Reinforcement Learning for Radio Resource Management of Hybrid-Powered Cellular Networks

Hadi Sayed, Ali El Amine, Hussein Al Haj Hassan, Loutfi Nuaymi, Roger Achkar

Reinforcement Learning for Radio Resource Management of Hybrid-Powered Cellular Networks

Hadi Sayed[1], Ali El-Amine[2], Hussein Al Haj Hassan [1], Loutfi Nuaymi[2], and Roger Achkar[1]
[1]Department of Computer and Communications, Faculty of Engineering, American University of Science and Technology, Beirut, Lebanon.
[2] IMT Atlantique, IRISA, UMR CNRS 6074, F-35700 Rennes, France
Email: hadii.sayed95@gmail.com, ali.el-amine@imt-atlantique.fr, hhajhassan@aust.edu.lb, loutfi.nuaymi@imt-atlantique.fr, rachkar@aust.edu.lb

***Abstract*—In this paper, we consider cellular networks powered by both renewable energy and the Smart Grid. We study the problem of minimizing the cost of on-grid energy while maximizing the satisfaction of users with different requirements. We consider patterns of renewable energy generation, traffic variation and real-time price of grid energy. Knowing that these patterns are all time related, we use Q-learning to extract a common pattern as well as to decide the number of radio resource blocks activated to maximize the users' satisfaction and minimize the on-grid energy cost. Results show that using Q-learning achieves a good tradeoff with more than 75% reduction in energy cost and negligible degradation in users' satisfaction.**

***Index Terms*—Cellular Networks, Reinforcement learning, Renewable energy, Smart grid**

I. Introduction

Cellular networks are witnessing exponential increase in mobile traffic with no sign of slowing down. Based on CISCO VNI, it is forecasted that the global mobile data traffic will witness an increase of seven-fold between 2016 and 2021 [1]. As a fast solution, mobile operators are deploying more base stations (BSs). This imposes serious challenges on mobile operators in terms of both operational and capital expenditures. Focusing on operation expenditure, the energy cost can reach more than 32% of the total operational cost with BSs as the most consuming part of the network [2].

Many approaches are proposed to reduce the energy consumption of cellular networks [3]. Motivated by the temporal variation in traffic, several studies propose activating/deactivating resource blocks of BSs such as [4]. In contrast to switching-off techniques that save energy by shutting down the BS (or switching it to a sleep state), deactivating resource blocks reduce the energy consumption without interrupting the service or causing coverage holes. In addition, mobile operators have already started deploying renewable energy (RE) sources and storage units at the level of BSs [5]. This will allow BSs to consume free and clean energy instead of totally relying on the grid energy. However, deciding the number of active resource blocks will become more challenging due to the intermittent nature of RE, the variation of grid energy price and the different users' requirement. In [6], the authors study the trade-off between user satisfaction and the ratio of RE usage in cellular networks. Determining the number of active resource blocks in a single BS powered by RE and power grid is studied in [7]. The authors aim at minimizing the average grid energy while satisfying users' Quality of Service (QoS), which is expressed in terms of outage probability. In [8], the authors propose a new resource allocation policy for RE-powered BSs to reduce the grid power consumption for mobile video download applications. For more studies please refer to [5], [9]. Although these studies succeed to reduce the on-grid energy consumption, they only consider one type of user traffic with a hard constraint on the achievable user rate as a satisfaction metric. In [10], we study the trade-off between grid energy consumption and users' satisfaction for a single BS. The study considers several types of users with different requirements, where the satisfaction of a user is calculated based on several utility functions depending on the user's type. However, [10] does not provide any algorithm for determining the number of active resource blocks. In [4], we use activating/deactivating resource blocks to help cellular networks provide the Smart Grid (SG) with ancillary services. However, this study considers only a heuristic to determine the number of active resource blocks.

In this paper, we employ Q-learning in order to determine the number of active resource blocks for a RE-powered BS while taking into consideration variation of grid energy cost, availability of RE and traffic variation. As all of these parameters vary based on time, Q-learning is used to extract a common pattern as well as decide the number of active resource blocks for each period of the day to minimize the grid energy cost and maximize the satisfaction of users.

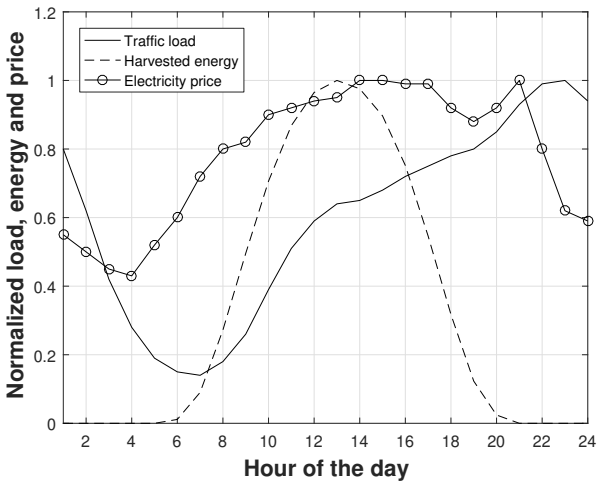

Fig. 1. Typical patterns of traffic [11], grid energy price variation for one day in France [12] and RE generation [13].

The rest of the paper is structured as follows: In Section II, we detail the system model and explain the considered scenario. The proposed approach is presented in Section III. Results are presented and discussed in Section IV before concluding in Section V.

II. System Model

We consider a BS powered by a RE source, a battery and connected to the SG. The BS is responsible of serving different types of users, where N_{HQ}, N_{SQ} and N_{BE} are respectively the number of hard-quality (e.g., voice), soft-quality (e.g., video streaming) and best-effort users (e.g., web). The number of users follows the traffic pattern presented in Figure 1. These users are randomly distributed in the targeted area.

A. Users' satisfaction

Utility functions have been used to measure the satisfaction of users in cellular networks with mixed traffic. Chen et al. proposed a unified utility function for various types of traffic using the sigmoid function [14]. The proposed utility is a sigmoid function that represents user satisfaction with respect to allocated resources, where each application is differentiated by specific utility parameters. The unified sigmoid function for different traffic types is expressed as:

$$u(r) = \frac{1}{\alpha + \beta e^{-\lambda(r-R_0)}} + \gamma \tag{1}$$

where α, β, λ and γ are pre-determined parameters to determine the shape of the utility function depending on the type of traffic. r is the allocated resources or achieved user rate. R_0 is the resource requirement (or required rate) of the user and represents the point of inflexion. For a given utility function, the utility is concave when the allocated resources are smaller than R_0, which means that the user strongly requires the resources of R_0. When the allocated resources are larger than R_0, the utility function is convex, which means that the user does not strongly require the resources. α, β and γ mainly affect the range of the utility. Moreover, adjusting these parameters for different traffic types is essential to have comparable values of the corresponding utilities.

We consider 3 types of traffic: Hard Quality (HQ), Soft Quality (SQ) and Best Effort (BE). HQ traffic, such as Voice, needs a minimum amount of resources. Allocating more resources for SQ traffic increases its utility until certain limit. BE traffic does not require a specific amount of resources ($R_0 = 0$). Based on [14], the utility function of HQ, SQ and BE traffic can be expressed as follows:

$$U_{HQ}(r) = \frac{1}{1 + e^{-\lambda_{HQ}(r-R_0)}} \tag{2}$$

$$U_{SQ}(r) = \frac{1}{1 + e^{-\lambda_{SQ}(r-R_0)}} \tag{3}$$

$$U_{BE}(r) = \frac{1}{1 + \beta e^{-\lambda_{BE} r}} - \frac{1}{1+\beta} \tag{4}$$

B. User rate computation

In our work, we consider that the rate is calculated based on 3GPP TR 36.942 [15]. The received signal power, P_r, to a user from a BS with transmitting power, P_t, is calculated as follows:

$$P_r = 10log_{10}(P_t \times 1000) - (Lo + G_T - G_R) \quad [dBm] \tag{5}$$

where Lo is the path loss calculating using Cost 231 extended Hata model as in [15]. G_T and G_R are respectively the gain of the transmitter and receiver. The detected signal-to-noise (SNR) ratio is calculated as:

$$SNR = P_r - N \quad [dB] \tag{6}$$

where N is the thermal noise power expressed by:

$$N = 10log_{10}(1000 \times KT_0 \times BW) + NF \quad [dBm]. \tag{7}$$

where KT_0 is the thermal density noise, BW is the user's allocated bandwidth, and NF is the noise figure. Using the calculated SNR, the user rate is calculated based on [15].

C. Energy models

1) Base station power model: The power model of the BS follows the EARTH model [11]. The power demand (input power), P_{in}, of a BS consists of a static part P_0 and a load dependent part related to the BS transmitted power P_T. Moreover, the power demand

depends on the number of active resource blocks. It can be expressed as:

$$P_{in} = N_{trx} \times (P_0 + \frac{n}{n_{max}} \times \Delta P \times P_T) \quad (8)$$

where N_{trx} is the number of transceiver chains of the BS, n is the number of active resource blocks, n_{max} is the maximum available resource blocks and ΔP is the slope of the load dependent power part. In order to control the power demand of a BS, we consider that the BS can activate or deactivate its radio resource blocks.

2) Renewable energy model: RE sources are known to have intermittent nature, i.e., energy is not available all the time. We consider the use of solar panels where RE is generated following the RE pattern shown in Figure 1. As shown by the figure, solar panels do not generate energy during the night. Moreover, the amount of energy generation depends on the the time of the day. Indeed, the exact shape of the pattern also depends on the location of the solar panel and the time of the year.

3) Energy storage model: Energy storage units are important in renewable systems to avoid the loss of excess generated RE as well as to compensate their unavailability. We consider that the storage is characterized by a capacity, which is defined as the maximum amount of energy that can be stored, and efficiency, which characterizes the loss due to the energy losses from (dis)charging. Degradation of battery is not considered in this work.

4) Grid energy: Due to the intermittent nature of RE, connecting the BS to the SG is important to ensure continuity of service. We consider that the grid has no constraint regarding the amount of energy procured by the network. However, the grid notifies the network by the price of grid energy each period (hourly in our case). We also consider that the network does not have any knowledge regarding the future price of energy. An example of a typical grid energy price throughout a day in France can be found in Figure 1.

D. Problem formulation

The objective of this work is to minimize the grid energy cost and maximize the utility of the users. We consider a greedy use of RE, i.e., RE is consumed when it is available, stored when the amount of energy harvested exceeds what the BS needs and battery is used when RE is not available. The energy demand of BS depends on the number of active resource blocks as presented in Section II-C1. Thus, the objective of this work is translated into determining the number of resource blocks as follows:

$$\max[\eta \times \frac{(\sum_{k=1}^{N_{HQ}} U_k + \sum_{l=1}^{N_{SQ}} U_l + \sum_{m=1}^{N_{BE}} U_m)}{(N_{HQ}.U_{HQ}^{max} + N_{SQ}.U_{SQ}^{max} + N_{BE}.U_{BE}^{max})} - (1-\eta) \times (\frac{E_{cost}}{E_{cost}^{norm}})] \quad (9)$$

U_K, U_l and U_m represent respectively the utility functions of HQ user k, SQ user l and BE user m. N_{HQ}, N_{SQ} and N_{BE} represent respectively the number of hard quality, soft quality and best effort users. U_{HQ}^{max}, U_{SQ}^{max}, and U_{BR}^{max} represents the theoretical maximum utilities of HQ, SQ and BE users, respectively. E_{cost} represents the cost of energy during the observation period, and $\eta \in [0,1]$ is a parameter that controls the trade-off between users satisfaction and energy cost. We calculate E_{cost} as follows:

$$E_{cost} = E_{grid} \times Price \quad (10)$$

E_{grid} and $Price$ represent the consumed grid energy and the price of grid energy, respectively. E_{cost}^{norm} is used to normalize the cost to be between 0 and 1. The consumed grid energy can be calculated as follows:

$$E_{grid} = \max(0, E - E_{all}) \quad (11)$$

E is the energy demand of the BS. E_{all} is the allocated energy from the battery and renewable source. Moreover, the maximization problem in Equation 9 is constrained as follows:

$$n_{ac} \leq n_{\max} \quad (12)$$

where n_{ac} is the amount of active resources and $n_{\max}$ is the maximum available resource blocks. Moreover, the RE causality constraint can be expressed as follows:

$$E_{all} \leq E_{av} \quad (13)$$

where E_{all} represents the allocated RE (during a period) and E_{av} represents the available RE (at the same period).

III. Proposed approach

In this section, we present the proposed Q-learning algorithm that determines the number of active resource blocks of the BS. The first phase of the algorithm consists of the training phase. During this phase, the BS explores the environment using historical data of traffic load, RE generation and price of the grid energy. These variations change in time following different but specific patterns. In other words, each of the parameters has its specific pattern that is related in time, and Q-learning is used to learn the best action by finding a common pattern. Once this step is complete, the system goes online.

We consider that the state of the BS is represented by periods of the day. The number of states is a trade-off between the accuracy of the decision and the complexity of the problem. On one hand, when a small step duration is considered, the Q-learning algorithm will be able to take more accurate decisions. On the other hand, this will increase the number of states thus, extending the learning time.

We define the set of possible actions $\mathcal{A}$ that corresponds to setting the number of active resource blocks.

The number of possible actions is related to the bandwidth of the BS. An episode starts at the beginning of the day and finishes at the end of the day. During each epoch of an episode (period of a day), the BS chooses an action, then stores a quality-value linking the states $s \in \mathcal{S}$ to the chosen action $a \in \mathcal{A}$. The action consists of choosing the number of active resource blocks in order to maximizes the reward r. We define the reward as the weighted-sum of the average utility of users U_{avg} and the cost of the using grid energy, E_{cost}.

$$r = (1-\eta)U_{avg} - \eta E_{cost}/E_{cost}^{norm} \tag{14}$$

Note that for $\eta = 0$, the problem reduces to maximizing the utility of the users without considering the energy cost, whereas for $\eta = 1$ the problem becomes minimizing the energy cost.

The reward is then used to update the Q-value locally, $Q(s^t, a^t)$, indicating the level of convenience of selecting action a^t when in state s^t (t represent the current period). The Q-value is updated following the update rule:

$$\begin{aligned} Q(s^t, a^t) \leftarrow Q(s^t, a^t) + \\ \alpha[r^t + \gamma \max_{a \in A} Q(s^{t+1}, a^{t+1}) - Q(s^t, a^t)] \end{aligned} \tag{15}$$

where α is the learning rate that represents the speed of convergence, and $\gamma \in [0, 1]$ is the discount factor that determines the current value of the future state costs. During the learning phase, the agent selects the corresponding action based on the ϵ-greedy policy, i.e., it selects with probability $1-\epsilon$ the action associated with the maximum Q-value, and with probability ϵ selects a random action (here y is a random variable uniformly distributed between 0 and 1):

$$a^t = \begin{cases} \arg\max_{a \in \mathcal{A}} Q(s^t, a^t), & \text{if } y > \epsilon \\ rand(\mathcal{A}), & \text{otherwise} \end{cases} \tag{16}$$

By implementing the ϵ-greedy policy, the BS would have explored all possible actions and avoided local minima. For more details on Q-learning please refer to [16]. Algorithm 1 details the general Q-learning training phase used in this mechanism.

Algorithm 1 : Q-Learning Algorithm: Training phase

Initialize Q(s,a) to zero for all $s \in S$, $a \in A(s)$
repeat(for each episode):
  Initialize s
  repeat(for each time t):
    Choose a^t using ϵ-greedy policy in (16)
    Compute r^t
    Update $Q(s^t, a^t)$ using the update rule in (15)
    $s^t \leftarrow s^{t+1}$
  until the end of the day
until convergence

IV. Simulations and Results

We consider a Macro-base station serving a 1200 m radius cell. Mobiles are distributed randomly with equal number of each type of users. The BS operates with 2 GHz carrier frequency and 10 MHz bandwidth. We adopt cost 213 Walfisch-Ikegami for path loss [15] and consider the noise figure as 9 dB. The amount of RE generated is 35% of the BS demand when operating at full load. Table I summarizes the parameters considered in the system.

TABLE I
VALUES AND ASSUMPTIONS.

Parametters	Values and assumptions
Number of sectors in a BS	3
Carrier frequency	2GHz
Number of resource blocks	50
Storage size	4000 Wh
Cell radius	1200 m
Path loss	Cost 231 extended Hata model
GT	15 dBi
GR	0 dBi
Noise figure	9 dB
BS power model	EARTH
Environment	Urban
Maximum number of users	10,20,30,40,50,60

We use the Q-learning algorithm to determine the number of active resource blocks. Patterns of RE generation, price of grid energy and traffic are used for the training phase. Indeed, minor changes in pattern may be observed due to specific events or small errors. Figure 2 presents the evolution of the sum of the maximum Q-value of each state (case $\eta = 0.2$ and maximum number of users = 10). As shown in the figure, the algorithm converges after a large number of episodes. This is not a problem as this step is done offline using patterns derived from historical data. The evolution of the average utility of users and the grid energy cost are presented in Figure 3 and Figure 4, respectively. At the beginning of the training phase, the average utility and the cost take wide range of values. Then the range (for each of utility and cost) starts to shrink. Indeed, the minor changes of average utility and cost after a large number of episodes is due to the random distribution of users.

Figure 5 presents the number of resource blocks in each period (hour) of the day when the maximum number of users is 10. From the figure we can see that the highest number of active resource blocks is observed at the time where traffic is high (19-23). Although we have high traffic at time 21, we observe lower number of active resource blocks. This is because the price of grid energy is still high and stored RE has been used. After time 21, we see that the number of active RBs

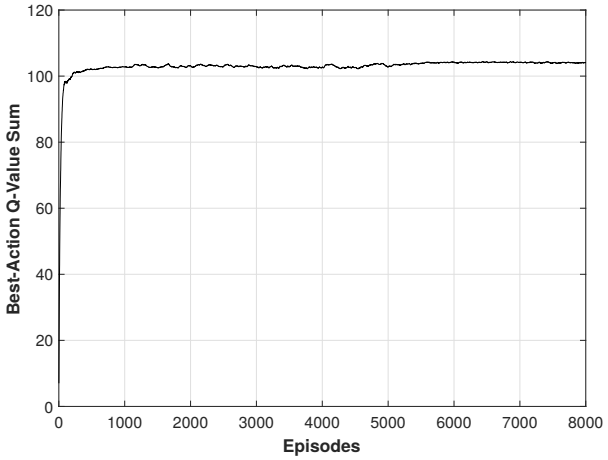

Fig. 2. Convergence of the Q-Learning algorithm for $\eta = 0.2$.

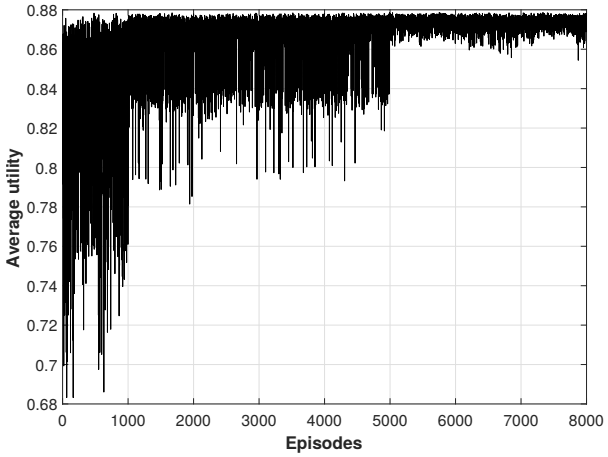

Fig. 3. Evolution of average utility of users for $\eta = 0.2$.

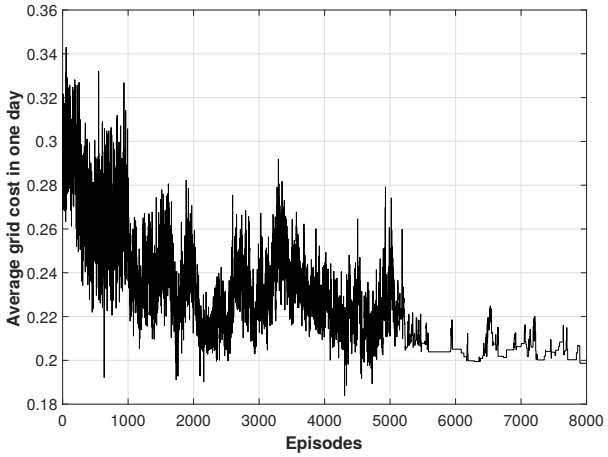

Fig. 4. Evolution of grid energy cost for $\eta = 0.2$

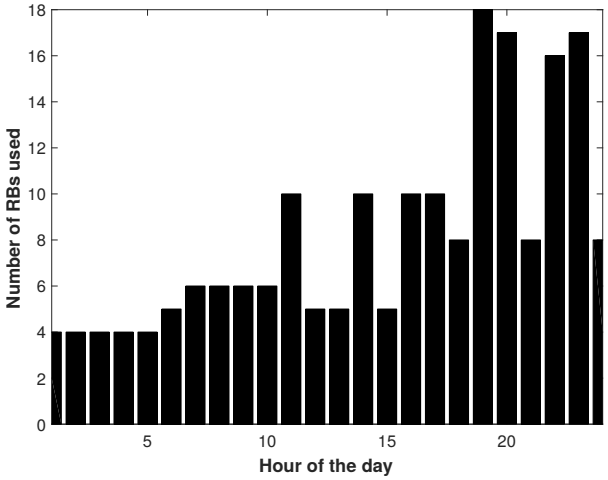

Fig. 5. Hourly number of resouce blocks activated by Q-learning algorithm.

increases again. This is due to the fact that the price of the grid energy decreased. It is worth noting that Figure 5 is one of the possible solutions. By repeating the training, we have noticed several possible solutions. These solutions are slightly different in terms of the activated radio resource blocks but lead to almost the same values of average utility and grid energy cost.

To study the effect of parameter η, we present the variation of cost gain, with respect to the case of no RE, and average utility of users in Figure 6 and Figure 7, respectively. We can notice that no significant change in the value of utility or cost for small values of η. For a specific number of users, the cost gain starts to increase and the average utility of users starts to decrease after a certain value of η (e.g., $\eta = 0.6$ for 10 users). This threshold decreases with the increase in the maximum number of users. An example of the tradeoff (case of maximum of 10 users) is shown in Figure 8.

Table II shows the utility of users and cost of grid energy for the proposed algorithm and existing algorithms (maximum number of users is 20). For the existing algorithms, we consider: 1) Reference: BS operates at the maximum number of available resource blocks without using RE; 2) Greedy: only RE is used in a greedy way (use when available); 3) Traffic, RE and Price of grid energy (TRP) aware algorthim: it determines the number of active resource blocks by discretization the state of a BS based on RE generation, price of grid energy and traffic, similar algorithms are found in [17] and [18]. From Table II, we can see that the best utility is achieved by the reference and greedy, since all the resource blocks are activated. However, the price is very high compared with other algorithms. TRP reduces the cost of grid energy, but this comes with a price in terms of average utility of users. The proposed algorithm ($\eta = 0.2$) achieves significant reduction in the cost with better value of average utility than TRP. The proposed algorithm ($\eta = 0.8$) further reduces the cost but with the worst utility among all algorithms. In summary, the proposed algorithm ($\eta = 0.2$) achieves the best tradeoff

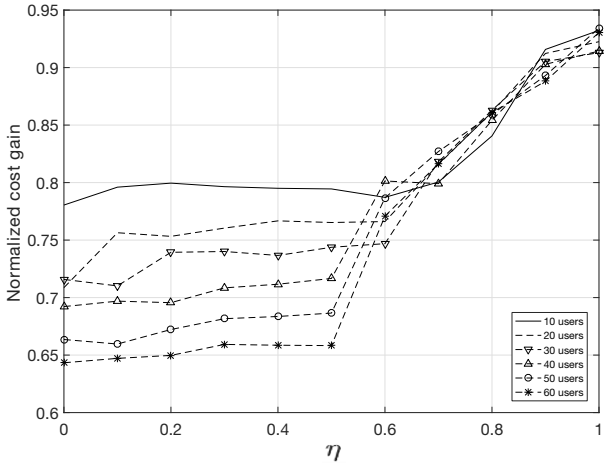

Fig. 6. Energy cost gain with respect to η for several numbers of maximum users.

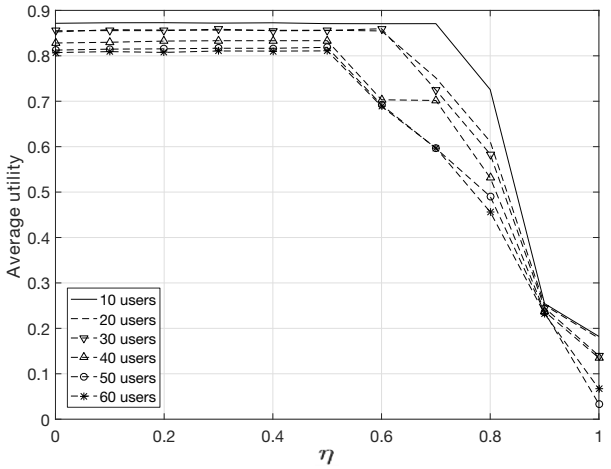

Fig. 7. Average utility of users with respect to η for several numbers of maximum users.

with large reduction in the cost and slight decrease in the user utility.

V. Conclusion

With the increase in traffic requirements and demand, satisfying users and minimizing the operational cost have became two important objectives to be achieved in future cellular networks. In this paper, we study the problem of determining the number of resource blocks that

TABLE II
Utility and cost of simulated algorithms.

Algorithm	$U_{average}$	$Ecost_{mean}$
Reference	0.89	1.491
Greedy	0.89	1.05
TRP	0.78	0.745
Qlearning ($\eta = 0.2$)	0.85	0.29
Qlearning ($\eta = 0.8$)	0.6	0.22

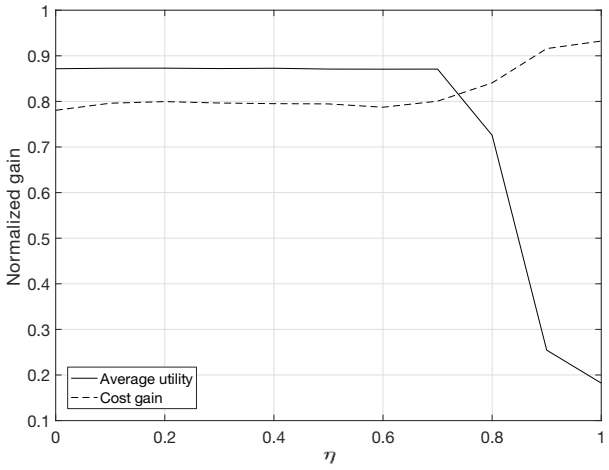

Fig. 8. Trade-off between cost gain and average utility of users with respect to η in case of maximum of 10 users.

combines these two objectives. We consider different types of users with different requirements. Motivated by the fact that parameters of the problem such as renewable energy generation, traffic load and price of grid energy vary with time following specific patterns, we use Q-learning to determine the hourly number of active resource block that fits these patterns. Results show that Q-learning can find a good tradeoff where significant cost reduction is achieved with negligible degradation in users' satisfaction.

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021*. Cisco, Tech. Rep., February 2017.

[2] Juejia Zhou et al. Energy source aware target cell selection and coverage optimization for power saving in cellular networks. In *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, Dec 2010.

[3] Luis Suarez et al. An overview and classification of research approaches in green wireless networks. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):1–18, 2012.

[4] Hussein Al Haj Hassan et al. A novel energy model for renewable energy-enabled cellular networks providing ancillary services to the smart grid. *IEEE Transactions on Green Communications and Networking*, 2019.

[5] Hussein Al Haj Hassan et al. Integrating cellular networks, smart grid, and renewable energy: Analysis, architecture, and challenges. *IEEE access*, 3:2755–2770, 2015.

[6] Mauro Carreno and Loutfi Nuaymi. Renewable energy use in cellular networks. In *2013 IEEE 77th vehicular technology conference (VTC Spring)*, pages 1–6. IEEE, 2013.

[7] Jie Gong et al. Energy-aware resource allocation for energy harvesting wireless communication systems. In *VTC spring*, 2013.

[8] Po-Han Chiang et al. Renewable energy-aware video download in cellular networks. In *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1622–1627. IEEE, 2015.

[9] Hussein Al Haj Hassan et al. Renewable energy in cellular networks: A survey. In *Online Conference on Green Communications (GreenCom), 2013 IEEE*, pages 1–7. IEEE, 2013.

[10] Hussein Al Haj Hassan et al. Grid energy consumption of mixed-traffic cellular networks with renewable energy sources. In *Green Communications (OnlineGreenComm), 2016 IEEE Online Conference on*, pages 1–6. IEEE, 2016.

[11] Gunther Auer and et al. How much energy is needed to run a wireless network? *Wireless Communications, IEEE*, 18(5), October 2011.

[12] Epex, european power exchange spot information,www.epexspot.com, last checked: 1 february 2019.

[13] Pvwatts calculator, national renewable energy laboratory, pvwatts.nrel.gov, last checked:1 april 2019.

[14] Li Chen et al. Utility-based resource allocation for mixed traffic in wireless networks. In *Computer communications workshops (INFOCOM WKSHPS), 2011 IEEE conference on*, pages 91–96. IEEE, 2011.

[15] 3gpp. evolved universal terrestrial radio access (e-utra); radio frequency (rf) system scenarios; v8.0.0 (release 8). tr 36.942, 3rd generation partnership project (3gpp),. 2008-09.

[16] Richard S Sutton et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[17] Hussein Al Haj Hassan et al. Classification of renewable energy scenarios and objectives for cellular networks. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, pages 2967–2972. IEEE, 2013.

[18] Ali El-Amine et al. Services kpi-based energy management strategies for green wireless networks. In *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–7. IEEE, 2018.