2021 IEEE PES/IAS PowerAfrica

Design and Testing of a Demand Response Q-Learning Algorithm for a Smart Home Energy Management System

Walter Angano
Department of Mechanical and Manufacturing Engineering
University of Nairobi
Nairobi, Kenya
walter.angano@students.uonbi.ac.ke

Peter Musau
Department of Electrical and Information Engineering
University of Nairobi
Nairobi, Kenya
pemosmusa@uonbi.ac.ke

Cyrus Wabuge Wekesa
School of Engineering
University of Eldoret
Eldoret, Kenya
cwekesa@uoeld.ac.ke

***Abstract*— Growth in energy demand stimulates a need to meet this demand which is achieved either through wired solutions like investment in new or expansion of existing generation, transmission and distribution systems or non-wired solutions like Demand Response (DR). This paper proposes a Q-learning algorithm, an off-policy Reinforcement Learning technique, to implement DR in a residential energy system adopting a static Time of Use (ToU) tariff structure, reduce its learning speed by introducing a knowledge base that updates fuzzy logic rules based on consumer satisfaction feedback and minimize dissatisfaction error. Testing was done in a physical system by deploying the algorithm in Matlab and through serial communication interfacing the physical environment with the Arduino Uno. Load curve generated from appliances and ToU data was used to test the algorithm. The designed algorithm minimized electricity cost by 11 % and improved the learning speed of its agent within 500 episodes.**

Keywords— Demand Response, Q-Learning, Reinforcement Learning, Smart Home Energy Management System, Time of Use

I. INTRODUCTION

Electrical energy has the advantage of versatility (can be put to multiple uses), cleanliness and can be transported at the speed of light. However, one major problem this form of energy faces is the expense of providing grid-scale storage. For this reason, the energy generated must simultaneously be consumed. That is, energy generation must balance energy demand plus energy losses at all times, a necessity that also facilitates support for system integrity (constancy of system frequency). The Kenya Least Cost Power Development Plan (LCPDP) report findings forecast an excess generation compared to demand in the coming years [1] and the consequence is an increase in electricity prices to meet costs due to excess generation.

Demand Side Management (DSM) has been demonstrated as an effective tool for promoting energy efficiency and balance between energy generation and demand. DSM as an overarching topic encourages energy consumers and utilities to be energy efficient. The elements of DSM include Load Management and Demand Response (DR). As one of the vehicles of DSM, DR refers to short-term responses to electricity market prices on the demand side/ by consumers [2]. DR programs are developed to encourage short-term load reductions by consumers when the energy pricing is high particularly during peak hours. DR programs are categorized into price and incentive-based [2]. Examples of DR price-based programs include the Static Time of Use (ToU) rates, Critical peak pricing (CPP), and Real-Time Pricing (RTP). Research on DR algorithms has evolved with Q-Learning agent-based algorithm being the predominant method.

This paper proposes an approach objectively to decrease the learning speed of a Q-learning agent and integrating consumer feedback on optimal policy by an agent subject to a static ToU. The rest of the paper is organized as follows; Background and Related Work, Methodology, Results and Discussion, Conclusion, Acknowledgement and References.

II. BACKGROUND AND RELATED WORK

A review of DR algorithms and modeling techniques by [3] illustrates Reinforcement Learning (RL) as a predominantly applied method in DR applications when problems are formulated as a Markov Decision Process (MDP). RL algorithm is considered more suitable in real-world applications, particularly DR. One of the RL algorithms widely used in DR is Q-learning which is agent-environment-based and seeks to establish an optimal policy from a set of actions. The authors concluded that most reinforcement learning algorithms have been performed in a simulation environment which has limited the implementation of such algorithms in residential and commercial buildings. Testing of algorithms in physical systems is a potential research path to measure the capability, flexibility and reliability of control by reinforcement learning agents. Limited publications considered human feedback through estimation of dissatisfaction function. Some algorithms are characterized with a curse of dimensionality problem particularly for large state-action where the speed of convergence is significantly reduced and subsequently learning speed by RL agent.

Other approaches explored include intelligent residential consumer systems that trade with an Energy Storage System (ESS) while non-intelligent consumers are given the option of purchasing energy from the ESS pool [4]. Intelligent residential systems have a smart agent that manages the ESS based on the pool price and neighborhood energy demand. The authors preferred a fuzzy inference system for the battery and price by setting a fuzzy logic where values of input vector through fuzzy rules are translated into corresponding output vector. The fuzzy rules represent the infinite states of energy price and the State of Charge as finite states.

2021 IEEE PES/IAS PowerAfrica | 978-1-6654-0311-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/POWERAFRICA52236.2021.9543244

978-1-6654-0311-5/21/$31.00 ©2021 IEEE

The authors [5] proposed a demand response scheme using RL with a single agent and integrates fuzzy reasoning to approximate values for reward functions. Human preference is considered in the control feedback as a state at each time step. Q-learning (an off-policy RL technique) was considered in selecting an optimal decision. The MDP constituted state-space with all the possible states in terms of power demand and electricity price signals. The reward function was implemented using fuzzy logic which approximates the numerical reward for a certain action and state. The actions with the highest reward values are considered optimal and corresponding actions implemented.

Multi-agent approaches included the design of a multi-agent RL intending to achieve an efficient home-based DR by modeling a one-hour ahead scheduling of smart appliances for a home energy management system with PV generation [6]. The proposed RL approach consists of two parts. The first part is a training of the Extreme Learning Machine (ELM) algorithm which is based on the feed-forward Neural Network. The ELM, using previous 24-hr data, predicts the 24-hr future trend on electricity prices and solar PV generation output. The predicted data is input to the second part which is a Q-learning algorithm designed to make hour-ahead decisions on energy consumption based on optimal policy. The optimal Q value is obtained using the Bellman equation. RL solution can be summarized to entail three algorithms, first algorithm the main function that initializes the parameters of the Q learning. The second algorithm is a feedforward NN with 24-hr data on electricity prices and solar generation as its input. The output is the predicted information on electricity price and solar generation for the next hour. The third algorithm is the Q-learning algorithm that makes scheduling decisions based on optimal policy.

Real-time DR was conducted to minimize the cost of electricity and maximize user comfort [7]. The authors presented an optimal scheduling strategy of appliances based on deep reinforcement learning (DRL) considering both discrete and continuous policies. An approximate policy was design based on the neural network (NN) to learn the optimal scheduling strategy from high-dimensional data of real-time pricing, states of an appliance, and outdoor temperature. The NN is trained using a policy search algorithm. The MDP structure consists of states as real-time electricity prices, outdoor temperature, and state of all appliances. Actions include binary control action variables/ discrete (deferrable appliances), continuous control variables (regulated appliances). Reward function modeled on three aspects: thermal comfort index, electricity cost, and consumer range anxiety. In solving the MDP, a neural network-based stochastic policy is adopted to determine the optimal policy. Bernoulli distribution and Gaussian distribution functions are used to estimate the approximate policy when the action is discrete and continuous, respectively. NN policy network determined the parameters for the distribution functions by learning them. The architecture of the NN takes in the input parameters (past electricity prices, outdoor temperatures, and states of all the appliances) and outputs the discrete and continuous actions by Bernoulli and Gaussian distribution functions respectively.

Most RL algorithms have been tested in simulation environments with limited testing in physical systems while others presented approaches that are considered complex for a simple residential system. In the context of integrating human feedback, the simulation environment limits actual feedback which is essential in understanding the performance of the agent. The curse of dimensionality has been addressed but learning speeds can still be significantly improved. Besides, the agent's action selection preference requires both exploitation and exploration of the environment. Multi-agent systems involved assigning an agent to each appliance which seems a complex system for small residential systems.

III. METHODOLOGY

A. Markov Decision Process (MDP) Model

1) Environment

The environment consists of non-schedulable appliances (mandatory) and schedulable (interruptible and non-interruptible) as the primary participants in DR. Schedulable appliances provide the leverage to deploy load management strategies per the ToU and realize energy savings. Load classification and load control level emanates from arranging load demand for the appliances according to consumer preference and priority and computing their cumulative load demand, respectively.

The total demand, P_t from all the appliances at any given time is given by equation

$$P_t = P_t^{NS} + P_t^{NI} + P_t^{IN1} + P_t^{IN2} \quad (1)$$

Load control levels (LCL) are defined cumulatively by adopting load demand for each appliance category.

$$LCL_1 = P^{NS} \quad (2)$$

$$LCL_2 = P^{NS} + P^{NI} \quad (3)$$

$$LCL_3 = P^{NS} + P^{NI} + P^{IN1} \quad (4)$$

$$LCL_4 = P^{NS} + P^{NI} + P^{IN1} + P^{IN2} \quad (5)$$

Where P^{NS}, P^{NI}, P^{IN1} and P^{IN2} is the total load demand for non-schedulable, non-interruptible, priority one and two interruptible appliances, respectively.

2) Agent

A single agent is designed and trained using data from a residential consumer and learns the environment for optimal policy output.

3) State Space

The set of state-space consists of the LCL and electricity static ToU.

4) Action Space

The action space consists of a set of load management strategies (load clipping, valley filling and load shifting) and status quo (no action). The balance in the action space is given in Fig 1. Load shifting and clipping actions are compensated by valley-filling.

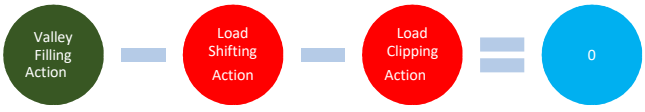

Fig. 1. Action space balance

The action space per the LCL and electricity price grid is assigned a weight which is essential when integrating feedback from consumers.

B. Reward Function

According to [8], the three main types of fuzzy logic systems commonly used include fuzzifier and defuzzifier, pure and Takagi-Segeno-Kang (TSK) fuzzy systems. A fuzzy system with fuzzifier and defuzzifier is commonly used as it eliminates problems associated with pure and TSK fuzzy systems. A Fuzzy logic system constitutes a crisp input (the LCL and static ToU) and crisp output is the numerical reward approximated by the system's fuzzy inference engine.

C. Fuzzy Rule Base and Fuzzy Inference Engine

A fuzzy rule base as the heart of the fuzzy system constitutes a set of IF-THEN rules. From Fig. 2, eight rules in a canonical form are defined in Table I. The fuzzy set **A** includes the LCL in a universe of discourse V equivalent to Load control 4 (LC4) and the status of electricity prices (whether Low or High) in a universe of discourse derived from historical tariff data.

TABLE I. THE CANONICAL FORM OF THE RULE BASE

Rule 1	If (LD is LCL1) and (ET is LP) then (SQ is NR)(LS is NR)(LC is NR)(VF is HR) (1)
Rule 2	If (LD is LCL1) and (ET is HP) then (SQ is HR)(LS is NR)(LC is NR)(VF is NR) (1)
Rule 3	If (LD is LCL2) and (ET is LP) then (SQ is LR)(LS is LR)(LC is NR)(VF is HR) (1)
Rule 4	If (LD is LCL2) and (ET is HP) then (SQ is HR)(LS is LR)(LC is NR)(VF is NR) (1)
Rule 5	If (LD is LCL3) and (ET is LP) then (SQ is HR)(LS is R)(LC is NR)(VF is NR) (1)
Rule 6	If (LD is LCL3) and (ET is HP) then (SQ is NR)(LS is HR)(LC is NR)(VF is NR) (1)
Rule 7	If (LD is LCL4) and (ET is LP) then (SQ is NR)(LS is NR)(LC is HR)(VF is NR) (1)
Rule 8	If (LD is LCL4) and (ET is HP) then (SQ is NR)(LS is NR)(LC is HR)(VF is NR) (1)

Fuzzy set B constitutes action space linguistic form Highly Recommended (HR), Recommended (R), Least Recommended (LR) and Not Recommended (NR).

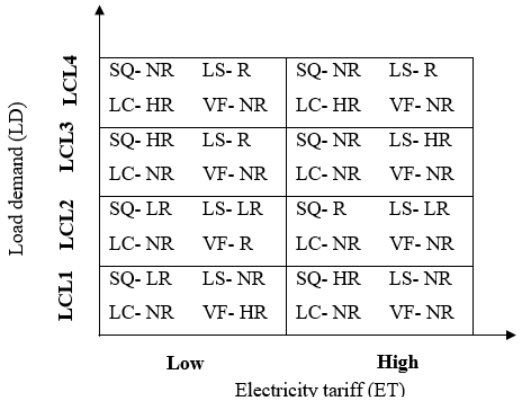

Fig. 2. Load control and electricity prices grid

Mamdani inference method, a type of composition-based inference, is adopted based on intuitive appeal. Mamdani combination is defined as a single fuzzy relation Q_M,

$$Q_M = \bigcup_{k=1}^{Y} Ru^k \tag{6}$$

A minimum inference engine is adopted in this research defined as,

$$\mu_{B^k}(y) = \underset{k=1}{\overset{Y}{max}} \left[\sup_{x \in U} min \begin{pmatrix} \mu_{A^k}(x), \mu_{A_1{}^k}(x_1), \dots \\ , \mu_{A_n{}^k}(x_n), \mu_{B^K}(y) \end{pmatrix} \right] \tag{7}$$

Triangular fuzzifier maps a real value $\boldsymbol{x^{max}} \in \boldsymbol{U}$ to a fuzzy set $A^k \, in \, U$ characterized by a triangular membership function,

$$\mu_A(x) = \begin{cases} \left(1 - \frac{|x_1 - x_{max}|}{b_1}\right) \dots \left(1 - \frac{|x_n - x_{max}|}{b_n}\right) \\ \quad if \ |x_i - x_{i,max}| \le b_i, i = 1,2,\dots,n \\ 0 \qquad\qquad\qquad if \ otherwise \end{cases} \tag{8}$$

This paper adopts the center of gravity (CoG) defuzzifier. The CoG defuzzifier specifies y^* as the area center covered a membership of B^k as

$$y^* = \frac{\int_V y(\mu_{B^k})(y)dy}{\int_V (\mu_{B^k})(y)dy} \tag{9}$$

The defuzzifier in this case outputs the approximate reward based on the crisp input (LCL and static ToU).

D. Introduction to Q-Learning Algorithm

Q-learning algorithm is a temporal difference learning algorithm and an off-policy reinforcement learning that aims to learn optimal policy and approximates the current optimal action-value q_* using the Bellman equation,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{10}$$

Q-learning algorithm computes the value of taking an action a in state s and determines the optimal policy, $q_*(s, a)$ from a set of actions $a \in A(s)$ for that particular state. The parameters α and γ represent the learning rate of the algorithm and the discount factor [9], [10].

E. Exploration and Exploitation

This paper adopts the $\varepsilon - greedy$ exploration technique which ensures actions are selected randomly (exploration) and greedily (exploitation) with a probability ε and $1 - \varepsilon$, respectively.

$$a_t = \begin{cases} random \ action & with \ probability \ \varepsilon \\ max Q_t(a) & with \ probability \ 1 - \varepsilon \end{cases} \tag{11}$$

F. Returns and Episodes

The primary goal of an agent is to maximize cumulative rewards in a particular time slot. Denote sequence of rewards as $R_{t+1}, R_{t+2}, R_{t+3}, \dots$ so that the expected return is maximized. The maximized return is considered a function of the sum of all rewards.

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T \tag{12}$$

Where T is the final time step or episode.

G. Integration of Consumer Feedback

Binary action vectors by the algorithm and consumer at time t are represented as $W_{Alg,t}$ and $W_{Cons,t}$, respectively. Then the magnitude of the vector can be used to determine consumer dissatisfaction. Consider the length of the action vector difference,

$$\Delta W = \| W_{Alg,t} - W_{Cons,t} \| \quad (13)$$

When, $\Delta W = 0$, the consumer is satisfied with the algorithm's optimal policy. However, the consumer shows dissatisfaction when $\Delta W > 0$. Consumer dissatisfaction with the algorithm's decision is handled by updating the fuzzy rules. However, the reward difference between consumer dissatisfaction and algorithm is minimized. The weighting method is used to assign the weights of the linguistic action space in Fuzzy set B as Highly Recommended (HR) – 0.4, Recommended (R) – 0.3, Least Recommended (LR)-0.2 and Not Recommended (NR) - 0.1. At $time\ t$, the algorithm's action vector and corresponding index are given as:

$$[\beta_{alg,t}, \lambda_{alg,t}] = max(W_{Alg,t}) \quad (14)$$

Consumer feedback is represented as

$$[\beta_{cons,t}, \lambda_{cons,t}] = max(W_{cons,t}) \quad (15)$$

The difference in reward needs to be greater than zero to guarantee an update to the rules. When $R_t(\lambda_{cons,t}) - R_t(\lambda_{alg,t}) > 0$, then the rules are updated depending on the weightage. The load demand and electricity price grid G is assigned the maximum of weighted Fuzzy set B.

$$G\left((ET_{n=1\ to\ 2}, LD_{m=1,\dots,4}), \lambda_{alg,t}\right) = G\left((ET_{n=1\ to\ 2}, LD_{m=1,\dots,4}), \lambda_{cons,t}\right) \quad (16)$$

$$G\left((ET_{n=1\ to\ 2}, LD_{m=1,\dots,4}), \lambda_{cons,t}\right) = \max(Fuzzy\ B) \quad (17)$$

An example of the fuzzy rule update is illustrated in Fig. 3.

SQ- 0.1	LS- 0.3	SQ- 0.1	LS- 0.3
LC- 0.4	VF- 0.1	LC- 0.4	VF- 0.1
SQ- 0.4	LS- 0.3	SQ- 0.1	LS- 0.4
LC- 0.1	VF- 0.1	LC- 0.1	VF- 0.1
SQ- 0.2	LS- 0.2	SQ- 0.3	LS- 0.2
LC- 0.1	VF- 0.3	LC- 0.1	VF- 0.1
SQ- 0.2	LS- 0.1	SQ- 0.4	LS- 0.1
LC- 0.1	VF- 0.4	LC- 0.1	VF- 0.1

→

SQ- 0.1	LS- 0.3	SQ- 0.1	LS- 0.3
LC- 0.4	VF- 0.1	LC- 0.4	VF- 0.1
SQ- **0.3**	LS- **0.1**	SQ- 0.1	LS- 0.4
LC- 0.1	VF- **0.4**	LC- 0.1	VF- 0.1
SQ- 0.2	LS- 0.2	SQ- 0.3	LS- 0.2
LC- 0.1	VF- 0.3	LC- 0.1	VF- 0.1
SQ- 0.2	LS- 0.1	SQ- 0.4	LS- 0.1
LC- 0.1	VF- 0.4	LC- 0.1	VF- 0.1

Fig. 3. Fuzzy rule update using the knowledge base

H. Time of Use Tariff Structure

The tariff structure is given in Fig.4. Historical tariff data for residential consumers in Kenya are distributed around the mean which is also the shoulder or mid-peak. ToU plans from Ireland Italy, Australia, Canada and Sri Lanka [12] are used in benchmarking.

I. Testing Setup

The testing set-up in Fig. 5 is implemented using the Arduino Uno kit. An Arduino program is preloaded in the microprocessor. This program checks if the serial port has changed for processing. Matlab program which is the agent communicates with the preloaded program through serial communication.

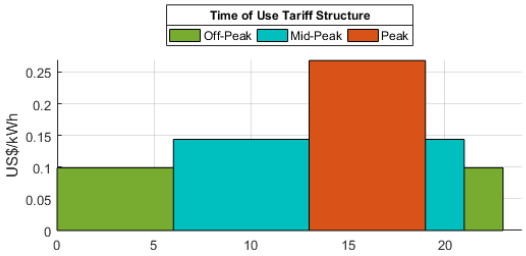

Fig. 4. Static Time of Use tariff plan

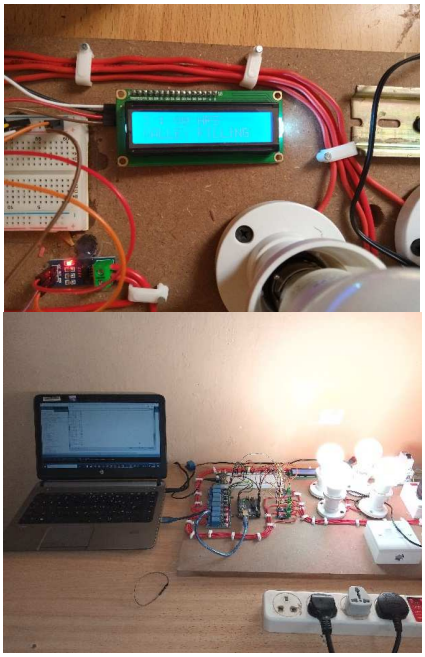

Fig. 5. Testing set-up for the algorithm

IV. Results and Discussion

Fig. 6 expresses the learning curve as the graph of mean cumulative rewards as a consequence of the agent's optimal policy selection against the number of episodes or training time taken. It was observed that the agent converged after 500 episodes.

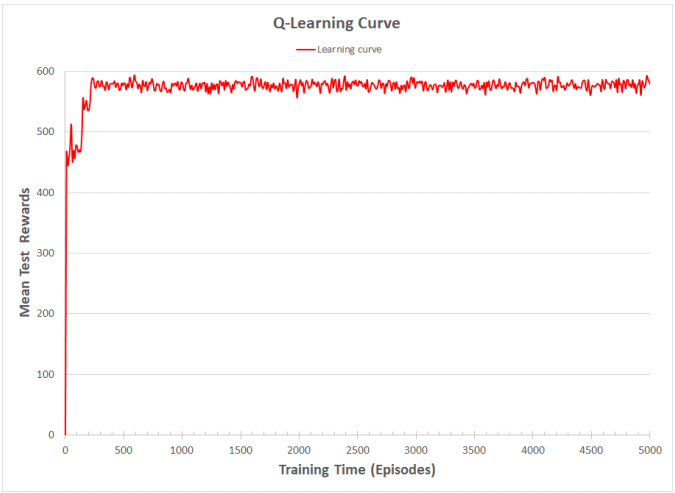

Fig. 6. The training curve for the algorithm

The overall effect of the agent's action is evaluated through an overlay of the recommended actions on the initial load curve. The applicable load management strategies include load shifting, valley-filling and status quo. Appliances in load category 3 were shifted during the peak times when the tariff is high. This happens between 07:00-16:00 and 18:00-21:00 HRS which are the mid-peak and peak hours

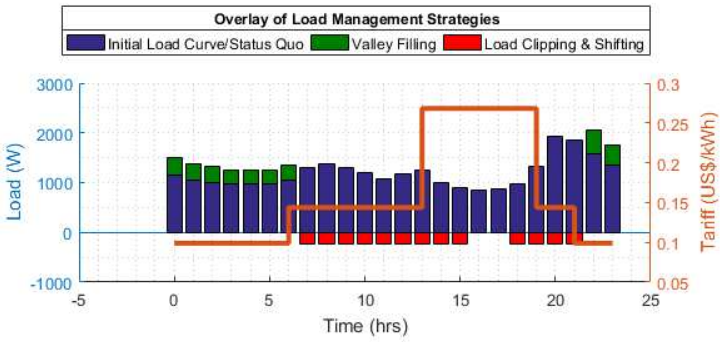

Fig. 7. Overlay of Load Management Strategies

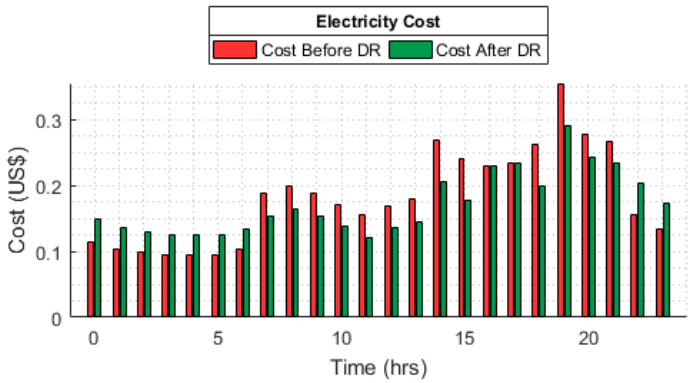

Fig. 8. Comparison of Energy Cost before and after demand response

For the status quo, the net cost is zero since no action is taken. Fig 7 shows that during valley filling, the demand response effect is an extra electricity cost on the consumer since appliances are added hence an increase in total load. The costs and savings as a result of the corresponding applicable load management strategies only at the region or time when they are applied as shown in Fig. 8. The net energy savings realized is 11 percent as in Fig. 9.

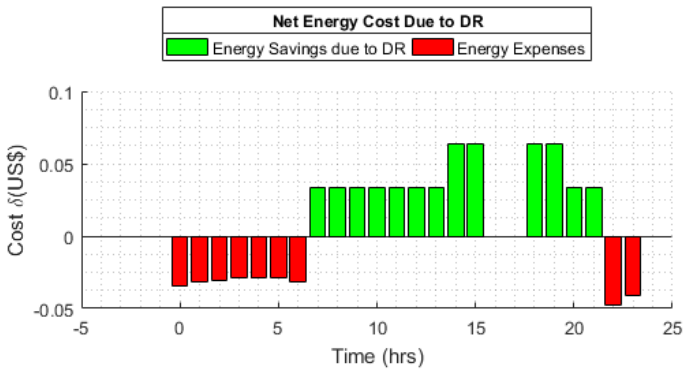

Fig. 9. Comparison of Energy Cost before and after demand response

V. Conclusion

Q-learning is a predominant tool of reinforcement learning that researchers have found resourceful when establishing optimal policy from a set of actions. This paper proposed improving some of the gaps by establishing a state-space action consisting of consumer-tailored load categories by grouping appliances according to their priority and usage frequency. A knowledge improvement base was developed to update the fuzzy rules and ensure the algorithm minimizes consumer dissatisfaction. This approach resulted in the agent's improved learning speed with convergence in 500 episodes and cost savings of 11 percent. A testing system was assembled and interfaced with a graphical user interface designer using app designer in Matlab. Through serial communication, the Arduino microprocessor received command signals from Matlab and either activated or deactivated a relay to turn the loads on or off.

Future research work will focus on developing consumer dissatisfaction models using Artificial Neural Network (ANN) and cloud-based technologies such as Microsoft Azure and integrating the models as a crisp input in fuzzy systems.

Acknowledgment

I acknowledge Virunga Power's support and encouragement during this research.

References

[1] Energy and Petroleum Regulatory Authority, "Updated Least Cost Development Plan Study Period 2020-2040," EPRA, Nairobi, 2020.

[2] United Nations Industrial Development Organization, "Sustainable Energy Regulation and Policy-making Training Manual - Demand Side Management," United Nations Industrial Development Organization, 2009.

[3] I. Hussain, S. Mohsin, A. Basit, Q. U. Khan ZA and N. Javaid, "A review on demand response: pricing, optimization, and appliance scheduling," *Procedia Computer Science,* vol. 52, p. 843–50, 2015.

[4] J. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied Energy,* vol. 235, pp. 1072-89, 2019.

[5] S. Zhou, Z. Hu, W. Gu, M. Jiang and X. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE Journal of Power and Energy Systems,* vol. 5, pp. 1-10, 2019.

[6] F. Alfaverh, M. Denaï and Y. Sun, "Demand Response Strategy Based on Reinforcement Learning and Fuzzy Reasoning for Home Energy Management," *IEEE Access,* vol. 8, pp. 39310-39321, 2020.

[7] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai and S. C. Lai, "IEEE Transactions on Smart Grid," *A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management,* vol. 11, pp. 3201-3211, 2020.

[8] H. Li, Z. Wan and H. He, "Real-Time Residential Demand Response," *IEEE Transactions on Smart Grid,* vol. 11, pp. 4144-4154, 2020.

[9] L. X. Wang, A Course in Fuzzy Systems and Control, Prentice Hall PTR, 1997.

[10] R. S. Sutton and A. G. Barto, "Reinforcement Learning," in *Finite Markov Decision Processes*, Cambridge, The MIT Press, 2018, pp. 47-71.

[11] B. Jang, M. Kim and G. Harerimana, "Q-Learning Algorithms: A Comprehensive Classification and Applications," *IEEE Access,* vol. 7, pp. 133653-133667, 2019.

[12] Electricity Supply Board, "ESB Customer Supply Proposal for Smart Metering Tariff," ESB, South City, Ireland, 2010.