UCB1 based reinforcement learning model for adaptive energy management in buildings

Rui Andrade[1], Tiago Pinto[1,2], Isabel Praça[1], Zita Vale[1]
[1]GECAD – Research group, Institute of Engineering, Polytechnic of Porto (ISEP/IPP), Porto, Portugal
{rfaar, tmcfp, icp, zav}@isep.ipp.pt
[2]BISITE Research Centre, University of Salamanca (USAL), Calle Espejo, 12, 37007 Salamanca, Spain
tpinto@usal.es

Abstract. This paper proposes a reinforcement learning model for intelligent energy management in buildings, using a UCB1 based approach. Energy management in buildings has become a critical task in recent years, due to the incentives to the increase of energy efficiency and renewable energy sources penetration. Managing the energy consumption, generation and storage in this domain, becomes, however, an arduous task, due to the large uncertainty of the different resources, adjacent to the dynamic characteristics of this environment. In this scope, reinforcement learning is a promising solution to provide adaptiveness to the energy management methods, by learning with the on-going changes in the environment. The model proposed in this paper aims at supporting decisions on the best actions to take in each moment, regarding buildings energy management. A UCB1 based algorithm is applied, and the results are compared to those of an EXP3 approach and a simple reinforcement learning algorithm. Results show that the proposed approach is able to achieve a higher quality of results, by reaching a higher rate of successful actions identification, when compared to the other considered reference approaches.

Keywords: adaptive learning, energy management in buildings, EXP3, reinforcement learning, UCB1

1 Introduction

During the last decade a centralized approach is being used in energy (and more specifically, in electricity) markets. Energy consumers are only connected to energy producers and thus the energy distribution is all cantered around one production point [1]. Alternatives to this traditional energy market are emerging and future energy markets are evolving towards a more distributed model. The biggest difference is the decentralization of the energy production, which has originated a new type of role in the

[1] This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 641794 (project DREAM-GO) and from Project SIMOCE (ANI|P2020 17690)

market: besides energy producers and energy consumers, consumers who are also actively producing energy become part of the market, and are referred to as "prosumers" [2]. With the emergence of prosumers, new possibilities open in the market, enabling the emergence of distributed energy markets. These markets are categorized by connecting consumers and prosumers in a grid while still being connected to centralized producers.

The new characteristics, and consequently the new role, of consumers in the energy ecosystem, force these players to pursue more intelligent and adaptive energy management solutions, in order to be able to take as much advantage from the environment as possible. House or building Energy Management Systems (EMS) are designed to manage the energy consumption and generation within the buildings, respond to energy requests from the grid and minimize the energy bill, while at the same time taking into consideration the comfort levels within the users. The objective is that the EMS will use a minimal amount of energy and still keep the user satisfied [3].

Creating an EMS for smart-houses is the aim in [4]. The system takes into consideration five a total of five possible electrical loads: fixed loads, lights, dishwasher, washing machine and dryer. This EMS also considers the desired temperature for the house which is set by the user. In [5] a similar concept is explored. The proposed EMS has the ability to control the energy consumption within the building and aims to shift the electricity usage depending on the current electricity prices in order to lower the electric costs. The research presented in [6] proposes an EMS for a smart-house focused on managing renewable energy sources such as solar and wind energy generation, Hybrid electric vehicles with batteries, supercapacitors (SCs), and the house itself. The system makes use of maximum power point tracking (MPPT) algorithms to control and optimize the energy storage, solar generation and wind generation.

In order to improve its performance, EMS should collect data and make changes in its behavior when necessary [3]. Artificial intelligence, and machine learning in particular, are promising solutions to improve the processes of self-evaluation and adaptation [7]. Some relevant work has already been accomplished in this domain, e.g. by using reinforcement learning [8], but much has yet to be explored in order to enable an effective and dynamic adaptation of EMS to the constantly changing environment and uncertainty associated to energy resources, such as consumption habits, renewable generation and market prices .

This paper proposes a novel model based on a Markov decision process for decision-making in the context of a smart house. A reinforcement learning approach is presented, in which the goal is to learn the best action for the user to take, considering the expected state of energy resources at each time. An adaptation of the Upper Confidence Bound (UCB1) algorithm (a well-known algorithm for multi-armed bandits [9], is applied to solve de modelled problem. Results are compared to those achieved by the Exponential-weight algorithm for exploration and exploitation (Exp3), also a commonly used algorithm for adversarial bandits problems [10]; and by a simple reinforcement learning algorithm that simply updates the confidence value in each *action-state* pair according to the given reinforcement value at each time. A case study using real data is presented, and shows that the proposed UCB1 based algorithm is able to achieve better results than the other reference algorithms.

2 Proposed Approach

The proposed approach aims at enabling a house EMS to learn and adapt to the dynamic changes in the environment. The objective is to learn which is the best action *a* to perform at each time *t*, considering the current state *s* of the surrounding environment. The proposed model considers a generic set of actions, which can be instantiated depending on each specific application scenario, e.g. as presented in the case study. These may represent the action to consume the energy stored in the battery, to sell the generated energy to the network, etc. Performing an action in a current state results in a specific reward for time *t*, which represents the value that this action brings in the corresponding state. This process is called a Markov Decision Process (MDP) for decision-making.

Different reinforcement learning algorithms can operate on top of an MDP model. This process can be described in a simple number of steps. A state is given as input, an action is selected and performed, the reward given to the action is used to determine how good that action is in that state and the resulting state is given as the new input and the cycle continues [11, 12], this set of steps is shown in Fig. 1.

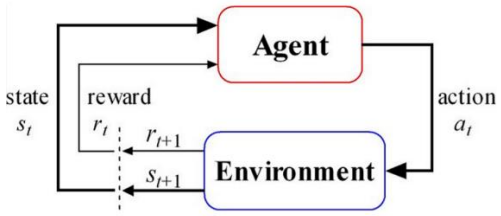

Fig. 1. Reinforcement learning algorithms learning process, [12]

Reinforcement learning algorithms are in constant learning and can adapt to changing environments, referred by [11] as non-stationary environments. This makes reinforcement learning an ideal option for problems that may need to deal with unpredictable changes, such as energy pricing.

There are many reinforcement learning algorithms, with different ways of processing the received rewards, which makes them have distinct results depending on the problem where they are being applied [12].

Multi-armed bandit algorithms [9] are reinforcement learning algorithms that try to find the best of playing in multiple slot machines, also known as “one-armed bandit” or simply “arm”. These may have a biased reward probability distribution picked a priori. Searching for the best arm is called the exploratory phase, and using that information to make the biggest possible profit is the exploitation phase. The aim on these algorithms is to try the different options until enough confidence is built on what option is the best. Upper Confidence Bound (UCB) algorithms are usually applied to solve this problem.

This work presents an adaptation of a UCB algorithm to solve the envisaged MDP problem. UCB1 combines the exploratory phase and an exploitation phase, in a way that the algorithm choses one of those two approaches in each iteration depending on

the received rewards. The algorithm also has a concept of a regret function that is used to try to find the loss correspondent with each arm. The arm with the lowest value in the regret function is considered the best option.

$$8 \cdot \left[\sum_{i:\mu_i<\mu^*} \left(\frac{\ln t}{\Delta_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{i=1}^{K} \Delta_i \right). \quad (1)$$

Exponential-weight algorithm for exploration and exploitation (Exp3) is an algorithm for adversarial bandits problems, which is similar to the multi-armed bandit problem, but instead of fixed distributions, adversarial bandits follow the idea that an "adversary" is changing the rewards distributions in each time step algorithms [10]. The algorithm uses a parameter called egalitarianism, $\gamma \in [0, 1]$, this parameter is used to balance the exploration. The objective with this parameter is to determine the amount of time $(1 - \gamma)$, in which the algorithm is doing a weighted exploration/exploitation. The weighted exploration/exploitation is based on the current estimated reward, and the rewards received from the weighted exploration/exploitation are immediately used to update the correspondent arm's weight with (2) where i indicates the arm, and Pi represents the received reward for the arm, and (3) is used to calculate the current probability for each arm.

$$w_{i,t} = w_{i,t-1} \cdot e^{\gamma \cdot \frac{\rho_i}{p_{i,t} \cdot K}} \quad (2)$$

$$p_{i,t} = (1 - \gamma) \frac{w_{i,t}}{\sum_{j=1}^{K} w_{j,t}} + \gamma \cdot \frac{1}{K}. \quad (3)$$

A simple reinforcement learning algorithm is also considered in this work, for benchmarking comparison purposes. This algorithm considers the updating of the confidence value in each action *a* in time *t*, through a direct increment of the confidence value *C* according to the reinforcement value *R*. The update of the values is expressed by (4).

$$C_{a,\ t+1} = C_{a,t} + R_{a,t} \quad (4)$$

3 Case Study

This case study aims at assessing the proposed approach and comparing the performance of the different reinforcement learning algorithms, by using a practical application case. The MDP model is instantiated as follows. 5 states are considered, combining different possibilities regarding the energy consumption, generation and retail market price, as shown in Table 1. The considered values for these three components are based on real data of a house studied in [13]. Each of the states is active during a specific period in each 24 hours cycle. The probabilities of transitions between states at the end of each period are also specified and presented in Table 1. These define the probability of each state occurring, based on a random distribution.

Conversely, 5 possible actions are also considered. These are presented in Table 2, together with the considered rewards for each State-Action pair. The proposed MDP model is executed for 10000 iterations for each of the three considered algorithms.

Table 1. Considered states

State	Description	Transition states and probabilities
GgtC_PL	Generation greater than Consumption with low Price	GgtC_PH 50% GetC 50%
GgtC_PH	Generation greater than Consumption with high Price	GgtC_PH 66.7% GgtC_PL 33.3%
GltC_PL	Generation less than Consumption with low Price	GltC_PL 75% GltC_PH 25%
GltC_PH	Generation less than Consumption with high Price	GgtC_PL 10% GltC_PL 10% GltC_PH 80%
GetC	Generation equal to Consumption (Price independent)	GgtC_PL 50% GltC_PL 50%

Table 2. Possible actions and rewards for each *s-a* pair

	GgtC_PL	GgtC_PH	GltC_PL	GltC_PH	GetC
Buy	0.1	0	1	0.6	0.2
Sell	0.8	1	0	0.1	0.2
Store	1	0.8	0.1	0	0.2
Use_stored	0.2	0.1	0.8	1	0.2
Consume	0.4	0.4	0.5	0.8	1

Fig. 2, Fig. 3 and Fig. 4 show the evolution of the rewards for each action over time, for the three algorithms. If the algorithm is able to learn ideally, all actions should converge to the reward value of 1, which is maximum reward value for each action.

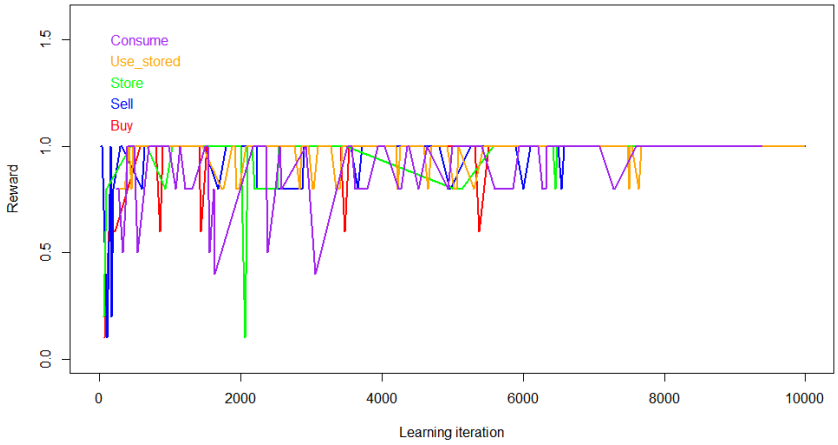

Fig. 2. Simple reinforcement learning algorithm

From Fig. 2 it is visible that in the first iterations, actions appear to be chosen randomly, however as the iterations increase, the algorithm manages to learn when the action should be used and in that way the algorithm converges around iteration 8000, and all actions start being used ideally.

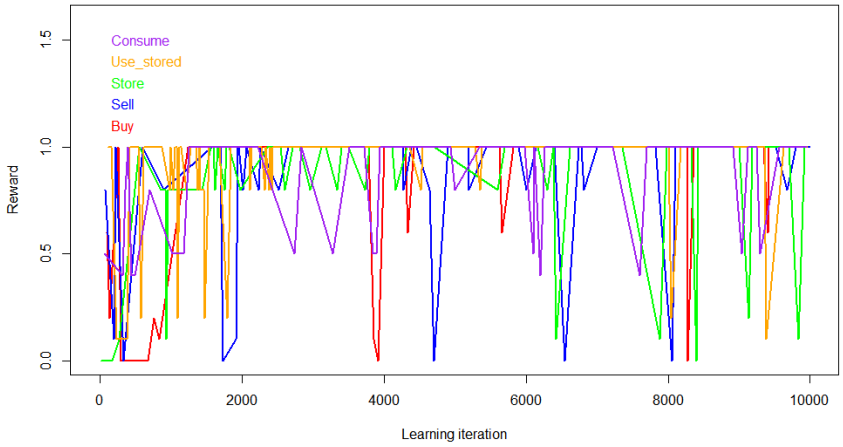

Fig. 3. EXP3

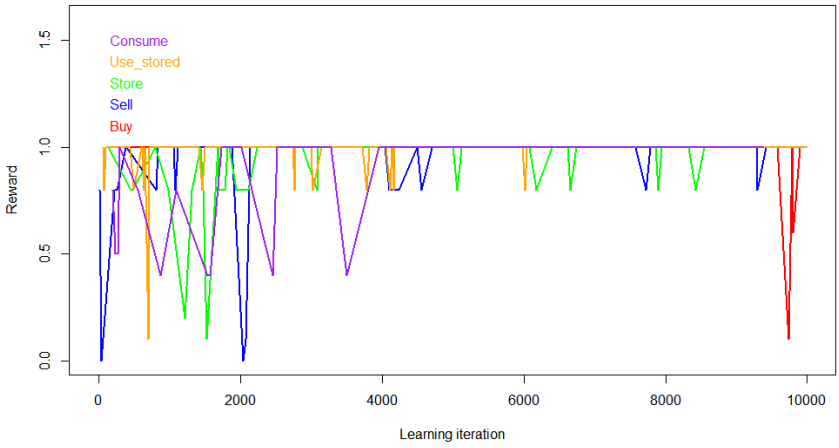

Fig. 4 - UCB1

EXP3 shows a different trend of results. It appears that algorithm is converging to the 1.0 reward value, which would indicate that actions are often being used ideally, however there also low points throughout the entire graph, which indicates that the algorithm is constantly exploring and trying to adapt to possible changes to the environment. On the other hand, UCB1 shows a much more stable behaviour. Most of the time actions are chosen ideally (exploitation), with a convergence to the reward value = 1 around iteration 5000. However, the algorithm still continues to explore other actions in a very small frequency, trying to adapt to possible changes.

Table 3 shows the ideal action frequency, which should be achieved if the algorithms would only exploit the best actions and not explore possible alternative actions. Table 4, Table 5 and Table 6 show the frequency of application of each action in each state, for all 3 algorithms. The highest frequency actions in each state are highlighted in green.

Table 3. Ideal action frequency

	Buy	**Consume**	**Sell**	**Store**	**Use_stored**
GetC	0	1	0	0	0
GgtC_PH	0	0	1	0	0
GgtC_PL	0	0	0	1	0
GltC_PH	0	0	0	0	1
GltC_PL	1	0	0	0	0

Table 4. Simple reinforcement learning algorithm action frequency

	Buy	**Consume**	**Sell**	**Store**	**Use_stored**
GetC	0.009	0.971	0.008	0.006	0.006
GgtC_PH	0.003	0.011	0.852	0.130	0.004
GgtC_PL	0.006	0.013	0.176	0.796	0.009
GltC_PH	0.043	0.119	0.004	0.002	0.831
GltC_PL	0.839	0.017	0.004	0.004	0.136

Table 5. EXP3 action frequency

	Buy	**Consume**	**Sell**	**Store**	**Use_stored**
GetC	0.054	0.833	0.029	0.042	0.042
GgtC_PH	0.026	0.042	0.762	0.138	0.033
GgtC_PL	0.040	0.036	0.159	0.725	0.040
GltC_PH	0.028	0.037	0.017	0.023	0.895
GltC_PL	0.796	0.040	0.027	0.029	0.109

Table 6. UCB1 action frequency

	Buy	**Consume**	**Sell**	**Store**	**Use_stored**
GetC	0.022	0.914	0.022	0.022	0.022
GgtC_PH	0.007	0.017	0.878	0.090	0.008
GgtC_PL	0.011	0.021	0.109	0.845	0.013
GltC_PH	0.022	0.065	0.005	0.005	0.903
GltC_PL	0.896	0.017	0.005	0.006	0.075

By comparing the three algorithms' action selection frequency in each state, it can be seen that the proposed UCB1 approach is able to achieve the best results, with the highest frequency of choice of the best action in 4 of the 5 considered states. Only for the state when the generation is equal to the consumption is the simple reinforcement learning algorithm able to reach a higher frequency for the Consume action, which means to simply consume the generated energy. The EXP3 algorithm reaches a low quality of results, giving a high priority to the exploration of alternative actions, and neglecting the exploitation of the best actions. The fact that the EXP3 algorithm does not consider the probability of transition between states, rather using a probability distribution for each state independently from the possible transitions, makes this algorithm disregard important information about the problem, which may be one of the main causes for its lack of success in this problem. The UCB1 approach, on the other hand is able to learn that the best action is to buy when the price is low and the generation is lower than the consumption; to sell when there is more generation than consumption and the price is high, but to store instead when the price is low; and to use the stored energy when price is high and the generation is not enough to meet the consumption.

4 Conclusion

Energy management in buildings is a central priority worldwide, due to the incentives to the increase of energy efficiency and renewable energy sources penetration. The uncertainty associated to the different resources makes this a hard problem to solve while considering its dynamic characteristics, and constantly changing nature.

This paper addresses this problem by proposing a solution modelled as a Markov decision process. A reinforcement learning model is applied for the intelligent energy management in buildings. A UCB1 based algorithm is applied, and the results are compared to those of an EXP3 approach and a simple reinforcement learning algorithm.

The results from the presented case study show that the proposed UCB1 approach is able to achieve a higher quality of results, by reaching a higher rate of successful actions identification, when compared to the other considered reference approaches.

References

1. Borlase, S.: Smart Grids: Infrastructure, Technology, and Solutions. (2012).
2. Rosen, C., Madlener, R.: Regulatory Options for Local Reserve Energy Markets: Implications for Prosumers, Utilities, and other Stakeholders. Energy J. 37, 39–50 (2016).
3. Fernandes, F., Morais, H., Vale, Z., Ramos, C.: Dynamic load management in a smart home to participate in demand response events. Energy Build. 82, 592–606 (2014).
4. Acone, M., Romano, R., Piccolo, A., Siano, P., Loia, F., Ippolito, M.G., Zizzo, G.: Designing an Energy Management System for smart houses. In: 2015 IEEE 15th International Conference on Environment and Electrical Engineering (EEEIC). pp. 1677–1682 (2015).
5. Zhou, B., Li, W., Chan, K.W., Cao, Y., Kuang, Y., Liu, X., Wang, X.: Smart home energy management systems: Concept, configurations, and scheduling strategies. Renew. Sustain. Energy Rev. 61, 30–40 (2016).
6. Afrakhte, H., Bayat, P., Bayat, P.: Energy management system for smart house with multi-sources using PI-CA controller. In: 2016 Iranian Conference on Renewable Energy Distributed Generation (ICREDG). pp. 24–31 (2016).
7. Pinto, T., Vale, Z., Sousa, T.M., Praça, I., Santos, G., Morais, H.: Adaptive learning in agents behaviour: A framework for electricity markets simulation. Integr. Comput. Aided. Eng. 21, 399–415 (2014).
8. Li, D., Jayaweera, S.K.: Reinforcement learning aided smart-home decision-making in an interactive smart grid. In: 2014 IEEE Green Energy and Systems Conference (IGESC). pp. 1–6 (2014).
9. Burtini, G., Loeppky, J., Lawrence, R.: A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit. arXiv1510.00757 [cs, stat]. (2015).
10. Bouneffouf, D., Féraud, R.: Multi-armed bandit problem with known trend. Neurocomputing. 205, 16–21 (2016).
11. Shimkin, N.: Reinforcement Learning – Basic Algorithms, (2011).
12. Xu, X., Zuo, L., Huang, Z.: Reinforcement learning algorithms with function approximation: Recent advances and applications. Inf. Sci. (Ny). 261, 1–31 (2014).
13. Faia, R., Pinto, T., Abrishambaf, O., Fernandes, F., Vale, Z., Corchado, J.M.: Case based reasoning with expert system and swarm intelligence to determine energy reduction in buildings energy management. Energy Build. 155, 269–281 (2017).