies exhibited by one super-capacitor.**
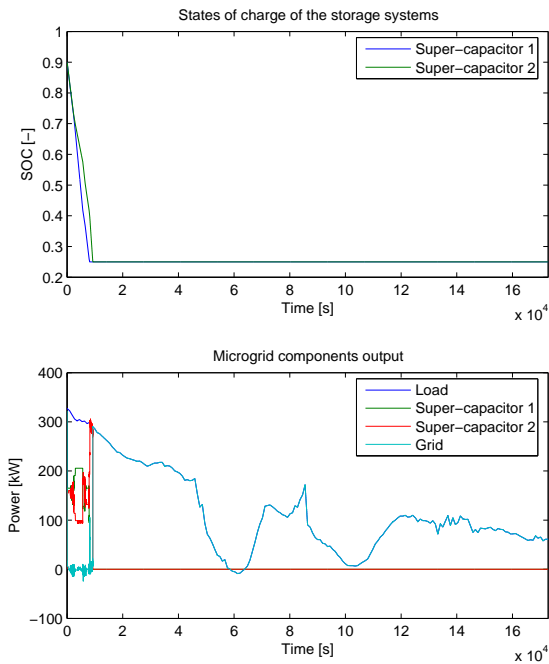
Figure 5: Learnt strategies exhibited by two batteries.
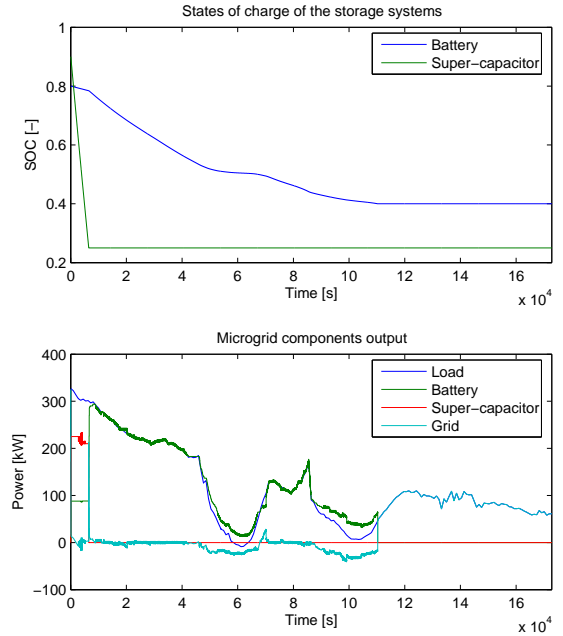

Figure 7: Learnt strategies exhibited by one battery and one super-capacitor.

Figure 6: Learnt strategies exhibited by two super-capacitors.
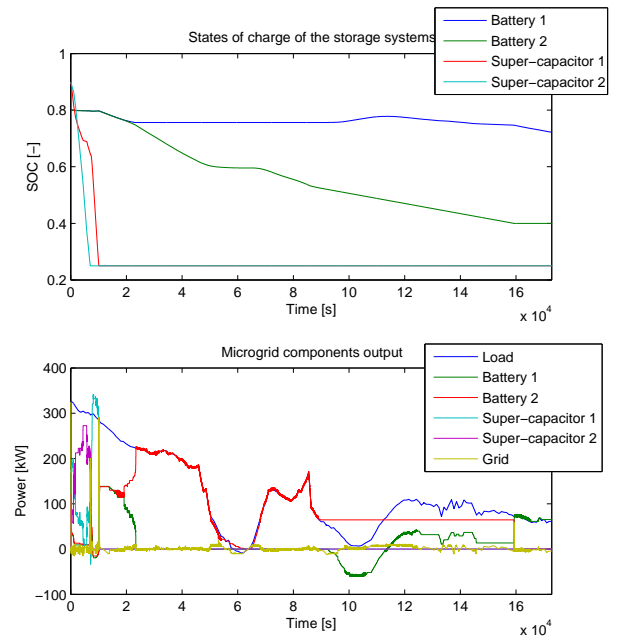
Figure 8: Learnt strategies exhibited by two batteries and two super-capacitors.

6. CONCLUSION AND FUTURE WORK

This paper has presented a multi-agent reinforcement learning algorithm, and has shown that it can be used for controlling a microgrid in a fully distributed manner, through multiple storage systems, and without requiring any centralized rule-based control. Simulation results have demonstrated both the feasibility and the reliabilty of the pro-

posed approach. This work is a preliminary step towards the elaboration of robust control mechanisms in microgrids involving RES and energy storage, and as such it can certainly be improved. Future work includes the investigation of actor-critic algorithms and bayesian learning methods for a smoother control of the storage systems (instead of using an algorithm that needs to discretize the action space like Q-Learning), the application of reinforcement learning techniques to other control problems related to microgrids, and the experimental validation on physical microgrids.

7. REFERENCES

[1] J.S. Albus. A new approach to manipulator control: The cerebellar articulation model controller (cmac). *Journal of Dynamic Systems, Measurement and Control*, 97(3), 1975.

[2] L. Buşoniu, R. Babuška, B. De Schutter, and D. Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2010.

[3] Y.-H. Chang, T. Ho, and L.P. Kaelbling. Mobilized ad-hoc networks: A reinforcement learning approach. In *ICAC*, pages 240–247, 2004.

[4] Z. Chongjie, S. Abdallah, and V.R. Lesser. Integrating organizational control into multi-agent learning. In *AAMAS*, pages 757–764, 2009.

[5] A.L. Dimeas and D. Hatziargyriou. Multi-Agent Reinforcement Learning for Microgrids. In *IEEE Power and Energy Society General Meeting*, pages 1–8, 2010.

[6] T. Konnerth, J. Chinnow, D. Grunewald, and S. Kaiser. Integration of Simulations and MAS for Smart Grid Management Systems. *Agent Technologies for Energy Systems*, 2012.

[7] L. Kuyer, S. Whiteson, B. Bakker, and N.A. Vlassis. Multiagent reinforcement learning for urban traffic control using coordination graphs. In *ECML/PKDD (1)*, pages 656–671, 2008.

[8] M.J. Mataric. Reinforcement learning in the multi-robot domain. *Auton. Robots*, 4(1):73–83, 1997.

[9] S.D.J. McArthur, E.M. Davidson, V.M. Catterson, A.L. Dimeas, N.D. Hatziargyriou, F. Ponci, and T. Funabashi. Multi-agent systems for power engineering applications–Part I: Concepts, Approaches, and Technical Challenges. *IEEE Transactions on Power Systems*, 22(4):1743–1752, 2007.

[10] S.D.J. McArthur, E.M. Davidson, V.M. Catterson, A.L. Dimeas, N.D. Hatziargyriou, F. Ponci, and T. Funabashi. Multi-agent systems for power engineering applications–Part II: Technologies, Standards, and Tools for Building Multi-agent Systems. *IEEE Transactions on Power Systems*, 22(4):1753–1759, 2007.

[11] S. Miller, S.D. Ramchurn, and A. Rogers. Optimal Decentralized Dispatch of Embedded Generation in Smart Grids. In *AAMAS*, pages 757–764, 2009.

[12] R. Roche, B. Blunier, A. Miraoui, V. Hilaire, and A. Koukam. Multi-agent systems for grid energy management: A short review. In *Proc. IECON 2010 - 36th Annual Conf. IEEE Industrial Electronics Society*, pages 3341–3346, 2010.

[13] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations*. Cambridge University Press, 2008.

[14] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, March 1998.

[15] N. Tesla. Transmission of power: polyphase system : Tesla patents. In *Westinghouse Electric and Manufacturing Co.*, 1893.

[16] P. Vytelingum, S.D. Ramchurn, A. Rogers, and N.R. Jennings. Agent-based homeostatic control for green energy in the smart grid. In 9th International Joint Conference on Autonomous Agents & Multi Agent Systems, Toronto, Canada, 2010.

[17] P. Vytelingum, T. Voice, S.D. Ramchurn, A. Rogers, and N.R. Jennings. Agent-based micro-storage management for the smart grid. In *AAMAS*, pages 39–46, 2010.

[18] G. Weiss. *Multiagent Systems, A Modern Approach to Distributed Artificial Intelligence*. MIT Press, 1999.

[19] M. Wooldridge. *An Introduction to MultiAgent Systems*. Wiley and Sons, 2002.

[20] J. Wu, X. Xu, P. Zhang, and C. Liu. A novel multi-agent reinforcement learning approach for job scheduling in grid computing. *Future Generation Comp. Syst.*, 27(5):430–439, 2011.