Q-Learning Algorithm and CMAC Approximation Based Robust Optimal Control for Renewable Energy Management Systems ⁑

Vy Huynh Tuyet, Luy Nguyen Tan

Faculty of Electronic Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
(e-mail: huynhtuyetvy@iuh.edu.vn, nguyentanluy@iuh.edu.vn)

Abstract: This paper investigates a robust optimal control algorithm for a renewable energy management system. The algorithm is obtained by developing a novel method based on the zero-sum games (ZSG) theory for H_∞ control and the well-known Q-learning algorithm. Firstly, the H_∞ performance index function is formed via real-time parameters including electricity price, load demand, solar energy, and battery lifetime. Secondly, a self-learning and control algorithm is established, and the value function solution is approximated by the cerebellar model articulation controller (CMAC). Finally, the algorithm guarantees that the disturbance compensation policy, optimal value function, and the optimal control strategy converge to the near-optimal values. Comparison with other methods in a numerical experiment using practically measured data is implemented to evaluate the effectiveness of the designed algorithm.

Keywords: Renewable energy, storage system, robust optimal control, Q-learning, CMAC, neural network.

1. INTRODUCTION

The human brain (Frackowiak & SJ, 2004) is able to deal with a huge set of problems with known patterns in a given context through the process of perception, memory, learning, and behavior generation. Artificial neural networks are rapidly being used to obtain a competitive advantage through data and automation in production, based on what is achievable with the human brain (Bannat, et al., 2010; Carvalho, et al., 2020; Dumitrache, et al., 2019). In recent times, the design controller for renewable energy storage systems has drawn the attention of the control community. One of the challenges of the problem that needs to be solved is that the feedback state from renewable energy is intermittent, random, and unpredictable. The controller does not merely charge/discharge to add more power to the grid system, increasing economic efficiency in the conventional sense, but also ensures the optimal economy over time, extends the maximum battery's lifetime, and avoids overcharging and over-discharging (Guerrero, et al., 2013; Angelis, et al., 2013; Lai & McCulloch, 2017; Liu, et al., 2016). In (Appen, et al., 2014; Shafiee, et al., 2014), an intelligent control scheme is designed, which allows the main components in the grid system to connect, and all operations of the system are fully coordinated with the goal of utilizing optimally the available energy, including renewable energy. In particular, in any smart grid, the energy storage system is an indispensable component. It is used to charge energy taken from the grid and discharge energy to the load, with the aim of reducing the electricity cost for the grid power supplier.

Adaptive dynamic programming (ADP) theory has been continuously developed and applied to the optimal control problem over the past several years (Lewis, et al., 2012; Yang, et al., 2015; Jiang & Jiang, 2014). ADP techniques showed the potential control capabilities and were quite widely used in many real applications, such as in communication (Liu, et al., 2018; Jiang, et al., 2019; Zhang & Xin, 2021), in medical diagnosis (Petousis, et al., 2019; Yu, et al., 2020; Mukherjee & Bohra, 2020), and particularly, in energy storage system (ESS) (Perera, et al., 2021; Ojand & Dagdougui, 2022). ESS has been considered as a solution to reduce power losses or as a backup, additional power supply in times of power shortage. ESS is often embedded in the distributed network, which is able to integrate renewable energy sources or not (Zhao & Ding, 2018; Duan, et al., 2019; Shuai, et al., 2019). In (Venayagamoorthy, et al., 2016; Lu & Wang, 2020; Berrueta, et al., 2020; Khan, et al., 2021), the battery structures in ESS were focused on being analyzed by the self-learning algorithms to improve the battery capacity and the management of various types of batteries' characteristics. In this paper, ESS is considered as a part of the electric grid in a household area.

The basic ADP algorithm (Werbos, 1977) is applied to control the energy management systems with optimal performances (Wei, et al., 2015; Wei, et al., 2017; Song, et al., 2014; Venayagamoorthy, et al., 2016). There exist several basic ADP algorithms, but the Q-learning algorithm is a typical ADP method having the capability of learning and controlling online without completely unknown system dynamics. In (Huang & Liu, 2011; Si & Wang, 2001), the basic Q-learning algorithm is combined with a neural network (NN) to design an optimal controller for a renewable energy management system connecting to the grid power, the storage system, and the load. In (M. Boaro et al., 2013), the Q-Learning algorithm can minimize the value function of a smart energy system, including wind and solar energy. To increase the training speed of NN, Q-learning, combined with the particle swarm optimization (PSO) technique, was proposed in (Fuselli, 2013). Wei et al. in (Q. Wei et al., 2015) introduced the Q-learning algorithm with two iterations to obtain optimal power control in a residential grid environment. However, the

⁑ This work was supported by the Industrial University of Ho Chi Minh City under Grant 22/HD-DHCN.

algorithm does not consider solar energy in the optimal control. In (Wei, et al., 2017), solar energy was considered sequentially, and the value iteration (VI) control method was designed for a battery in an energy storage system instead of the Q-learning algorithm. Although algorithms ensure that the value function converges to a near-optimum, two iterative loops are required, increasing the computing complexity.

There are various approaches to solving the optimization problem that can be applied to finding the optimal control policy in an energy management system (Zamfirache, et al., 2022). However, the Q-learning and VI algorithms listed above have not been considered as part of the energy management system with external disturbance yet. In practice, the power disturbance from renewable energy can cause the instability of controlled systems. Therefore, the robust control problem needs to be given attention. H_∞ techniques have played an important role in analyzing and designing robust control algorithms that can be applied to linear systems by solving the Riccati equation (Li, et al., 2014) or to nonlinear systems by solving the equation of Hamilton-Jacobi-Isaacs (HJI) (Van der Schaft, 1992). In (Rigatos, et al., 2017), H_∞ is used to compensate for the disturbances that are created due to the linearizing of the oxygenator's dynamic model and to improve robustness of the system against modelling uncertainty and external disturbances. Although modern control theory has been developed strongly to solve H_∞ solutions for nonlinear systems (Basar & Bernhard, 2008), finding a saddle point by analytical solution, which includes the optimal control policy and disturbance compensation policy, has proven to be impossible (Van der Schaft, 1992; Basar & Bernhard, 2008; Wu & Biao, 2012). The ADP method is a strong and effective technique to approximate the saddle point. In (Abu-Khalaf & Lewis, 2008; Abu-Khalaf, et al., 2007), based on the ZSG theory, the Nash equilibrium is approximated online by three NN structure; one for approximating the value function, and the other two for approximating the optimal control strategy and disturbance compensation policy, respectively.

In the Q-learning method, at every iteration step, $Q(x_k, u_k)$ is updated and memorized with all x_k and u_k (x_k: system state, u_k: control signal). Therefore, applying this method to control renewable energy management systems has some disadvantages as follows: 1) Can only be applied to a system with finite points in a state space and a finitely quantized control signal set. Because there are an infinite number of points that need to be discretized in a continuous system, the computational cost does not allow one to go through all of these points to explicitly update and store the evaluation function. 2) The cost of storage and computation grows exponentially $\Omega_x^{|U_x|}$ as the number of points and quantized signals in the state space and control space increases. This leads to a combinatorial explosion. The cost of computation and storage is reduced if the function values of the un-updated points in the state space can be interpolated from the function values of their updated neighbors. The approximator is one of the effective tools that can solve this problem (Xu, et al., 2014). Up to now, there have been many studies and applications of approximation applied to ADP, such as radial basis function (RBF) with versions of normalized RBF (NRBF), resource allocating RBF (RARBF), adaptive normalized RBF (ANRBF), multi-layer perceptron (MLP), and cerebellar model articulation controller (CMAC). In these approximation methods, CMAC has outstanding advantages in terms of computational efficiency and convergence speed (Tham, 1994).

In this paper, starting from the above analysis, we first design a Q-learning based robust optimal control (Q-ROC) algorithm for the full energy management system, including grid, renewable power, storage, and load. This is also the first time external disturbances are taken into consideration to resolve the robust optimal control problem for renewable energy storage systems. The following is a list of the paper's main contributions: 1) Unlike the methods in (Q. Wei et al., 2015; Wei, et al., 2017), Q-ROC can compensate for external disturbances optimally. 2) Instead of approximating $Q(x_k, u_k)$ by MLP as in (Q. Wei et al., 2015), Q-ROC uses the approximator CMAC to increase computational efficiency and speed up convergence. 3) The Q-ROC algorithm, which is executed on real-time data, including solar power, load, and price, is compared its effectiveness to that of the existing algorithms. The advantage of the proposed algorithm in the paper compared to those in our previous works (Luy, 2017; Luy, 2018) or the existing works (Wei, et al., 2017; Wei, et al., 2018) is that Q-Learning is a model-free control method versus the actor-critic algorithm. Therefore, identification of the structure and the parameters for the renewable energy management system is not needed. It is worth noting that the nonlinear model in the paper is just used in simulation; it is not used in control design.

The rest content is divided into following sub-parts. Section 2 formulates the problem, Section 3 designs Q-ROC algorithm, Section 4 gives a numerical experiment, and Section 5 draws a brief conclusion.

2. PROBLEM FORMULATION

This section describes the system model, assumptions, and control objective and establishes the performance index function with external disturbance.

2.1 Smart grid system

A smart grid system, which is described in Fig. 1, consists of grid power, solar energy, battery power, load and a controller.

In this system, the power flows between components are defined as follows: (i) Solar energy can be used to meet demand $T_{RL,k}$, while also charging the storage system $T_{RB,k}$. Solar energy, on the other hand, is free and is connected directly to load and storage system; (ii) Grid power is a one-way power flow that is used to meet load demand and charge storage system $T_{GB,k}$. Power flow from solar to grid or energy storage to grid is not allowed; (iii) Storage system can exist in three states: charged by solar energy or grid, discharged to meet load demand, and idle; (iv) Load can receive power from one or more resources simultaneously, including solar to load $T_{RL,k}$, grid to load $T_{GL,k}$, and storage system to load $T_{BL,k}$.

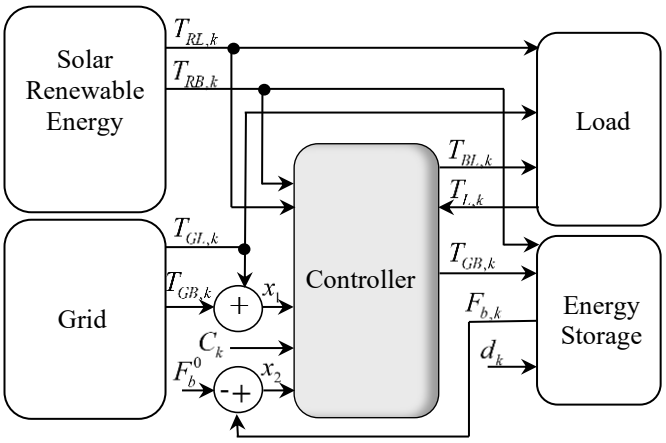

Fig. 1. Smart grid system.

The inputs of the controller include grid power, load demand, and storage power, which change continuously with solar energy and control law. Based on the parameters, including price C_k, load demand $T_{L,k}$, grid power $T_{G,k}$, solar power to storage system $T_{RB,k}$, storage power $F_{b,k}$, and storage demand F_b^0, performance index function is established and minimized to get control law. As a result, the control law charges/discharges the storage system to ensures the load balance. Note that due to the instability of load, disturbance from grid, and solar power, which is dependent on weather, season, and time a day, there always exists disturbance d_k affecting the system. $x_{1,k}$, $x_{2,k}$, which are variables of system state, are expressed in subsection 2.2, and equation (8)

Accordingly, the balance of the load and the grid power is described as:

$$T_{L,k} = T_{RL,k} + T_{GL,k} + T_{BL,k} \tag{1}$$

$$T_{G,k} = T_{GL,k} + T_{GB,k} \tag{2}$$

The balance of the solar energy is defined as:

$$T_{R,k} = T_{RL,k} + T_{RB,k} \tag{3}$$

In this system, the battery, which is utilized as a storage system, is designed not to charge or discharge simultaneously, and its battery model is expressed as (Huang & Liu, 2011):

$$\begin{aligned} F_{b,k+1} &= F_{b,k} - T_{BL,k}(0.898 - 0.173 T_{BL,k}/T_{rate}) \\ &+ (T_{BL,k} + T_{GB,k})0.898 - 0.173(T_{RB,k} + T_{GB,k}/T_{rate}) \end{aligned} \tag{4}$$

where the battery power $F_{b,k}$ has storage limit in range: $F_b^{\min} \le F_{b,k} \le F_b^{\max}$, $F_b^{\min}$ is the battery's minimum storage energy, $F_b^{\max}$ is the battery's maximum storage energy, $T_{rate} > 0$ is the battery's nominated power output. Note that self-discharge is prohibited. As a result, we define the battery discharge by $F_{b,k} > 0$, charge by $F_{b,k} < 0$, and idle by $F_{b,k} = 0$.

2.2 Nonlinear model and control objectives

Due to the costlessness of solar energy, it is prioritized to satisfy load demand first, and the rest is stored by charging the battery. Hence, the solar energy to load and to battery are described as follows:

$$T_{RL,k} = \begin{Bmatrix} T_{R,k}, T_{L,k} \ge T_{R,k} \\ T_{L,k}, T_{L,k} < T_{R,k} \end{Bmatrix} \tag{5}$$

$$T_{RB,k} = \begin{cases} 0, T_{L,k} \ge T_{R,k} \\ T_{R,k} - T_{L,k}, T_{L,k} < T_{R,k} \end{cases} \tag{6}$$

Accordingly, the load balance (1) is rewritten using (2) and (6) as:

$$P_{L,k} = T_{G,k} + (T_{BL,k} - T_{GB,k}) \tag{7}$$

Let system state vector be $x_k = \left[x_{1,k}, x_{2,k} \right]^T$, $x_{1,k} = T_{G,k}$, $x_{2,k} = F_{b,k} - F_b^0$. Let the control policy be $u_k = T_{RB,k} + T_{BL,k} - T_{GB,k}$. The equation of the nonlinear discrete dynamic with disturbance d_k, which is derived from the storage model and the equations (1)-(7), are expressed as:

$$\begin{aligned} x_{k+1} &= F(x_k, u_k, d_k, k) \\ &= \begin{pmatrix} P_{L,k} - u_k + d_k \\ x_{2,k} - (u_k - P_{R,k})\vartheta(u_k - P_{R,k}) \end{pmatrix} \end{aligned} \tag{8}$$

where

$$\vartheta(u_k - P_{R,k}) = 0.898 - 0.173\left| u_k - P_{R,k} \right| / T_{rate}$$

For convenience of analysis, assumptions are described as follows:

Assumption 1 (Wei, et al., 2017)*:* Given the period of λ = 24 hours and time step of 1 hour, the price, load energy, and renewable energy are discrete-time as:

$$C_k = C_{k+\lambda}; T_{L,k} = T_{L,k+\lambda}; T_{R,k} = T_{R,k+\lambda} \tag{9}$$

Assumption 2 (Wei, et al., 2017)*:* The power transfer from renewable solar or battery to grid is not taken into account, thus we specify $T_{G,t} > 0$.

Assumption 3: The battery is only in one of three states: charge, discharge, or idle. That is, it can't charge and discharge at the same time. Then, $T_{BL,k} > 0$ and $T_{GB,k} > 0$ imply $T_{GB,k} = 0$, and $T_{BL,k} = 0$, respectively.

Based on the H_∞ control theory and the required performance of renewable energy management systems in optimal control (Wei, et al., 2017), the cost function in the paper is presented as follows:

$$J(x_k)=\sum_{k=0}^{\infty}\gamma^k\Big(\alpha(C_kT_{GL,k})^2+\beta(F_{b,k}-F_b^0)^2 +\delta(T_{RB,k}+T_{GB,k}-T_{BL,k})^2-\mu\big(d_k\big)^2\Big) =\sum_{k=0}^{\infty}\gamma^k\Big(\alpha(C_kx_{1,k})^2+\beta x_{2,k}^2+\delta u_k^2-\mu\big(d_k\big)^2\Big) \tag{10}$$

where $0<\gamma\le 1$ is discount factor, $\alpha,\beta,\delta>0$ is performance index factors, d_k is disturbance compensation factor. The performance index function (10) describes the total cost that must be paid for the electricity supplier in the first term. The second term ensures that the battery's stored power is maintained not far away from the average level of the battery storage $F_b^0=(F_b^{\min}+F_b^{\max})/2$. The third term prohibits the battery from being fully charged or fully discharged. The second term and the third one are both designed to increase the battery's life. The last term relates to the disturbance compensation in H_∞ control problem with $\mu\ge\mu^*$. μ^* is defined as the minimum compensation factor so that the closed system remains stable (Van der Schaft, 1992).

Optimization objective: Under the influence of the unstable components, including solar energy, disturbance from grid, load, a robust optimal method is developed to minimize the cost function so that the economic efficiency is increased, the battery lifetime is extended, and the load balancing is maintained.

3. Q-ROC ALGORITHM

3.1 Q-learning in the energy management system

The cost function (10) is written as:

$$Q(x_0,u_0,w_0)=\sum_{k=0}^{\infty}\gamma^kS(x_k,u_k,w_k) \tag{11}$$

where $S(x_k,u_k,w_k)=x_k^TM_kx_k+\delta u_k^2-\mu d_k^2$, x_0 is the initial state, and $M_k=\begin{bmatrix}\alpha C_k^2, 0;0\ \beta\end{bmatrix}$.

According to Watkins (Watkins, 1989) and ZSG theory in H_∞ control (Basar & Bernhard, 2008; Wu & Biao, 2012; Wei, et al., 2018), the optimal value function is expressed by:

$$Q^*(x_0,u_0,w_0)=S(x_k,u_k,w_k)+\gamma\min_{u_{k+1}}\max_{w_{k+1}}Q^*(x_{k+1},u_{k+1},w_{k+1}) \tag{12}$$

The optimal cost function (10) fulfills the Bellman's principle as:

$$J^*(x_k)=\min_{u_k}\max_{w_k}\Big[S\left(x_k,u_k,w_k\right)+\gamma J^*(x_{k+1})\Big] \tag{13}$$

The disturbance compensation law w_k (first player) is utilized to maximize the cost function (13), meanwhile the control strategy u_k (second player) is utilized to minimize its value. If there is a saddle point in the control method, the control law and disturbance compensation law are defined as follows:

$$u_k^*=\arg\min_{U_k}Q^*(x_k,U_k,w_k) \tag{14}$$

$$w_k^*=\arg\max_{W_k}Q^*(x_k,u_k,W_k) \tag{15}$$

From (12) and (13), we get the optimal value function:

$$J^*(x_k)=Q^*(x_k,u_k^*,w_k^*) \tag{16}$$

However, the result of $Q^*(x_k,u_k,w_k)$ equation can't be obtained explicitly by the mathematical expressions. An iterative algorithm is developed to approximate this result. The value of Q-function is updated at the iterative step l by (Watkins, 1989):

$$Q^l(x_k,u_k,w_k)=Q^{l-1}(x_k,u_k,w_k)+\chi\Big(S(x_k,u_k,w_k)+\gamma\min_{U_k}\max_{W_k}Q^{l-1}(x_{k+1},U_k,W_k)-Q^{l-1}(x_k,u_k,w_k)\Big) \tag{17}$$

where $0<\chi<1$ is the updating speed. Accordingly, the control law and disturbance compensation law are at the iterative step l:

$$u_k=\arg\min_{U_k}Q^l(x_k,U_k,w_k) \tag{18}$$

$$w_k=\arg\max_{W_k}Q^l(x_k,u_k,W_k) \tag{19}$$

Equations (17) – (19) proceed iteratively until obtaining the convergence condition $\left\|Q^l(x_k,u_k,w_k)-Q^{l-1}(x_k,u_k,w_k)\right\|<\delta$, where δ is a small positive constant.

3.2 Q-ROC with CMAC approximator

From the equation (17), at the iterative step l, the storage and computation costs increase exponentially ($\Omega_{x_k}^{\|U_k\|\times\|W_k\|}, x_k\in\Omega_{x_k}$) between the number of explicit points in state space and the number of control signals and disturbance signals of each state. It leads to a combinatorial explosion. To overcome this disadvantage, function approximation is employed.

The Q-ROC learning structure is proposed in Fig.2. The CMAC approximator is to find $Q^l(x_k,u_k,w_k)$. The control strategy u_k, and the disturbance compensation policy w_k are approximated by two two-layer perceptron (2-LP) approximators.

Remark: Although the 2-LP method is utilized, the $Q(x_k,u_k,w_k)$ function is approximated by CMAC to speed up the convergence and decrease the computational complexity.

The CMAC network simulates the information processing model in the human cerebellum, consisting of many cells stacked on top of each other (Si & Wang, 2001). When receiving external information, only certain cells in the cerebellum are stimulated to interpolate the output using information stored in memory. The value range of input i is

quantized by the CMAC into B_i elements that have an equal resolution width.

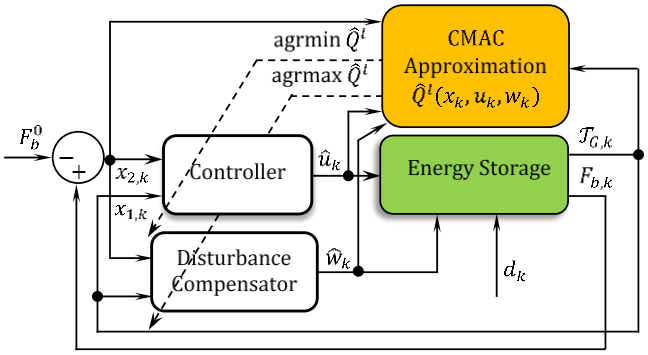

Fig. 2. Q-ROC learning structure.

Next, they are replicated onto K layers, stacked, and then slided over each other by a distance. Every input is successively mapped into quantized values at all layers, and this mapping occurs at each dimension of the input vector.

Figure 3 details an architecture and operation of the CMAC for Q-Learning algorithm. The CMAC is used to approximate a function $Q(x_k, u_k, w_k): \xi \to Q$, $\xi = [x_k, u_k, w_k]^T \in \mathbb{R}^4$ and $Q \in \mathbb{R}$. The CMAC algorithm consists of two mapping steps for determining the value of the output, i.e., $P: A \to Q$, where A is dimensional association space. With $\xi = \left[\xi_1, \xi_2, \xi_3, \xi_4\right]^T$, define $\left[\xi_{i\min}, \xi_{i\max}\right] \in \mathbb{R}, 1 \le \forall i \le 4,$ select the number of resolution elements $N_i, i = 1, ..., 4,$ which is strictly increasing integer numbers.

In the paper, to reduce the computational complexity, CMAC uses receptive field functions as unit impulse functions, then multi-dimensional receptive-field functions, which was used in (Kim & Lewis, 2000), is omitted. When the input ξ receives the values, the CMAC performs a mapping to calculate the output value of the approximation function:

$$\hat{Q}^l(x_k, u_k, w_k) = \hat{W}_c^T \xi \tag{20}$$

where $\hat{W}_c^T$ is the set of the active resolution elements of input vector for layers. It is worth noting that the operating range of the inputs is quantized into values with equal intervals in this paper.

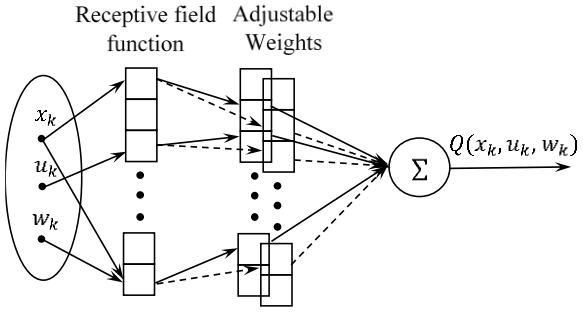

Fig. 3. Architecture of a CMAC for Q-Learning.

Figure 4 describes how to divide the input space into the hyper rectangles, and map the input values into the memory cells in the specific CMAC network with two dimensions and four layers. The weights are stored in the memory cells, and are optimized throughout training. The algebraic total of the weights in all the memory cells activated by the input point is the output of a CMAC. When the value of the input point changes, the number of activated hyper-rectangles changes, and the number of memory cells participating in the CMAC output changes as well.

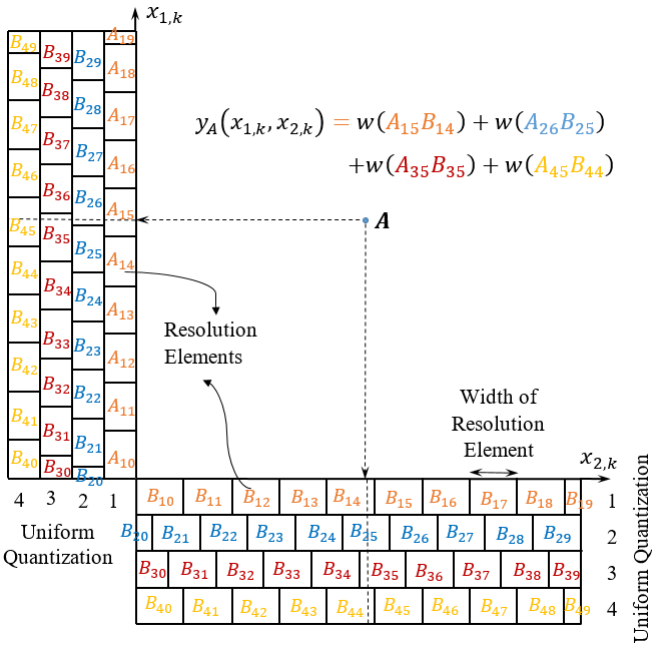

Fig. 4. Two-input CMAC with four layers.

The control law and disturbance compensation law are expressed by:

$$\hat{u}_k = \hat{W}_1^T \phi(V_1^T x_k) \tag{21}$$

$$\hat{w}_k = \hat{W}_2^T \phi(V_2^T x_k) \tag{22}$$

where V_i, $\hat{W}_i$, $i = 1,2$ are the weight vectors of the 2-LP network.

The activate function $\phi(.)$ is chosen as the sigmoid function. The NN weights of the input layer V_i, $i = 1,2$, are not required to update while the weights of the output layer at the iterative step $l, l = 1, 2...$ are updated as follows (Si & Wang, 2001):

$$\hat{W}_i^j(l+1) = \hat{W}_i^j(l) - \alpha_i \frac{\partial E_i^j(l)}{\partial \hat{W}_i^j(l)}, \; i = 1,2 \tag{23}$$

where $E_1^j(l) = 1/2(\hat{u}_k - \arg\min_{U_k} Q^j(x_k, U_k, w_k))^2$, $E_2^j(l) = 1/2(\hat{w}_k - \arg\max_{W_k} Q^j(x_k, u_k, W_k))^2$ and α_i, $i = 1,2,$ be learning rate.

The update law the CMAC weights is designed by:

$$\hat{W}_c^j(l+1) = \hat{W}_c^j(l) - \frac{\alpha_c}{K}\frac{\partial E_c^j(l)}{\partial \hat{Q}^{j,l}(\xi)}\frac{\partial \hat{Q}^{j,l}(\xi)}{\partial \hat{W}_c^j(l)} \tag{24}$$

where α_c is the learning rate, and

$$E_c^j(l) = \frac{1}{2}\Big(S^l(x_k,u_k,w_k) + \gamma \min_{U_k}\max_{W_k} Q^{j-1,l}(x_{k+1},U_k,W_k) - Q^{j-1,l}(x_k,u_k,w_k)\Big)^2 \tag{25}$$

3.3 Algorithm

Based on the preceding preparations, the Q-ROC algorithm, which comes with the structure in Fig. 2, as well as the laws for updating weights in (23), (24), is explained in detail and shown in Algorithm 1.The algorithm's objective is to keep the energy storage system running at its best under solar energy conditions.

Algorithm 1: Q-ROC

Step 1: Initiate $\hat{W}_c(0), \hat{W}_i(0), V_i, i = 1,2, 0 < \gamma < 1$, learning rate $0 < \alpha_c$, α_1, α_2, $\chi < 1$, $\hat{Q}^0 = x_0^T P x_0$ where P is a nonnegative matrix.

Step 2: Let $l = 0$. Put x_k into 2-LP network. The signal from input nodes goes through the hidden nodes forward to the output nodes to obtain results $\hat{u}_k, \hat{w}_k$ for all $i = 1,2$, the input of the hidden unit $k, k = 1, \ldots, n_h$ at hidden layer:

$$net_{k,j}(l) = \sum_{j=1}^{n_i} V_{k,j,i}(l) x_{k,j} \tag{26}$$

Output of the hidden unit:

$$\Phi_{k,i}(l) = a(net_{k,i}(l)) = \frac{1}{1+e^{-net_{k,i}(l)}} \tag{27}$$

Input of the unit $h, h = 1, \ldots, n_0$ at the output layer:

$$net_{h,i}(l) = \sum_{k=1}^{n_h,i} \hat{W}_{h,k,i}(l)\Phi_{k,i} \tag{28}$$

Input of the unit $h, h = 1, \ldots, n_o$ (control law and disturbance compensation law) at the output layer:

$$\hat{u}_k = net_{k,1}(l), \hat{w}_k = net_{k,2}(l) \tag{29}$$

Exploration procedure: $\varepsilon \leftarrow rand[0,1]$. If $\varepsilon < 0.1$, excite the systems using:

$$\hat{u}_k = \hat{u}_k + \xi \tag{30}$$

$$\hat{w}_k = \hat{w}_k + \xi \tag{31}$$

where $\xi = 0.1 rand[-1,1]$.

Step 3: Update the CMAC weights:

$$\hat{Q}^{l+1}(x_k,\hat{u}_k,\hat{w}_k) = \hat{Q}^l(x_k,\hat{u}_k,\hat{w}_k) + \chi(S(x_k,\hat{u}_k,\hat{w}_k) + \gamma\hat{Q}^l(x_{k+1},\hat{u}_{k+1},\hat{w}_{k+1}) - \hat{Q}^l(x_k,\hat{u}_k,\hat{w}_k)) \tag{32}$$

Step 4: Compute:

$$u_{\min} = \arg\min_{\forall u_k} \hat{Q}^{l+1}, \quad w_{\max} = \arg\max_{\forall w_k} \hat{Q}^{l+1} \tag{33}$$

Step 5: Update the weights of 2-LP network:

$$\hat{W}_1(l+1) = \hat{W}_1(l) + \alpha_1(u_{\min} - \hat{u}_k)\Phi^T(x_k), \quad \hat{W}_2(l+1) = \hat{W}_2(l) + \alpha_2(w_{\max} - \hat{w}_k)\Phi^T(x_k) \tag{34}$$

Step 6: If $\left\|\hat{Q}^{l+1}(.) - \hat{Q}^l(.)\right\| < \delta$ (δ: a small positive constant as the condition to stop), stop. Otherwise, $l = l+1$ go to step 2.

Remark 2: In Q-Learning, the system explore procedure is required to make the parameters converge to global optimal values. This procedure is called the ε-greedy policy (Sutton & Barto, 1998). *To this end, the probe noise is used to persistently excite the control and disturbance policies in (30) and (31). By applying the ε-greedy action selection for the online data-driven approach, the overfitting is avoided.*

4. SIMULATION RESULTS

In this section, the practical data is collected, the parameters are set up for the simulations, and the results are analyzed. The Q-ROC algorithm is executed by simulating on the smart grid system, including solar, grid, battery and load. The disturbance compensation policy and optimal control strategy are gained by applying the Q-ROC algorithm. The results are evaluated and compared with another algorithm.

4.1 Data and parameters for simulation

The simulation data, which includes the electricity price, load demand, and solar power, is gathered from statistical sources as real-time data. Following assumption 1, the electricity price, load demand, and solar power are discrete-time periodic functions with a period of 24 hours, so the data utilized for simulation needs to be standardized every 24 hours from the real-time data every 168 hours by taking an average. The sets of the electricity price and load power in 168 hours, which are from ComEd Company (ComEd, 2019) and National Renewable Energy Laboratory (NREL, 2019), respectively, are depicted in Fig. 5a and Fig. 6a. The average sets of electricity price and load power in 24 hours are extracted from these in 168 hours and depicted in Fig. 5b and Fig. 6b. The solar energy depicted in Fig.6a was recorded at the first week of July 2019 in San Francisco (NREL, 2019). The average solar energy is depicted in Fig. 7b (the power generated by a photovoltaic (PV) panel detailed in (Q. Wei et al., 2015).

To compare with the method without considering external disturbance, the battery parameters are chosen as those in (Wei, et al., 2015). Thus, the battery is chosen with the capacity be 14KWh, the rated power $C_{rate} = 3$ KWh, the initial level be 9KWh, the storage limitations (upper and lower) $F_b^{\max} = 10$ KWh, $F_b^{\min} = 2$ KWh, respectively. Let $\gamma = 0.95$, $\alpha = 1$, $\beta = 0.3$, $\delta = 0.2$, $\mu = 5$. Choosing the initial state as

$x_0 = [1,4]^T$ and the positive definite matrix $P = [2.05, 0.11; 0.11, 8.07]$, then the initial Q-function is $\hat{Q}^0 = x_0^T P x_0$. Note that without loss of generality, P can be chosen equal to zero. The external disturbance from grid and solar is $d_k = 0.1\sin(T_{G,k})\cos(T_{R,k})$, of which the amplitude is suitable with the observed data from (NREL, 2019).

From the system (8), the structures of 2-LP networks are chosen with three inputs and one output. One can choose a hidden layer with ten neuron units. It is emphasized that the more number of hidden units we choose the more accuracy is achieved. However, we need balance between accuracy and computational complexity as well as the convergence rate. As usual the weights the networks are initialized in range (0,1), the learning rate is chosen $\alpha_i = 0.001$. Note that for the larger learning rate, the weights quickly converge to their near-optimal values, but the control performance is reduced.

According to (C.S & H., 1995), there is no perfect method for determining the optimal parameters for the CMAC network, such as the number of layers and solution elements. To guarantee a trade-off between the accuracy of the output and the required memory size, we choose the CMAC network with 4 layers, $K = 4$, and 10 resolution elements on each layer, $B_i = 10, i = 1,2$. Therefore, the number of required storage locations is $N_w = K\prod_{i=1}^{2} B_i = 4\times 100 = 400$. It can be seen that although the number N_w for CMAC is large, it requires significantly less storage than a lookup table. On the other hand, despite of larger N_w, the CMAC's converge rate is much faster than the networks of MLP and RBF families.

To guarantee the parameters to converge to the global optimal values, the exploration procedure is required. To this end, the probe noise is chosen by the trial and error method. The suitable probe noise for the exploration in this case is chosen as $\xi = 0.1 rand[-1,1]$ with the probability $\rho = 0.1$. Note that according to the actual system, we can choose the larger exploration probability and gradually reduce it in time. However, if ρ is near to 1, the system may be unstable and control performance may be poor in the early stages. Conversely, if ρ is too small, the control performance will be trapped at local optimal values. In other words, the choice of exploration-exploitation trade-off is difficult and depends on the experience of the designer regarding a particular system.

4.2 Simulation Results

The simulation results of the Q-ROC algorithm with the data and parameters described above are compared with those of the Q-Learning optimal control (Q-OC) algorithm, which does not consider the external disturbance (Q. Wei et al., 2015). The simulation data and parameters of Q-OC algorithm are also set up the same as the Q-ROC algorithm.

Figure 8 and Fig. 9 show the optimal control stratery of the Q-ROC algorithm and the Q-OC algorithm, respectively. The battery's charging/discharging power is accompanied by this optimal control law, which is achieved based on the real-time load demand, price, and solar power in Fig. 5b - Fig. 7b. The battery is in charging state when $\hat{u}_k$ is negative and in discharging state when $\hat{u}_k$ is positive.

Figure 10 and Fig. 11 show the battery's optimal power of the Q-ROC algorithm and the Q-OC algorithm, respectively. With the battery's average power $F_b^0 = \frac{1}{2}(10+2) = 6$ KWh and one of the optimal objectives in the cost function is to guarantee the battery's charging/discharging around F_b^0 to prolong the battery lifetime, the Q-ROC algorithm gives the result that satisfies this objective better than the Q-OC algorithm. The battery's power changes not much higher (charge) or lower (discharge) than F_b^0 in the Q-ROC algorithm, while it changes further away from F_b^0 in the Q-OC algorithm.

Figure 12 and Fig. 13 show the optimal grid power of the Q-ROC algorithm and the Q-OC algorithm, respectively. The load can receive power simultaneously from the solar, the grid, and the battery. Depending on the battery's and load's feedback signals, when the solar power is zero at night, the load requirement is satisfied by the grid power and battery power, or only the grid power satisfies both the load requirement and the battery charge sequentially.

Figure 14 and Fig. 15 show the load demand and total power required to meet the load demand from grid power, solar power, and battery power in two cases of the Q-ROC algorithm and the Q-OC algorithm, respectively. It can be seen that the total power is consistent with the charge and discharge of the battery. Because the Q-ROC algorithm has ability to compensate for external disturbance, total power satisfies the load demand and the most optimal value is obtained. Otherwise, the Q-OC algorithm without the disturbance compensator, total power is higher than the load demand, which leads to wasted power.

5. CONCLUSION

This paper employs the Q-learning algorithm and CMAC to establish a robust optimal control scheme for the renewable energy management system. The controller has the ability to

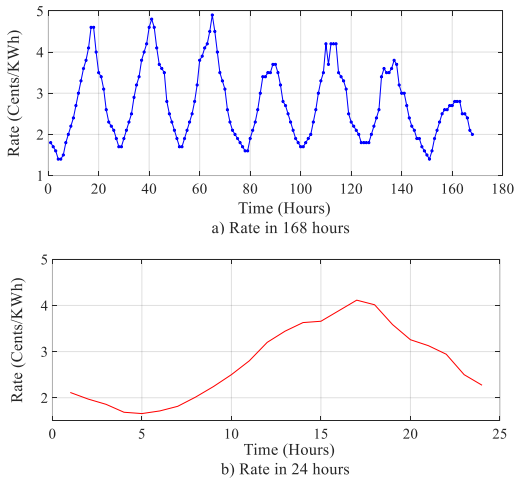

Fig. 5. Electricity price.

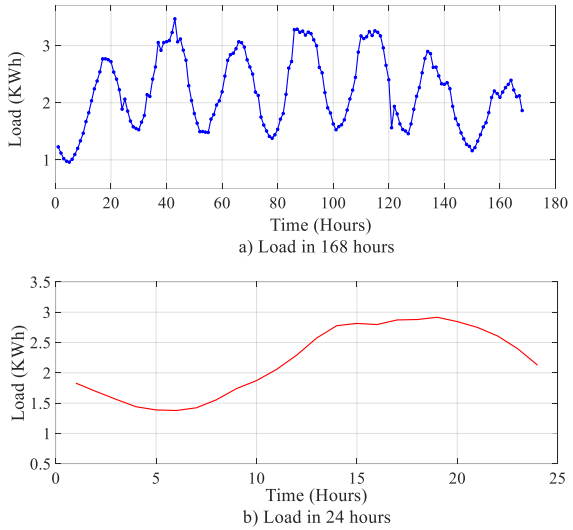

Fig. 6. Load demand.

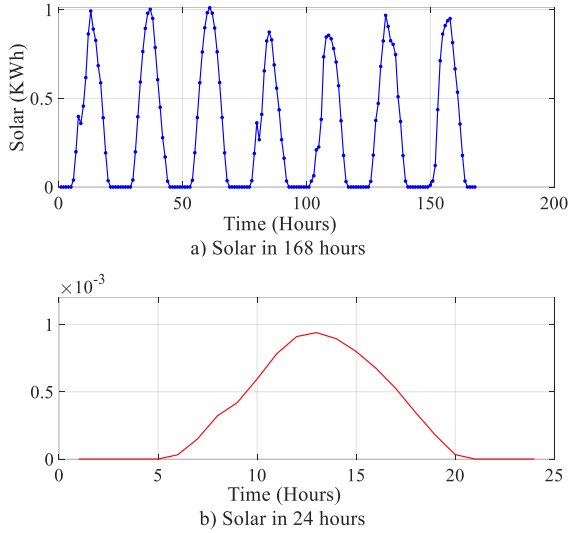

Fig. 7. Solar power.

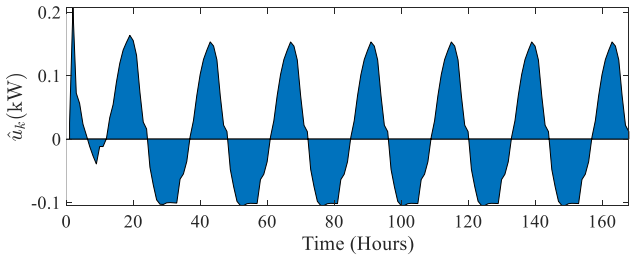

Fig. 8. Optimal control law in Q-OC algorithm.

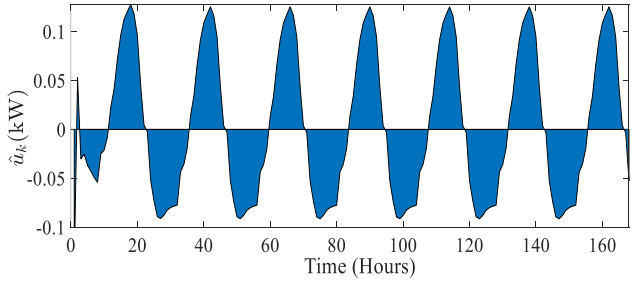

Fig. 9. Optimal control law in Q-ROC algorithm.

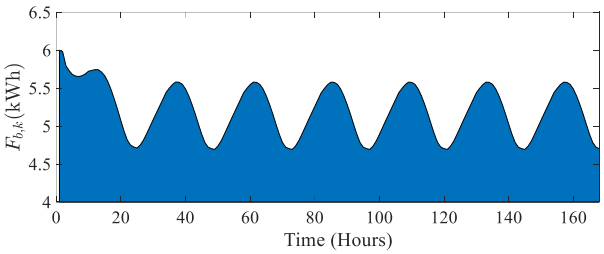

Fig. 10. Battery power in Q-OC algorithm.

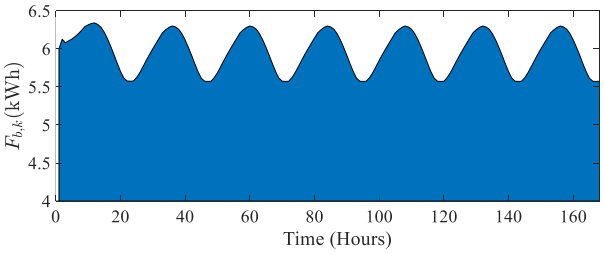

Fig. 11. Battery power in Q-ROC algorithm.

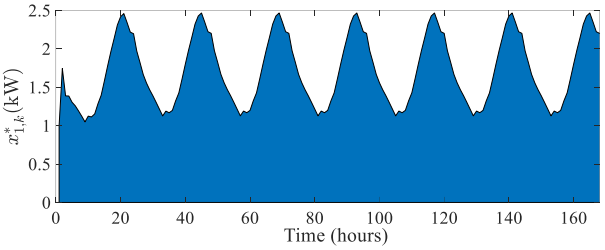

Fig. 12. Optimal grid power in Q-OC algorithm.

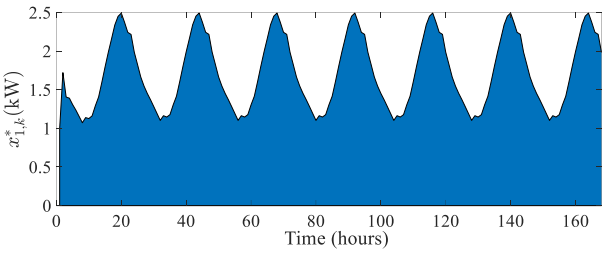

Fig. 13. Optimal grid power in Q-ROC algorithm.

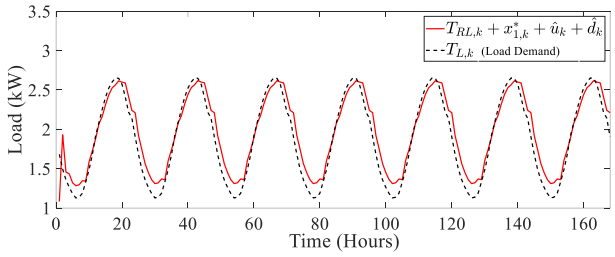

Fig. 14. Optimal load balance in Q-OC algorithm.

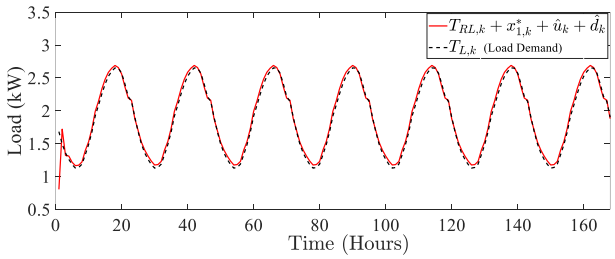

Fig. 15. Optimal load balance in Q-ROC algorithm.

charge and discharge optimally according to the sum of squares of electricity price, battery lifetime, control signals, and disturbance compensation signals. By utilizing the advantages of CMAC, the computational complexity is reduced and the convergence is speeded up, which are required as the important conditions in online control. The external disturbance is compensated for by utilizing ZSG theory in H_∞ control. As a result, the saddle point, including the control policy and disturbance compensation policy, is approximated. According to the practically measured data, the results from simulation for the system, including solar energy, grid, and battery, compared to another method without disturbance rejection, justify the proposed algorithm. Distributed control for the multi-renewable energy system or the multi-battery system will be concentrated on in future work.

REFERENCES

Abu-Khalaf, Lewis, M., F.L. & Huang, J., 2007. H_∞ *Automatic Control, IEEE Transactions on,* 51(12), pp. 1989 - 1995 doi: 10.1109/TAC.2006.884959.

Abu-Khalaf, M. & Lewis, F., 2008. Neurodynamic Programming and Zero-Sum Games for Constrained Control Systems. *IEEE Transactions on Neural Networks,* 19(7), pp. 1243-1252 doi: 10.1109/TNN.2008.2000204.

Angelis, F. D. et al., 2013. Optimal Home Energy Management Under Dynamic Electrical and Thermal Constraints. *IEEE Transactions on Industrial Informatics,* 9(3), pp. 1518-1527 doi: 10.1109/TII.2012.2230637.

Appen, J. v., Stetz, T., Braun, M. & Schmiegel, A., 2014. Local Voltage Control Strategies for PV Storage Systems in Distribution Grids. *IEEE Transactions on Smart Grid,* 5(2), pp. 1002-1009 doi: 10.1109/TSG.2013.2291116.

Bannat, A. et al., 2010. Artificial cognition in production systems. *IEEE Transactions on automation science and engineering,* 8(1), pp. 148-174 doi: 10.1109/TASE.2010.2053534.

Bannat, A. et al., 2010. Artificial cognition in production systems. *IEEE Transactions on automation science and engineering,* Volume 8, pp. 148-174.

Basar, T. & Bernhard, P., 2008. H_∞ Optimal Control and Related Minimax Design Problems DOI: 10.1007/978-0-8176-4757-5.

Berrueta, A. et al., 2020. Identification of Critical Parameters for the Design of Energy Management Algorithms for Li-Ion Batteries Operating in PV Power Plants. *IEEE Transactions on Industry Appl,* 56(5), pp. 4670-4678 doi: 10.1109/TIA.2020.3003562.

C.S, L. & H., K., 1995. Selection of learning parameters for CMAC-based adaptive critic learning. *IEEE Transactions on Neural Networks,* 6(3), pp. 642-647 doi: 10.1109/72.377969.

C.S, L. & H., K., 1995. Selection of learning parameters for CMAC-based adaptive critic learning. *IEEE Transactions on Neural Networks,* Volume 6, pp. 642-647.

Carvalho, A. V., Chouchene, A., Lima, T. M. & Charrua-Santos, F., 2020. Cognitive Manufacturing in Industry 4.0 toward Cognitive Load Reduction: A Conceptual Framework. *Applied System Innovation,* 3(4), p. 55 DOI: 10.3390/asi3040055.

Carvalho, A. V., Chouchene, A., Lima, T. M. & Charrua-Santos, F., 2020. Cognitive Manufacturing in Industry 4.0 toward Cognitive Load Reduction: A Conceptual Framework. *Applied System Innovation,* Volume 3.

ComEd, C., 2019. *Data of electricity rate from ComEd Company.* [Online] Available at: https://hourlypricing.comed.com/pricing-tabletoday [Accessed 2022].

Duan, J. et al., 2019. Reinforcement-Learning-Based Optimal Control of Hybrid Energy Storage Systems in Hybrid AC–DC Microgrids. *IEEE Transactions on Industrial Informatics,* 15(9), pp. 5355-5364 doi: 10.1109/TII.2019.2896618.

Dumitrache, I., Caramihai, S. I., Moisescu, M. A. & Sacala, I. S., 2019. Neuro-inspired Framework for cognitive manufacturing control. *IFAC-PapersOnLine,* 52(13), pp. 910-915 DOI: 10.1016/j.ifacol.2019.11.311.

Dumitrache, I., Caramihai, S. I., Moisescu, M. A. & Sacala, I. S., 2019. Neuro-inspired Framework for cognitive manufacturing control. *IFAC-PapersOnLine,* Volume 52, pp. 910-915.

Frackowiak & SJ, R., 2004. *Human brain function.* 2th ed. s.l.:Elsevier.

Fuselli, D. e. a., 2013. Action dependent heuristic dynamic programming for home energy resource scheduling. *International Journal of Electrical Power and Energy Systems,* 06, Volume 48, pp. 148-160 doi.org/10.1016/j.ijepes.2012.11.023.

Guerrero, J. M., Chandorkar, M., Lee, T. & Loh, P. C., 2013. Advanced Control Architectures for Intelligent Microgrids—Part I: Decentralized and Hierarchical Control. *IEEE Transactions on Industrial Electronics,* 60(4), pp. 1254-1262 doi: 10.1109/TIE.2012.2194969.

Huang, T. & Liu, D., 2011. A self-learning scheme for residential energy system control and management. *Neural Computing and Applications - NCA,* 02, 22(2), pp. DOI: 10.1007/s00521-011-0711-6.

Jiang, C., Liu, F., Chen, S. & Xiao, W., 2019. *Adaptive Dynamic Programming Based optimization Scheduling for Wireless Mobile Charging.* s.l., 15th International Conference on Control and Automation (ICCA).

Jiang, Y. & Jiang, Z., 2014. Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems. *IEEE Transactions on Neural Networks and Learning Systems,* 25(5), pp. 882-893 doi: 10.1109/TNNLS.2013.2294968.

Khan, N. et al., 2021. Batteries State of Health Estimation via Efficient Neural Networks With Multiple Channel Charging Profiles. *IEEE Access,* Volume 9, pp. 7797-7813 doi: 10.1109/ACCESS.2020.3047732.

Kim, Y. & Lewis, F., 2000. Optimal design of CMAC neural-network controller for robot manipulators. *IEEE Trans. Syst. Man Cybern. Part C,* Volume 30, pp. 22-31 DOI:10.1109/5326.827451.

Lai, C. S. & McCulloch, M. D., 2017. Sizing of Stand-Alone Solar PV and Storage System With Anaerobic Digestion

Biogas Power Plants. *IEEE Transactions on Industrial Electronics,* 64(3), pp. 2112-2121 doi: 10.1109/TIE.2016.2625781.

Lewis, F. L., Vrabie, D. & Vamvoudakis, K. G., 2012. Reinforcement Learning and Feedback Control: Using Natural Decision Methods to Design Optimal Adaptive Controllers. *IEEE Control Systems,* 32(6), pp. 76-105 doi: 10.1109/MCS.2012.2214134.

Li, H., D., L. & Wang, D., 2014. Integral Reinforcement Learning for Linear Continuous-Time Zero-Sum Games With Completely Unknown Dynamics. *IEEE Transactions on Automation Science and Engineering,* 11(3), pp. 706-714 doi: 10.1109/TASE.2014.2300532.

Liu, F., Jiang, C., Chen, S. & Xiao, W., 2018. *Multi-sensor scheduling for target tracking based on constrained ADP in energy harvesting WSN.* s.l., 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1579-1584.

Liu, H. et al., 2016. Droop Control With Improved Disturbance Adaption for a PV System With Two Power Conversion Stages. *IEEE Transactions on Industrial Electronics,* 63(10), pp. 6073-6085 doi: 10.1109/TIE.2016.2580525.

Lu, X. & Wang, H., 2020. Optimal Sizing and Energy Management for Cost-Effective PEV Hybrid Energy Storage Systems. *IEEE Transactions on Industrial Informatics,* 16(5), pp. 3407-3416 doi: 10.1109/TII.2019.2957297.

Luy, N., 2017. Adaptive dynamic programming-based design of integrated neural network structure for cooperative control of multiple MIMO nonlinear systems. *Neurocomputing,* Volume 237, pp. 12-24 DOI: 10.1016/j.neucom.2016.05.044.

Luy, N., 2018. Distributed cooperative H_∞ optimal tracking control of MIMO nonlinear multi-agent systems in strict-feedback form via adaptive dynamic programming. *International Journal of Control,* 91(4), pp. 952-968 DOI: 10.1080/00207179.2017.1300685.

M. Boaro et al., 2013. Adaptive Dynamic Programming Algorithm for Renewable Energy Scheduling and Battery Management. *Cognitive Computation,* 06, Volume 5, pp. 264-277 doi.org/10.1007/s12559-012-9191-y.

Mukherjee, S. & Bohra, S. U., 2020. *Lung Cancer Disease Diagnosis Using Machine Learning Approach.* s.l., 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS).

NREL, 2019. *Data of load demand from National Renewable Energy Laboratory (NREL), USA.* [Online] Available at: https://data.openei.org/submissions/153 [Accessed 2022].

NREL, 2019. *Data of solar energy from National Renewable Energy Laboratory (NREL), USA.* [Online] Available at: https://www.nrel.gov/grid/solar-resource/renewable-resource-data.html [Accessed 2022].

Ojand, K. & Dagdougui, H., 2022. Q-Learning-Based Model Predictive Control for Energy Management in Residential Aggregator. *IEEE Transactions on Automation Science and Engineering,* 19(1), pp. 70-81 doi: 10.1109/TASE.2021.3091334.

Perera, A., Kamalaruban & Parameswaran, 2021. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews,* Volume 137, p. 110618 DOI: 10.1016/j.rser.2020.110618.

Petousis, P. et al., 2019. Using Sequential Decision Making to Improve Lung Cancer Screening Performance. *IEEE Access,* Volume 7, pp. 119403-119419 doi: 10.1109/ACCESS.2019.2935763.

Q. Wei et al., 2015. A novel dual iterative Q-learning method for optimal battery management in smart residential environments. *IEEE Transactions on Industrial Electronics,* 62(4), pp. 2509-2518 doi: 10.1109/TIE.2014.2361485..

Rigatos, G., Siano, P. & Selisteanu, D., 2017. Nonlinear Optimal Control of Oxygen and Carbon Dioxide Levels in Blood. *Intelligent Industrial Systems,* 3(2), pp. 61–75 doi: 10.1007/s40903-016-0060-y.

Rigatos, G., Siano, P. & Selisteanu, D., 2017. Nonlinear Optimal Control of Oxygen and Carbon Dioxide Levels in Blood. *Intell Ind Syst,* Volume 3, p. 61–75.

Shafiee, Q. et al., 2014. Robust Networked Control Scheme for Distributed Secondary Control of Islanded Microgrids. *IEEE Transactions on Industrial Electronics,* Volume 61, pp. 5363-5374.

Shuai, H. et al., 2019. Optimal Real-Time Operation Strategy for Microgrid: An ADP-Based Stochastic Nonlinear Optimization Approach. *IEEE Transactions on Sustainable Energy,* 10(2), pp. 931-942 doi: 10.1109/TSTE.2018.2855039.

Si, J. & Wang, Y.-T., 2001. Online learning control by association and reinforcement. *IEEE Transactions on Neural Networks,* 12(2), pp. 264-276 doi: 10.1109/72.914523.

Song, R., Xiao, W., Zhang, H. & Sun, C., 2014. Adaptive Dynamic Programming for a Class of Complex-Valued Nonlinear Systems. *IEEE Transactions on Neural Networks and Learning Systems,* 25(9), pp. 1733-1739 doi: 10.1109/TNNLS.2014.2306201.

Sutton, S. R. & Barto, G. A., 1998. *Reinforcement learning-an introduction..* s.l.:Cambridge, MA: MIT Press..

Tham, C., 1994. *Modular online function approximation for scaling up reinforcement learning.* s.l.:University of Cambridge.

Van der Schaft, A., 1992. L_2-gain analysis of nonlinear systems and nonlinear state-feedback H_∞ control. *IEEE Transactions on Automatic Control,* 37(6), pp. 770-784 doi: 10.1109/9.256331.

Venayagamoorthy, G. K., Sharma, R. K., Gautam, P. K. & Ahmadi, A., 2016. Dynamic energy management system for a smart microgrid. *IEEE Transactions on Neural Networks and Learning Systems,* 27(8), pp. 1643-1656 doi: 10.1109/TNNLS.2016.2514358.

Venayagamoorthy, G. K., Sharma, R. K., Gautam, P. K. & Ahmadi, A., 2016. Dynamic Energy Management System for a Smart Microgrid. *IEEE Transactions on Neural Networks and Learning Systems,* 27(8), pp. 1643-1656 doi: 10.1109/TNNLS.2016.2514358.

Watkins, C., 1989. *Learning from Delayed Rewards.* s.l.:Cambridge University.

Wei, Q., Liu, D., Lin, Q. & Song, R., 2018. Adaptive Dynamic Programming for Discrete-Time Zero-Sum Games. *IEEE Transactions on Neural Networks and Learning Systems,* 29(4), pp. 957-969 doi: 10.1109/TNNLS.2016.2638863.

Wei, Q., Liu, D., Shi, G. & Liu, Y., 2015. Optimal multi-battery coordination control for home energy management systems via distributed iterative adaptive dynamic programming. *IEEE Transactions on Industrial Electronics,* Volume 62, pp. 4203-4214.

Wei, Q., Shi, G., Song, R. & Liu, Y., 2017. Adaptive Dynamic Programming-Based Optimal Control Scheme for Energy Storage Systems With Solar Renewable Energy. *IEEE Transactions on Industrial Electronics,* 64(7), pp. 5468-5478 doi: 10.1109/TIE.2017.2674581.

Werbos, P., 1977. Advanced Forecasting Methods for Global Crisis Warning and Models of Intelligence. *General Systems Yearbook,* 01.Volume 22.

Wu, W. & Biao, L., 2012. Neural Network Based Online Simultaneous Policy Update Algorithm for Solving the HJI Equation in Nonlinear H_∞ Control. *IEEE Transactions on Neural Networks and Learning Systems,* 23(12), pp. 1884-1895 doi: 10.1109/TNNLS.2012.2217349.

Xu, X., Zuo, L. & Huang, Z., 2014. Reinforcement learning algorithms with function approximation: Recent advances and applications. *Information Sciences,* 03, Volume 261, p. 1–31 DOI: 10.1016/j.ins.2013.08.037.

Yang, Q., Jagannathan, S. & Sun, Y., 2015. Robust Integral of Neural Network and Error Sign Control of MIMO Nonlinear Systems. *IEEE transactions on neural networks and learning systems,* 09, 26(12), pp. 3278-3286 doi: 10.1109/TNNLS.2015.2470175.

Yu, H., Zhou, Z. & Wang, Q., 2020. Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model. *IEEE Access,* Volume 8, pp. 86400-86410 doi: 10.1109/ACCESS.2020.2992645.

Zamfirache, I. A., Precup, R.-E., Roman, R.-C. & Petriu, E. M., 2022. Policy Iteration Reinforcement Learning-based control using a Grey Wolf Optimizer algorithm. *Information Sciences,* Volume 585, pp. 162-175 doi.org/10.1016/j.ins.2021.11.051.

Zamfirache, I. A., Precup, R.-E., Roman, R.-C. & Petriu, E. M., 2022. Policy Iteration Reinforcement Learning-based control using a Grey Wolf Optimizer algorithm. *Information Sciences,* Volume 585, pp. 162-175.

Zhang, S. & Xin, H., 2021. *Multi-sensor Scheduling Method for Cooperative Target Tracking Based on ADP.* s.l., 2021 International Conference on Wireless Communications and Smart Grid (ICWCSG).

Zhao, T. & Ding, Z., 2018. Cooperative Optimal Control of Battery Energy Storage System Under Wind Uncertainties in a Microgrid. *IEEE Transactions on Power Systems,* 33(2), pp. 2292-2300 doi: 10.1109/TPWRS.2017.2741672.