On-line building energy optimization using deep reinforcement learning

Citation for published version (APA):

DOI:
10.1109/TSG.2018.2834219

Document status and date:
Published: 01/07/2019

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 19. Mar. 2020
On-line Building Energy Optimization using Deep Reinforcement Learning

Elena Mocanu, Decebal Constantin Mocanu, Phuong H. Nguyen, Antonio Liotta, Michael E. Webber, Madeleine Gibescu, J.G. Slootweg

Abstract—Unprecedented high volumes of data are becoming available with the growth of the advanced metering infrastructure. These are expected to benefit planning and operation of the future power system, and to help the customers transition from a passive to an active role. In this paper, we explore for the first time in the smart grid context the benefits of using Deep Reinforcement Learning, a hybrid type of methods that combines Reinforcement Learning with Deep Learning, to perform on-line optimization of schedules for building energy management systems. The learning procedure was explored using two methods, Deep Q-learning and Deep Policy Gradient, both of them being extended to perform multiple actions simultaneously. The proposed approach was validated on the large-scale Pecan Street Inc. database. This highly-dimensional database includes information about photovoltaic power generation, electric vehicles as well as buildings appliances. Moreover, these on-line energy scheduling strategies could be used to provide real-time feedback to consumers to encourage more efficient use of electricity.

Index Terms—Deep Reinforcement Learning, Demand Response, Deep Neural Networks, Smart Grid, Strategic Optimization

I. INTRODUCTION

THERE is an energy transition underway since the start of the millennium, comprised primarily of a push towards replacing large, fossil-fuel plants with renewable and distributed generation. It results in increased uncertainty and complexity in both the business transactions and in the physical flows of electricity in the smart grid. Because the built environment is the largest user of electricity, a deeper look at building energy consumption holds a promise for improving energy efficiency and sustainability. Understanding such individual consumption behavior based on the knowledge transfer from the fusion of extensive data collected from the Advanced Metering Infrastructure (AMI) is an essential step to optimize building energy consumption and consequently the effects of its use.

This work is motivated by the hypothesis that an optimal resource allocation of end-user patterns based on daily smart electrical device profiles could be used to smoothly reconcile differences in future energy consumption patterns and the supply of variable sources such as wind and solar [1]–[3]. It is expected that a cost minimization problem could be solved to activate real-time price responsive behavior [4]. A wide-range of methods have been proposed to solve the building energy and cost optimization problems, including linear and dynamic programming, heuristic methods such as Particle Swarm Optimization (PSO), game theory, fuzzy methods and so on [1]–[7]. Therein, both centralized and decentralized solutions exist, but they fail to consider on-line solutions for large-scale, real databases [6]. More concretely, any time when an optimization is needed, these methods have to compute completely or partially all the possible solutions and to choose the best one. This procedure is time consuming. In the big data era, more and more machine learning methods appear to be suitable to overcome this limitation by automatically extracting, controlling and optimizing the electrical patterns. This can be done by performing successive transformation of the historical data to learn powerful machine learning models to cope with the high uncertainty of the electrical patterns. Then, these models will be capable of generalization and they could be exploited in an on-line manner (i.e. few milliseconds) to minimize the cost or the energy consumption in newly encountered situations. Among all these machine learning models, the ones belonging to the Reinforcement Learning (RL) area are the most suitable for the cost minimization problem, as they are capable to learn an optimal behavior, while the global optimum is not known.

Thus, in the remaining of this paper we focus on RL methods, such as Q-learning [8], and their latest developments. The building environment is modeled using a Markov Decision Process [9] and it can be used to find the best long-term strategies. Prior studies showed that RL methods are able to solve stochastic optimal control problems [10] in the power system area as well as an energy consumption scheduling problem [11] with dynamic pricing [12]. A batch reinforcement learning method was introduced in [13] to schedule a cluster of domestic electric water heaters, and further on applied for smart home energy management [14]. Owing to the curse of dimensionality, these methods fail for large-scale problems. More recently, there has been a revival of interest in combining deep learning with reinforcement learning. Lately, in 2015, an application of Q-learning to deep learning has been successful at playing Atari2600 games at expert human levels [15]. In 2016, another one has defeated for the first time in history the world champion at the game of Go. Complementary with our work, Francois-Lavet et al. has proposed the use of Deep Q-learning for storage scheduling.
in microgrids [16]. The above methods represent the starting point of a new research area, known as Deep Reinforcement Learning (DRL), which has evolved through the intersection of reinforcement learning and neural networks. At the same time, in our previous work, we showed that Reinforcement Learning using Deep Belief Networks for continuous states estimation can successfully perform unsupervised energy prediction [17].

Our contribution: In this paper, inspired by the above research developments, we propose for the first time the use of the Deep Policy Gradient method, as part of Deep Reinforcement Learning algorithms, in the large-scale physical context of smart grid - smart building, as follows.

- We propose a new way to adapt DRL algorithms to the smart grid context, with the aim of conceiving a fast algorithm to learn the electrical patterns and to optimize on-line either the building energy consumption or the cost.
- We investigate two DRL algorithms, namely Deep Q-learning (DQN) [15] and Deep Policy Gradient (DPG).
- DPG in its current form is capable to take just one action at a specific time. As in the building context multiple actions have to be taken at the same moment, we propose a novel method to enhance DQN with the capability of handling multiple actions simultaneously.

We evaluate our proposed methods on the PecanStreet database at both the building and aggregated level. In the end, we prove that our proposed methods are able to efficiently cope with the inherent uncertainty and variability in the generation of renewable energy, as well as in the peoples’ behavior related with their use of electricity (i.e. charging of electric vehicles). Specifically, we show that the enhanced DPG is more appropriate to solve peak reductions and cost minimization problems than DQN.

The remaining of this paper is organized as follows. Section II describes the problem formulation and Section III we introduce the background and preliminary concepts. Section IV describes our proposed method followed by implementation details in Section V. Results and discussions are provided in Section VI. Finally, we conclude with some directions for future research.

II. Problem formulation

In this context, we aim to reduce load peaks as well as to minimize the cost of energy. Let $B$ denote the set of buildings, such that $B_i \in B, \forall i \in N$ representing the index of the building analyzed. The total building energy consumption $E_i$ is a sum over all power generation $P^{++}$ and consumption in a specific interval of time $\Delta t$. Therein, based on the shifting capabilities of appliances present in a building we differentiate between flexible power $P^F$, e.g. electric devices $d \in \{1, \ldots, m\}$, and fixed consumption $P^-$.

a) Cost minimization problem: In this paper, we assume two price components over the space of $B$, such that $\lambda^+_t$ is the price value set by the utility company for the time-slot $t$ and $\lambda^-_t$ represents the price value at which the utility company buys energy from end-users at time-slot $t$. Therefore, the optimal cost associated with customer $i$ at time $t$ for an optimization time horizon $T$ can be calculated as

$$\min \sum_{t=1}^{T} \left( \sum_{i=1}^{n} \lambda^+_t \cdot P^+_i(t) + \sum_{i=1}^{n} \lambda^-_t \cdot \left( P^-_i(t) + \sum_{d=1}^{m} a_{i,d,t} \cdot P^-_{i,d}(t) \right) \right)$$

s.t. \( \sum_{t=1}^{T} P^+_i(t) \leq E_i, \forall i \in N, \forall t \in N, \)

\( \sum_{t=1}^{T} P^-_i(t) \leq P^F_i, \forall d \in N, \forall t \in N, \)

\( a_{i,d,t} \in \{0, 1\}, \forall a \in A, \forall i \in N, \forall d \in N, \forall t \in N, \)

\( \lambda^+_t \geq 0, \forall t \in [1 : T], \lambda^-_t \geq 0, \forall t \in [1 : T] \in N. \)

where $a_{i,d,t} = 1$ if the electrical device is on at that specific moment in time, and 0 otherwise. Please note that, in our proposed method, computing $a_{i,d,t}$ is equivalent with the estimation of the actions (see Fig.1).

b) Peak reduction problem: In the special case of constant price, for electricity generation and consumption, with $\lambda^+_t = \lambda^-_t$, the cost minimization problem becomes a peak reduction problem, defined as

$$\min \sum_{t=1}^{T} \left( \sum_{i=1}^{n} P^+_i(t) - \sum_{i=1}^{n} \left( P^-_i(t) + \sum_{d=1}^{m} a_{i,d,t} \cdot P^-_{i,d}(t) \right) \right)$$

Consequently, the constraints following Eq.1 will remain valid for both problems. However, based on the differences between different types of electrical devices the full range of constraints is larger as explained in the next sections.

c) Electrical device constrains: We are assuming three types of consumption profiles. Firstly, we consider the time-scaling load. In respect to this we confine our analysis to the air conditioning load ($d_{AC}$), as a representative part of a larger set of electrical devices in every building which could be switched on-off for a limited number of times during an optimization horizon, e.g. lights, television, refrigerator. Prior studies show that short-term air conditioning curtailments have a negligible effect on end-user comfort [18]. Secondly, we include the time-shifting load, also called deferrable load, that must consume a minimum amount of power over a given time interval. Therein, we model the dishwasher ($d_{DW}$) as an uninterruptible load, which requires a number of consecutive time steps. Finally the electric vehicle($d_{EV}$) was considered as both a time scaling and shifting load. A more rigorous formulation of the building electrical components and their associated constraints could be found in [19]. In our case, a complementary probabilistic perspective over the time dependent devices constrains $a_{d,t}$ give us the following assumptions:

A 1: For all $d$, with $P^d_{-}$ time-scaling loads, there $\exists \delta_d \in \mathbb{R}_+$ constants over the optimization horizon such that

$$\sum_{t=1}^{T} P^d_{-} \leq \delta_d \quad \text{if} \quad p(P^d_{-} = 0|t) \in (0, 1]$$

$$\sum_{t=1}^{T} P^d_{-} = \delta_d \quad \text{if} \quad p(P^d_{-} = 0|t) = 0 \quad (8)$$

where $p(P^d_{-} = 0|t)$ is the probability of the electrical device $d$ to be active at any moment in time $t$, for all $t = [1 : T] \in N.$

A 2: For all $d$, with $P^d_{-}$ time-shifting loads, there $\exists \delta_d$ constants such that $\sum_{t=1}^{T} P^d_{-} = \delta_d$, for all $t = [1 : T] \in N.$
**Observation 1:** In this paper, $P^+$ (e.g. PV generation) is considered a non-curtailable resource.

**Observation 2:** All electrical vehicles, $d_{EV}$, and their associated consumption $P^d$, were considered as time scaling and shifting loads working under the conditions imposed by Assumption 1 and 2.

### III. BACKGROUND AND PRELIMINARIES

In this section, we provide a brief overview of reinforcement learning, Markov decision formalism, and deep neural networks.

#### A. Reinforcement Learning

In a Reinforcement Learning (RL) context, an agent learns to act using a (Partial Observable) Markov Decision Process (MDP) formalism. MDPs are defined by a 4-tuple $(S, A, T(·, ·), R(·, ·))$, where:

- $S$ is the state space, $∀s ∈ S$.
- $A$ is the action space, $∀a ∈ A$.
- $T : S × A × S → [0, 1]$ is the transition function given by the probability that by choosing action $a$ in state $s$ at time $t$, the system will arrive at state $s'$ at time $t + 1$, such that $p_a(s, s') = p(s_{t+1} = s' | s_t = s, a_t = a)$, and
- $R : S × A × S → R$ is the reward function, were $R_a(s, s')$ is the immediate reward received by the agent after it performs the transition to state $s'$ from state $s$.

The agent aims to optimize a stochastic policy $π : S × A → R$. Under structure assumption of the environment (i.e., finite states and actions) the Markov decision problem is typically solved using dynamic programming. However, in our built environment, the model has a large (continuous) states space. Therefore, the state space is given by the building energy consumption and price at every moment in time, while the action space is highly dependent on the electrical device constrains. The success of every action $a$ is measured by a reward $r$. Learning to act in an environment will make the agent to choose actions to maximize future rewards. The value function $Q^π(s, a)$ is an expected total reward in state $s$ using action $a$ under a policy $π$. Currently, one of the most popular reinforcement learning algorithm is Q-learning [8].

#### B. Deep Neural Networks

The topology of a Deep Neural Network (DNN) architecture is based on multiple layers of neurons. In general, a neuron is a non-linear transformation of the linear sum of its inputs. The first layer models directly the data. A hidden layer in the neural network architecture is built as an array of neurons taking the inputs from the previous layer. The activation function of a neuron on top of $k$ stacked layers in the architecture is using composite functions, such as $x ⊗ h_1 ⊗ h_2 ⊗ ⋯ ⊗ h_k$.

In 2011, it was shown that supervised training of a very deep neural network with hard non-linearities is faster if the hidden layers are composed of Rectified Linear Units (ReLU) [20]. Recently, the logistic sigmoid and the hyperbolic tangent activation are outperformed by ReLU [15], [21], [22]. Formally, ReLU is defined as a function $f(x_i) = \max(0, x_i)$.

![Fig. 1. The closed loop general architecture of Deep Reinforcement Learning, built as a combination of Reinforcement Learning and Deep Neural Network.](image)

In this section, we propose the use of Deep Reinforcement Learning (DRL) as an on-line method to perform optimal building resource allocation at different levels of aggregation. The general architecture of our proposed method is depicted in Fig. 1. DRL (RL combined with DNNs of $k$ hidden layers) can learn to act better than the standard RL by automatically extracting patterns, such as those of electricity consumption. Overall, we can represent the DNN method, from a very general perspective, as a black box model with good generalization capabilities over a given input distribution as follows:

$$\frac{\text{Input data}}{\text{DNN}(k)} \rightarrow \frac{\text{Output}}{\text{Data estimation}}$$

In the remaining of this section we will introduce two DRL methods, namely Deep Q-learning (DQN) and Deep Policy Gradient (DPG).
DRL = \begin{cases} \text{Deep Q-learning} \\
\text{Deep Policy Gradient} \end{cases}

In contrast to value-based methods (e.g. DQN), policy-based model free methods (e.g. DPG) directly parameterize the policy \( \pi(a|s; \theta) \) and update the parameters \( \theta \) by performing, typically approximate, gradient ascent on the expected long-term reward \[^{[24]}\].

A. Deep Q-learning (DQN)

Learning in DRL is done as follows. The DNN is trained with a variant of the Q-learning algorithm, using stochastic gradient descent to update its parameters \[^{[15]}\]. Firstly, the value-function from the standard RL algorithm is replaced by a deep Q-network with parameters \( \theta \), given by the weights and biases of DNN, such that \( Q(s,a,\theta) \approx Q^*(s,a) \). This approximation is used further to define the objective function by mean-squared error in Q-values:

\[
L(\theta) = \mathbb{E} \left[ (r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \theta) - Q(s_t, a_t, \theta))^2 \right] \tag{11}
\]

Leading to the following Q-learning gradient

\[
\frac{\partial L(\theta)}{\partial \theta} = \mathbb{E} \left[ (r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \theta) - Q(s_t, a_t, \theta)) \frac{\partial Q(s_t, a_t, \theta)}{\partial \theta} \right] \tag{12}
\]

Usually, this standard Q-learning algorithm used in synergy with neural networks oscillates or diverges, mainly because data are sequential. To overcome this limitation of correlated data and non-stationary distributions, we use an experience replay mechanism which randomly samples previous mini-batch of transitions \( (s_t, a_t, r_t, s_{t+1}) \) from the dataset \( \mathcal{D} \), and therefore smoothes the training distribution over many historical data. It is straightforward to integrate the above Deep RL approach into Eq.1. The binary action vector \( a_t \in A \) is augmented to \( \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \theta) \) and therefore \( \sum_{d=1}^D a_t P_{d,t} \) is optimally controlled. Specifically, rather than enforcing the constraints on the time window required by a specific device \( d \) and the comfort of end-users considered in \( A1, Eq.8 \), our idea is to encapsulate them in the reward function, \( r_t(\lambda^+, \lambda^-, P_{t,d}) \), which is further detailed in Section IVA.

B. Deep Policy Gradient (DPG)

Recently, it has been shown that policy gradient methods are able to decrease the time needed for convergence in continuous games \[^{[24],[25]}\]. From an architectural perspective, the neurons from the output layer of the DNN (with parameters \( \theta \)) corresponding to DPG, instead of estimating the \( Q(s_t, a_t, \theta) \), \( \forall a \in A \) as in DQN, they estimate the probability to take action \( a \) in a specific state \( s_t \), such as \( p(a|s_t, \theta) \), \( \forall a \in A \). This offers a clear advantage to DPG over DQN when there is a need to perform multiple actions simultaneously, as all actions can be sampled and executed simultaneously in the game using their own probability.

In the policy gradient context, the approximate optimization problem defined in Eq. \[^{[1]}\] or Eq. \[^{[7]}\] is equivalent of maximizing the total expected reward of a parameterized model under a policy \( \pi \), as follows

\[
\text{maximize } \mathbb{E}_{x \sim p(x|\theta)} [R|\pi] \tag{13}
\]

In the DPG context, the parameterized model is the DNN. Thus, the DNN becomes a probability density function over its inputs (the game states), i.e. \( f(x) \), leading Eq. \[^{[13]}\] to the following optimization problem

\[
\text{maximize } \mathbb{E}_{x \sim p(x|\theta)} [f(x)] \tag{14}
\]

As shown in \[^{[26],[27]}\], the unbiased gradient estimation uses \( f(x) \) as a score function yielding

\[
\nabla_\theta \mathbb{E}_x[f(x)] = \nabla_\theta \int dx \ p(x|\theta)f(x) = \int dx \ \nabla_\theta p(x|\theta)f(x) = \int dx \ p(x|\theta) \nabla_\theta \log p(x|\theta)f(x) = \mathbb{E}_x[f(x) \nabla_\theta \log p(x|\theta)] \tag{15}
\]

where \( \frac{\partial}{\partial \theta} = \nabla_\theta \) denote the first-order partial derivative over the output data. Intuitively, to solve the gradient of Eq. \[^{[15]}\] first we have to take samples of \( x_t \sim p(x|\theta) \) and to compute the estimated gradient, such that \( \frac{\partial}{\partial \theta} f(x_t) = f(x_t) \nabla_\theta \log p(x_t|\theta) \). Moving in the \( \hat{g}_t \) direction increases the log-probability of that particular sample \( x_t \) proportional with the reward associated with it, \( f(x_t) \). In other words, this practically shows how good is that sample. As in policy gradient the reward is available at the end of a game, these samples are collected in a trajectory, i.e. \( \tau = (s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}) \). To compute the gradient of a trajectory, we need to calculate and differentiate the density \( p(\tau|\theta) \) with respect to \( \theta \) as follows:

\[
p(\tau|\theta) = p(s_0) \prod_{t=0}^{T-1} [p(a_t|s_t, \theta) p(s_{t+1}|s_t, a_t)] \tag{16}
\]

By taking the log-probability of Eq. \[^{[16]}\], we obtain

\[
\log p(\tau|\theta) = \log p(s_0) + \sum_{t=0}^{T-1} \left[ \log p(a_t|s_t, \theta) + \log p(s_{t+1}|s_t, a_t) \right] \tag{17}
\]

Taking the derivative of Eq. \[^{[17]}\] with respect to \( \theta \) leads to

\[
\frac{\partial}{\partial \theta} \log p(\tau|\theta) = \frac{\partial}{\partial \theta} \sum_{t=0}^{T-1} \log p(a_t|s_t, \theta) \tag{18}
\]

Finally, we can write the gradient update \( \hat{g}_t^\theta \) for parameters \( \theta \) after considering a trajectory \( \tau \) as

\[
\hat{g}_t^\theta \propto \mathcal{R}_\tau \frac{\partial}{\partial \theta} \sum_{t=0}^{T-1} \log p(a_t|s_t, \theta) \tag{19}
\]
V. IMPLEMENTATION DETAILS

A. Network Architecture

Aiming to have a fair comparison between DQN and DPG, the architecture of the deep neural networks used is similar for both models and it has the following characteristics. Each reinforcement learning state (encoded in the input layer), is given by a time-window of two consecutive time steps. Thus, in the case of the peak reduction problem the input layer has 11 neurons, i.e. time step $t$, and base load, PV, AC state, EV and dishwasher at $t-1$ and $t$. Please note that with the exception of base load and generation which are fixed, the other state components are given by the dynamically adapted values (the ones obtained during the learning process) and not the initial ones measured by the smart meters. For the cost reduction problem, the input layer has an extra neuron which is used to encode the ToU tariff. Furthermore, the networks have three layers of hidden neurons, each layer having 100 neurons with Rectifier Linear Units (ReLU) as activation function.

The output layer differs for DQN and DPG. For DQN the output layer has 8 neurons, each neuron representing the Q-value of a combined action. Each combined action is a possible combination of the actions of the three flexible devices i.e. stop air conditioner ($a_1$), electric vehicle on/off ($a_2$), dishwasher on/off ($a_3$). By contrast, the DPG output layer has just three neurons, each neuron representing a device action. More precisely, it gives the probability to perform the action associated with the flexible device for the specific input layer. This is a clear advantage of DPG over DQN as it scales linearly with the number of flexible devices.

Hyper-parameters settings: In all experiments performed, the learning rate is set to $\alpha = 10^{-2}$, the discount factor to $\gamma = 0.99$, and $\eta = 0.01$. We train the models for 5000 episodes, where an episode is composed by 20 randomly chosen days. The weights update is performed after every two episodes. The final policy is kept as the output of the learning process.

B. The reward vectors for DRL

Regarding the multi-objective optimization problems solved in this paper, an accurate reward function is computed at the end of the day, instead of at each time step of the day. Thus, we derived a simple multiple-task joint reward with three reward components:

**Component 1:** For all $\tau = (s_1, a_1, r_1)$ the reward vectors will be able to control the actions of the three types of flexible consumption, and therefore the total switchable and scalable load in a household $\sum_{d=1}^{d} a_d P_{d,t}$, using differentiated rewards

$$r_{a_1} = \begin{cases} -n_{a_1} & \text{if } n_{a_1} > 10 \\ \zeta_1 & \text{if } n_{a_1} \in [1,10] \\ \zeta_2 & \text{if } n_{a_1} < 1 \end{cases}$$

$$r_{a_2} = \begin{cases} -4|n_{a_2} - n_{a_2}^{+}| & \text{if } n_{a_2}^{+} \neq n_{a_2}^{+}, \forall n_{a_2}^{+} \in \mathbb{N} \\ n_{a_2}^{+} & \text{if } n_{a_2}^{+} = n_{a_2}^{+}, \forall n_{a_2}^{+} \in \mathbb{N} \end{cases}$$

**Component 2:** Controlling the total energy consumption defined in Eq. 7 is done as follows

$$ r = \begin{cases} -3\zeta_2 + 4[\max(P^+) - \max(P^-)] & \text{if } \max(P^-) < \max(P^+) \\ -3\zeta_1 - 1 & \text{otherwise} \end{cases} $$

Further on, shifting the consumption through the time when there is more generation Eq. 6

$$ r = \begin{cases} \frac{\zeta_1}{\gamma} - |\min(P^-)| & \text{if } \bar{P}^- < 0 \\ \frac{-\zeta_1}{\gamma} & \text{otherwise} \end{cases} $$

The control of AC under the $A_2$ in Eq. 8 is given by

$$ r = \begin{cases} \frac{\zeta_1}{\gamma} + 2[\max(P^-_{AC}) - \max(P^+)_{AC}] & \text{if } \bar{P}^-_{AC} < 0 \\ \frac{-\zeta_1}{\gamma} & \text{otherwise} \end{cases} $$

**Component 3:** Controlling the total cost $C$, defined in Eq. 9

$$ r = \begin{cases} 5|\bar{C} - \bar{C}^{+}| & \text{if } \bar{C} < \bar{C}^{+} \\ -3\zeta_1 - 1 & \text{otherwise} \end{cases} $$

The agent must learn multiple tasks consecutively with the goal of optimizing performance across all previously learned tasks. So, we used for solving Eq. 7 the Component 1 and 2 of the reward, while for Eq. 9 the Component 1 and 3.

The joint reward components could be easily generalized to perform an arbitrary number of tasks. However, the range intervals for $n_{a_1}^{+}$ and $n_{a_2}^{+}$ considered in Eq. 20 as well as the positive and negative coefficients (i.e. $\zeta_1$ and $\zeta_2$) used in Eq. 21, 22, 23, 24 are dependent on the application. Also in Eq. 20 the range of $n_{a_1}^{+}$ and $n_{a_2}^{+}$ may be enlarged if comfort limits are relaxed. Algorithm 1 exemplifies on DPG, how DRL can be implemented. We have implemented both methods, DQN and DPG, in Python.

---

\textbf{Algorithm 1 Deep Policy Gradient (DPG) - estimating Eq. 13}

1: initialize model: hyper-parameters ($\alpha$, $\gamma$, $\zeta$)
2: initialize model: DNN with random weights $\theta$
3: initialize game: first state $s$ from a random day
4: for iteration = 1 to arbitrary number do
5: sample actions $p(a_1, a_2, a_3|\theta, s)$ with DNN
6: collect probabilities $p(a_1, a_2, a_3|\theta, s)$ in $A$
7: collect hidden neurons values in $H$ from DNN
8: collect $s$ in $S$
9: execute actions $a_1, a_2, a_3$ and move to next state $s'$
10: collect reward $r$ in $R$ from game
11: if episode is finished then
12: compute discounted rewards $R_d$ from $R$
13: estimate gradients from $A + R_d, \theta, S$, and $H$ (Eq. 19)
14: update $\theta$ with the estimated gradient
15: empty $A, R, S$, and $H$
16: end if
17: if current day ends then
18: reset game: first state $s'$ from a random day
19: end if
20: set $s \leftarrow s'$
21: end for

---

The number of neurons in the output layer of DQN is exponentially correlated with respect to the number of flexible devices.
A. Data set characteristics

1) Buildings pattern: To validate our proposed method we used the Pecan Street dataset. The disaggregated customer energy data contains up to 90 million unique electricity consumption records per day, which are used in order to build specific device patterns. Figure 2(a), 2(b), and 2(c) show three different building patterns averaged over the year 2015 with 15 minutes resolution. In these patterns the solar generation uncertainty as well as the people behavior characteristics are notable, e.g. even if all three buildings have an electric vehicle, just in the case of the third building, Fig. 2(c), it is used frequently. In our experiments, we have used the data between 27th October 2012 and 3rd September 2016.

2) Price data: We use the time-of-use (ToU) tariff provided by the local grid operator Austin Energy for customers who live inside the City of Austin, Texas. The summer rates are composed by on-peak, mid-peak and off-peak hours and the winter tariff has the mid-peak and off-peak hours components. There is also a difference between weekend and working-days tariff. Additionally, the self-generating customers are receiving an amount that is being paid by the utility for solar generation, called the value of solar tariff (VOST).

B. Numerical results - Peak reduction problem

The numerical results in terms of peak reduction at the single building level are shown in Table I and Figure 2 for three different buildings ($B_1$, $B_{11}$ and $B_{111}$), over one year with 15 minutes resolution.

C. Numerical results - Cost minimization problem

The results for the cost minimization problem are summarized in Table I. The difference between the cost minimization solutions obtained for every building correlated with their average electrical patterns (see Figure 2) give as a first indication about the individual capabilities of the end-users to adopt a more conservative behavior.

VI. Results and Discussion

In this section, we validate our proposed methods and we analyze their performance on a large real-world database recorded by Pecan Street Inc. First, the database is described. Then, numerical results are given for both problems, i.e. peak reduction and cost minimization for various number of buildings.

As it can be observed in Table II and in Figure 2, a secondary advantage of solving the cost minimization problem is its impact on solving of the peak reduction problem also. Therein, the best results in terms of both, peak reduction and cost reduction, are obtained for building $B_{111}$ using DPG.

D. Scalability and learning capabilities of DRL

To test whether good estimations occur in practice and at scale, we investigate the performance of our proposed methods, in three cases with different numbers of customers using data from the Pecan Street smart grid test-bed. Specifically, we are investigating and analyzing the corresponding results using DQN and DPG methods for 10, 20 and 48 buildings, respectively. Tables III and IV show that our proposed approach is scalable for both, peak reduction and cost minimization, respectively. More than that, they show that at the aggregated level, when more customers are taking the cost minimization problem into consideration, this solves implicitly also the peak reduction problem. Same as at the building level, DPG is more stable than DQN, and achieves a better performance. Overall, in the case of the cost reduction problem for 48 buildings, DPG reduces the peak with 26.3% and minimizes the cost with 27.4%, while DQN reduces the peak with just 9.6% and minimizes the cost with just 14.1%. To visualize how DPG performs, in Figure 3 we depict the unoptimized and the optimized annualized energy costs for each of the 48 buildings.

To visualize how DPG performs, in Figure 3 we depict the unoptimized and the optimized annualized energy costs for each of the 48 buildings. We can observe that the buildings behave very differently and in some cases DPG is capable to halve the yearly cost, while in other cases it succeeds to reduce the cost with just a few percentage points.

Convergence capabilities of DPG: The convergence is assessed through many iterations over episodes. For example, the
learning capabilities of DPG method in terms of peak reduction and their corresponding reward function for a building are shown in Figure 4. Each episode represents an average value over 20 randomly chosen days. Initially, we may observe that the reward increases fast, while after about 1000 episodes the reward, as expected, increases much slower. Therefore, after approximately 1000 episodes the average peak value and the optimized average peak value using the Deep Policy Gradient method converge. Still, the long-term reward expectation, as was expressed in Eq. (13), is increasing until approximately 2500 episodes.

**Computational time requirements:** Both DRL variants have the advantage of handling naturally much larger continuous state spaces, leading to better performance. In comparison with heuristic methods (e.g. PSO), after DRL learns how to act, it can make decisions (e.g. choosing the optimal control action)
in a few milliseconds, while PSO needs to re-run the costly optimization process for every decision.

VII. CONCLUSIONS

In this paper, we proposed the use of Deep Reinforcement Learning, as a hybrid method which combines Reinforcement Learning with Deep Learning, with the aim of conceiving an on-line optimization for the scheduling of electricity consuming devices in residential buildings and aggregations of buildings. We have shown that a single agent, empowered with a suitable learning algorithm, can solve many challenging tasks. We proposed two optimization methods, Deep Q-learning and Deep Policy Gradient, to solve the same sequential decision problems at both the building level and the aggregated level. At both levels, we showed that Deep Policy Gradient is more suited to perform on-line scheduling of energy resources than Deep Q-learning. We explored and validated our proposed methods using the large Pecan Street database. Both methods are able to successfully perform either the minimization of the energy cost or the flattening of the net energy profile. For the minimization of the energy cost, a variable electricity price signal is investigated to incentivize customers to shift their consumption to low-price, off-peak periods.

ACKNOWLEDGMENT

This research has been partly funded by the NL Enterprise Agency under the TKI SG-BEMS project of Dutch Top Sector and by the European Union’s Horizon 2020 2020 project INTER-IoT (grant number 687283).

REFERENCES
