Reinforcement Learning-Based Resource Allocation for M2M Communications over Cellular Networks

Sree Krishna Das*, Md. Siddikur Rahman†, Lina Mohjazi‡, Muhammad Ali Imran‡, and Khaled M. Rabie⋆
*Dept. of Electrical, Electronic and Communication Engineering, Military Institute of Science and Technology, Bangladesh
†Dept. of Electrical and Electronic Engineering, American International University-Bangladesh, Bangladesh
‡James Watt School of Engineering, University of Glasgow, UK
⋆Department of Engineering, Manchester Metropolitan University, UK
Email: skshna@yahoo.com, siddikur.sagor@gmail.com, l.mohjazi@ieee.org, muhammad.imran@glasgow.ac.uk, k.rabie@mmu.ac.uk

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771998

***Abstract*—The spectrum efficiency can be greatly enhanced by the deployment of machine-to-machine (M2M) communications through cellular networks. Existing resource allocation approaches allocate maximum resource blocks (RBs) for cellular user equipments (CUEs). However, M2M user equipments (MUEs) share the same frequency among themselves within the same tier. This results in generating co-tier interference, which may deteriorate the MUE's quality-of-service (QoS). To tackle this problem and improve the user experience, in this paper, we propose a novel resource utilization policy, which exploits reinforcement learning (RL) algorithm considering the pointer network (PN). In particular, we design an optimization problem that determines the optimal frequency and power allocation needed to maximize the achievable rate performance of all M2M pairs and CUEs in the network subject to the co-tier interference and QoS constraints. The proposed scheme enables the user equipment (UE) to autonomously select an available channel and optimal power to maximize the network capacity and spectrum efficiency while minimizing co-tier interference. Moreover, the proposed scheme is compared with traditional spectrum allocation schemes. Simulation results demonstrate the superiority of the proposed scheme than that of the traditional schemes. Moreover, the convergence of the proposed scheme is investigated which reduces the computational complexity (CC).**

***Index Terms*—M2M communications, resource allocation, throughput, pointer network, reinforcement learning.**

I. Introduction

By 2020, 4 billion devices are linked over 25 billion embedded intelligent systems creating 50 trillion GB of data [1]. Following these figures, internet of things (IoT), in particular wireless IoT, are potential candidates for future smart world. Its vast adoption puts forward several technical challenges which include network design and storage architecture for smart devices, effective information transmission protocols, proactive IoT device identification, malicious attack prevention, technology standardization and appliance interfaces. As a result, machine-to-machine (M2M) communications is deemed as a promising paradigm in addressing these issues and offering efficient operation of beyond fifth generation (B5G) and sixth generation (6G) cellular networks. Besides, unlike conventional communications, M2M communications involve direct links transmission with the evolved node B (eNB), resulting in many benefits, such as enhancing user data rate, decreasing power consumption, and reducing end-to-end (E2E) latency [2]. For M2M user equipments (MUEs) coexisting with cellular user equipments (CUEs), there are two different types of deployment such as (i) overlaying mode, and (ii) underlaying mode. CUEs and MUEs share the same radio resources through the underlaying mode. They suffer interference with each other. In the overlaying mode, dedicated spectrum resources are allocated without creating cross-tier interference. However, since a high number of users exist in the wireless network, the radio resources are usually inadequate [3], [4]. In order to entirely exploit the possible facilities of underlaid M2M communications, it is essential to provide the proper power for each UE by using the power control scheme and designing an efficient machine learning (ML)-based resource utilization policy that mitigates the co-tier and cross-tier interference between MUEs and CUEs.

A. Related Works

In recent years, there has been a great deal of attention to stochastic optimization and robust optimization methods because of the needs to handle the unpredictable value of CSI in M2M communications. The studies in [4] addressed CSI by maximizing the predicted linking capacity through the use of stochastic optimization. In addition, accurate CSI is often not possible or may require high feedback rates which makes the channel condition uncertain. [2], [5] studied how to maximize the resource efficiency (RE) in M2M communications by using the technique of joint power control and spectrum allocation to improve the user's data rate and prolong the battery lifetime of UE by facilitating the reuse of radio resources between MUEs and CUEs. In real time signal transmission, the size and shape parameters of the uncertainty set would fluctuate with the channel conditions [6], [7]. That is, CSI frequently changes due to the high mobility of CUEs and MUEs. However, the preceding works presented so far emphasize on M2M communications with ideal CSI [3]. Therefore, it is critical to investigate how to meet the increasing demand of higher transmission rate in M2M communications.

B. Motivation

Next generation wireless networks will generate a tremendous amount of data related to network statistics, such as user traffic, channel occupancy, channel quality, etc. This will induce unmanageable overhead that largely increases delay, computation, and energy consumption of network elements [8]. Neural network (NN) can leverage this data to develop automated and fine-grained schemes to optimize network radio resources. The multipurpose pointer network (PN) can predict sequences over variable length input dictionaries when it integrates with NN resulting in improvement of sequences with the assistance of input attention and generalization of variable size output dictionaries. However, the PN based on the NN is not taken into account in [7], and interferences are not taken into account in [4]. Furthermore, resource allocation techniques based on the conventional optimization theory, such as multidimensional knapsack problems (MKP), greedy algorithm, and heuristic algorithm [6], [7], are generally highly complex and not feasible for real-time applications. Reinforcement learning (RL) is deemed as a promising algorithm to solve cellular communication problems, especially for spectrum allocation, data offloading, adaptive modulation, power control and interference mitigation more efficiently compared to supervised and unsupervised learning algorithms. However, RL algorithms reveal low convergence speed as well as overall efficiency while working with large state–action spaces in complex networks. Moreover, in combined resource sharing and power controlling schemes, RL is unable to manage large action spaces and state space. Therefore, this paper considers RL-based PN for solving multi-dimensional state space and complexity discrete action space problems and proposes a joint power and spectrum allocation algorithm. The key contributions of this paper can be summarized as follows.

- This paper proposes a RL-based resource utilization policy for M2M communications over cellular networks.
- We adopt a mixed integer non-linear programming problem and NP-hard. Then, we assign the orthogonal subfrequenices for different MUEs and CUEs to maximize the sum rate of the network. Moreover, the M2M pairs are permitted to reuse resources to better use the scarce resources, when the co-tier interference is below than threshold interference. Besides, MUEs can choose a number of channels and proper power to transmit services as soon as possible without affecting the traditional communication of CUEs.
- This paper considers the RL algorithm empowered PN architecture which is based on a low-complexity process to effectively utilize the spectrum resources. Besides, the PN is a prominent type of NN which efficiently solves combinatorial optimization related problems in M2M communications. Furthermore, this method achieves the goal of significantly improving the QoS of the system, such as optimizing system capacity and simultaneously reducing interference.
- Extensive simulations are carried out for evaluating the resource utilization policy of the proposed scheme. It is also found that the proposed scheme reduces the computational complexity (CC). Also, the proposed scheme provides better network performance than that of the traditional schemes.

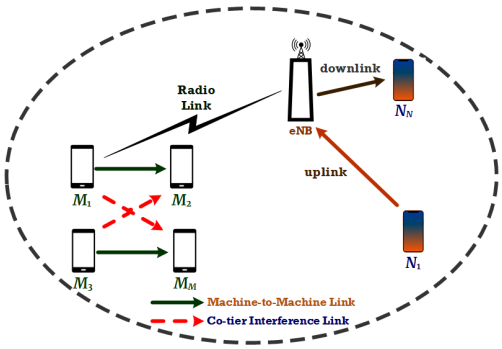

Fig. 1. Proposed system model of M2M communications over cellular networks.

The rest of the paper is organized as follows. Section II illustrates the system model. We put forward the radio resource utilization policy and develop the RL algorithm with low CC in section III. Section IV presents simulation results, followed by conclusions in Section V.

II. Proposed Network Model

A. System Model

The system model considers the downlink data transfer situation when the eNB is located in the center of the cellular cell. The M2M communication comprises of a couple of devices those are able to directly transmit data without the help of the eNB. On the other hand, the CUEs are mobile terminals that can only be connected via the eNB. There are M M2M pairs and N CUEs deployed randomly in the coverage area of the eNB as shown in Fig. 1. UEs operate in orthogonal frequency bands following orthogonal frequency division multiple access (OFDMA) technique. In this case, MUEs do not share the spectrum with CUEs. The unavailability of resource blocks (RBs) allows the MUEs to share the same frequency but it also causes co-tier interference between MUEs. Moreover, we assume that the complete CSI is accessible [9]. In other words, for simplicity, the eNB is capable of obtaining the full CSI between CUEs and M2M pairings.

B. Performance Metrics

As OFDMA is incorporated for CUE and MUE transmissions, MUE receivers are subject to the interference caused only by other MUE transmitters that reuse the same frequency. In this sense, the co-tier interference for the MUE receiver at subfrequency k is given as follows

$$I_k = \sum_{m=1}^{M} v^*_{m,k} p^*_{m,k} h^*_{m,k}, \tag{1}$$

where $v^*_{m,k} \in \{0,1\}$, designates whether the subfrequency k is allocated to M2M pair m, if the m M2M pair reuses the

subfrequency k, $v^*_{m,k}$ sets 1 as its value, else sets 0. Also, $p^*_{m,k}$ denotes the transmit power while $h^*_{m,k}$ is the channel gain on subfrequency k between the MUE receiver and other MUE transmitters including the 3^{rd} generation partnership project (3GPP) path loss (PL) model, channel fading on both M2M pairs and CUEs that follows the Rayleigh distribution with uniform variance. The signal to interference-plus-noise ratio (SINR) expression for CUE n and the M2M pair m at subfrequency k can be written as

$$\gamma_n^k = \frac{p_n^k h_n^k}{B_n N_0}, \tag{2}$$

and

$$\gamma_m^k = \frac{p_m^k h_m^k}{I_k + B_m N_0}, \tag{3}$$

respectively, where h_n^k and h_m^k represent the channel power gain on subfrequency k between M2M pair and CUE and eNB, respectively and N_0 is the noise power. Besides, B_n and B_m denote the allocated spectrum resources of the CUE and M2M pair, respectively. Moreover, p_n^k and p_m^k denote the transmit power of n CUE and m M2M pair, respectively. Therefore, the total achievable data rates for n^{th} CUE and m^{th} M2M pair, respectively can be stated as

$$R_n = \sum_{n=1}^{N}\sum_{k=1}^{K} B_n w_n^k \log_2(1+\gamma_n^k), \tag{4}$$

and

$$R_m = \sum_{m=1}^{M}\sum_{k=1}^{K} B_m v_m^k \log_2(1+\gamma_m^k), \tag{5}$$

respectively, where w_n^k and v_m^k denote whether or not the subfrequency k is assigned to n^{th} CUE and m^{th} M2M pair, respectively.

III. RL-Based Resource Allocation Policy

A. Overview

RL-based resource optimization: This subsection first formulates an optimization problem that maximizes the proposed network performance in terms of spectrum efficiency (SE) [9]. In addition, a RL-enabled lower CC algorithm is presented to address the resource scarcity.

Problem formulation: Since the M2M pairs dispatched information opportunistically [7], a significant amount of control signals are required by a proper power control system. We assume that an M2M pair consumes constant power to transmit data during its operation. In other words, MUEs transmitting power allocated at each subfrequency is given by

$$p_m^k = \begin{cases} 0, \text{if M2M pair is inoperative at } k \\ P, \text{if M2M pair is operative at } k \end{cases} \tag{6}$$

We designate $S_k = \{m | v_m^k = 1\}$ to present the set of working M2M pairs at subfrequency k. To maximize the system performance of both MUE and CUE, the optimization problem is expressed as

$$OP^1 : \max_{\{W,V,P\}} \{R_m + R_n\} \tag{7}$$

$$\sum_{k=1}^{K}\sum_{n=1}^{N} w_n^k \le 1, \forall n \in N, \forall k \in K \tag{8}$$

$$w_n^k \in \{0,1\}, \forall k, n, \tag{9}$$

$$v_m^k \in \{0,1\}, \forall k, m, \tag{10}$$

$$0 \le \sum_{n=1}^{N}\sum_{k=1}^{K} p_n^K \le P^{max}, \tag{11}$$

$$\gamma_n^k \ge \gamma_{th}, \forall k, n, \tag{12}$$

$$\gamma_m^k \ge \gamma_{th}, \forall k, \forall m \in S_k, \tag{13}$$

where $W = [w_n^k]_{N\times K}$ is the N by K subfrequency allocation matrix for CUEs, $V = [v_m^k]_{M\times K}$ is the M by K subfrequency allocation matrix for M2M pairs and $P = [p_n^k]_{N\times K}$ is the matrix of transmission power of CUEs at all subfrequencies. Constraints (8) and (9) ensure that a maximum of one CUE is allocated each subfrequency. Moreover, constraint (11) is utilized to limit the UE's transmit power. On the other hand, constraints (12) and (13) present the SINR constraints for the CUEs and MUEs, respectively. Here, γ_{th} is the threshold for indicating the minimum required SINR to guarantee the QoS requirements for the CUEs and M2M pairs, respectively.

B. Proposed Low Complexity Algorithm

From (7), it can be seen that OP^1 is a non-linear programming optimization and mixed integer problem, and, it is generally NP-hard [9]. In particular, the proposed RL technique can solve complex network resource optimization problems and take judicious control decisions with only limited information about the network statistics rather than existing ML techniques. Consequently, there are two subproblems, namely OP^2 and OP^3, which can be formulated based on to leverage (OP^1), and evolve a lower CC algorithm to solve it. The proposed RL-based resource allocation policy comprises the allocation for CUEs within the constraints of the eNB transmission power and orthogonal subfrequency initially, and after that the allocation for M2M pairs under threshold interference.

1) Spectrum allocation for each CUE: The first subproblem (OP^2) seeks to optimize the overall performance of each CUE by presuming that orthogonal subfrequencies do not suffer from co-tier and cross-tier interference resulting from M2M pairs, i.e.

$$OP^2 : \max_{\{W,P\}} R_n, \tag{14}$$

$$\text{s.t. } (8), (9), \text{ and } (11).$$

We can maximize the transmission rate of CUEs for each subfrequency to acquire the maximum throughput of all CUEs. Hence, the maximum SINR constraint is achieved for CUEs. In addition, the CUE with the leading channel gain is allowed to transfer data on each subfrequency. Furthermore, the transmission power of CUE n^* is managed with the open loop

power control technique to mitigate the interference at the M2M receivers, such as

$$p_{n^*}^k \geq \frac{\gamma_{th}(I_{max}^k + N_0)}{h_{n^*}^k}, \tag{15}$$

where I_{max}^k is the highest permitted interference at each subfrequency.

2) Spectrum allocation for each MUE: After finding the allocation for each CUE, then the subfrequency assignment for M2M pairs can be given as

$$OP^3 : \max_{\{V\}} R_m, \tag{16}$$

$$\text{s.t. } (10), (12), \text{ and } (13).$$

From equation (2), (12) is rewritten as follows

$$\min_n \left\{ \frac{p_n^k h_n^k}{\gamma_{th}} - N_0 B_n \right\} \geq 0. \tag{17}$$

Similarly, (13) is rewritten as

$$\sum_{m=1}^{M} v_{m,k}^* p_{m,k}^* h_{m,k}^* \leq \min_{m \in S_k} \left\{ \frac{P h_m^k}{\gamma_{th}} - N_0 B_m + v_m^k P h_m^k \right\}. \tag{18}$$

After the alteration of (12) and (13), (OP^3) is transfigured into a two-dimensional (2D) knapsack problem (KP), especially, there are two dimensions for the weights of the KP. It is applied to achieve the optimal solution for resource allocation in M2M communications under power and threshold interference constraints [9]. This type of issue is a part of the MKP [9] and only water-filling algorithms can be applied to address this type of NP-hard problems. However, efficient water-filling algorithms require high computation time, making them unsuitable for real-time applications [5]. In this article, the PN is employed to successfully handle the problem of combinatorial optimization.

C. Proposed Algorithm Representation

1) PN architecture: RL-based optimal resource allocation scheme is proposed in Algorithm 1. As above, a 2D KP is a resource optimization problem for each subfrequency k. Given the CUEs' resource distribution state, each of the M2M pairs is a characteristics vector of 3D (v, x, y), where v is the achievable data rate for M2M pairs according to (5), x and y are the weights on the 2D KP limitations, which may be obtained from (17) and (18), individually. The PN is a special type of recurrent NN which differentiates the encoder and the decoder via distinct colors. The input of v, x and y should be in a sequence that comprises the 3D characteristics vectors as specified, because the PN is built on the model of sequence-to-sequence. The output is also a sequence and may be obtained from the PN by using the pointing mechanism which re-arranges the input. The output is a collection of valid entities that meet the requirements. In particular, we cross the solution series and terminate when the obtained entities exceed the limitations of (17) and (18). The identified entities are the solution to the KP. We name it solution o and designate $V(o)$ as the total value of the corresponding set of the entities. The details of the PN framework are provided in [9].

Algorithm 1: RL-based resource utilization policy

```
Step 1: Spectrum distribution for CUEs;
Set n = [1, 2, ..., N], m = [1, 2, ..., M] and entire
  number of simulation times;
Initialize : R_m = 0, R_n = 0 and p_n^k = P^max / N;
for k = [1, 2, ..., K] do
    Obtain n* on subfrequency k;
    Update the optimal transmission power of each
      CUE according to (15);
    Update the data rate of each CUE according to
      (17);
end
Step 2: Spectrum distribution for MUEs;
Assign the training set S, amount of training phases T,
  batch size Q and PN parameter P;
for t = [1, 2, ..., T] do
    Choose a batch of sample s_b for b ∈ [1, 2, ..., Q];
    Trial solution o_b based θ_p(.|s_b) for b ∈ [1, 2, ..., Q];
    Calculate value: V(o_b|s_b);
    Update the PG (g_p) according to (22);
    Update the parameter of the PN according to (23);
    Update the baseline function
      q(s_b) = q(s_b) + α(V(o_b|s_b) − q(s_b)) for
      b ∈ [1, 2, ..., Q];
end
for k = [1, 2, ..., K] do
    Utilize the PN to calculate the v_m^k for m^th MUE;
    Update the date rate for each MUE;
end
```

2) RL algorithm: RL algorithm is considered as a proper method to train NN while solving combinatorial optimization problems. Our proposed low complexity policy-empowered RL aims the parameter optimization a PN, which is symbolized as P. Besides, the expected tour length expressed by an input graph s is given below [10]

$$J(P|s) = E_{o \sim \theta_p(.|s)} V(o|s). \tag{19}$$

The graphs generate from distribution S while training, where the overall training target includes sampling from the graph distribution and written as follows

$$J(P) = E_{s \sim S} J(P|s). \tag{20}$$

For optimizing the parameters, we recourse to policy gradient (PG) techniques and stochastic gradient descent. Using the RL algorithm, the gradient of (20) can be expressed as [10]

$$\nabla_p J(P|s) = E_{o \sim \theta_p(.|s)}[(V(o|s) - q(s)) \nabla_p \log \theta_p(o|s)], \tag{21}$$

where $q(s)$ signifies the baseline act in the training procedure that does not lie in the arrangement of the order in the proposed network and calculates the predicted value to decrease the divergence of the gradients [10]. The popular RL algorithm has been utilized to extract the gradient for improving the

network parameters using the Adam method [10]. The proposed resource allocation scheme uses the RL algorithm in a further practicable approach with Monte Carlo sampling. After assigning the batch size Q, by producing Q independently and identically allocated sample KP, the gradient is stated in a randomly determined mean form as

$$g_p = \frac{1}{Q}\sum_{b=1}^{Q}(V(o_b|s_b) - q(s_b))\nabla_p \log \theta_p(o_b|s_b), \quad (22)$$

$$P = ADAM(P, g_p), \quad (23)$$

The baseline function is an exponential moving mean value of the system rewards achieved through the network upon time to justify the detail that the strategy enhances with training.

3) CC analysis: Here we will analyze the CC of the proposed scheme. According to Step 1, the CC of the suggested algorithm 1 is $\mathcal{O}(N)$. The long-short term memory (LSTM) units with attention are the elementary modules in the PN of the Step 2. The CC for every given fine-tuned PN is $\mathcal{O}(M^2)$ whereas the attention variable calculation is done M times at every creation time [9], [10]. Thus, the entire CC of the proposed scheme is $\mathcal{O}(M^2 + N)$.

A model-free policy-based RL algorithm optimizes the parameters without knowledge of the environment. This algorithm measures the time of making model inference, that is, the step for the trained model to make decisions which minimizes the error on training samples, while keeping the bound on its model complexity small. Thus, computation time of the proposed scheme is faster than existing ML schemes which are analyzed in Fig. 4.

IV. Simulation Results and Analysis

A. Simulation Setup

The MATLAB based simulation results for the suggested scheme are presented in this section. Furthermore, the CUEs and M2M pairs are randomly placed in the network. In both CUEs and M2M pairs the channel fading follows a uniform variance of the Rayleigh distribution. Table I lists the main parameters of the simulation in detail. We compare the proposed algorithm with two traditional schemes, such as M2M mode and cellular mode. In the M2M mode, the M2M pair communicates with each other without using learning-based resource utilization [5]. Furthermore, in the cellular mode, the CUE communicates without using learning-based resource utilization in the proposed network [5]. Moreover, the proposed mechanism selects the suitable communication mode using RL empowered PN-based resource utilization policy without considering the UE selection.

B. The Throughput Comparison of M2M Pairs

Fig. 2 shows the average throughput versus the number of M2M pairs. The average throughput achieved by the proposed scheme rises as the number of M2M pairs rises. But, the rise is not remarkable due to the interference limitation in the cellular mode. From the figure, it can be observed that the proposed approach achieves a significant improvement in the network performance while mitigating the co-tier interference of the M2M Links. The M2M mode maximizes the SINR of M2M links compared to traditional methods, leading to only a slightly better performance than the cellular mode.

TABLE I
The Parameters of Simulation

Parameter	Value
Cell radius	250m
Carrier frequency	2 GHz
System bandwidth	6 MHz
Per RB bandwidth	200 kHz
Number of M2M pairs	30
Number of CUEs	10
M2M communication distance	50m
Maximum transmission power [2]	30 dBm
Noise power	-174 dBm/Hz
PL model between eNB and CUE [2]	$128.1 + 37.6\log(d[km])$
PL model between M2M pair [2]	$148 + 40\log(d[km])$
Threshold SINR	10 dB
Training samples	2000
Testing samples	500
Batch size	64
Hidden layer	64

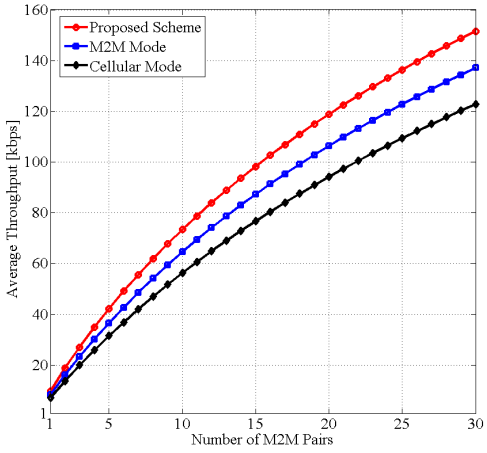

Fig. 2. Average throughput versus number of M2M pairs.

C. The Convergence of the Proposed Mechanism

Fig. 3 shows the reward comparison of the proposed scheme, M2M mode and cellular mode. As illustrated in the figure, our proposed scheme outperforms other existing M2M distribution algorithms in the proposed network. It shows that when the iteration increases, the network performance of users is improved, and our proposed method is much better than other methods. Compared to traditional resource distribution approaches, MUEs achieve reasonable system performance while co-tier interference is appropriately managed. Moreover, we can observe that the proposed scheme converges very rapidly. This is because the proposed scheme adopts the PN which reduces the CC. The numbers of M2M pairs and CUEs are set to be 30 and 10, respectively, which are randomly

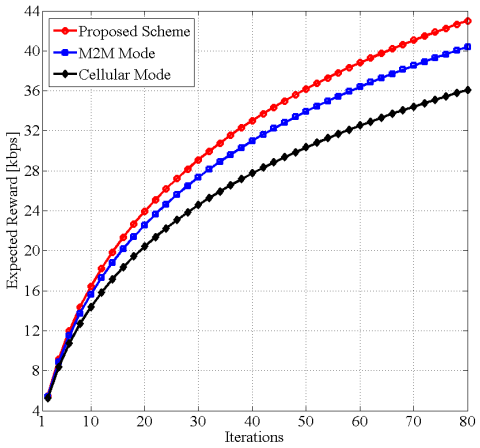

Fig. 3. The convergence of the proposed mechanism.

located in this figure. Additionally, if the number of UEs rises, the convergence rate becomes gradually slow. The estimated reward decreases as the number of UE rises, which indicates the system performance to be better if there are fewer UEs in the proposed network. The reason is that the interference generated by M2M links as a result of co-tier transmissions increases. Therefore, the estimated reward of a network with a lower number of UE is greater than at with the many UEs in M2M communications.

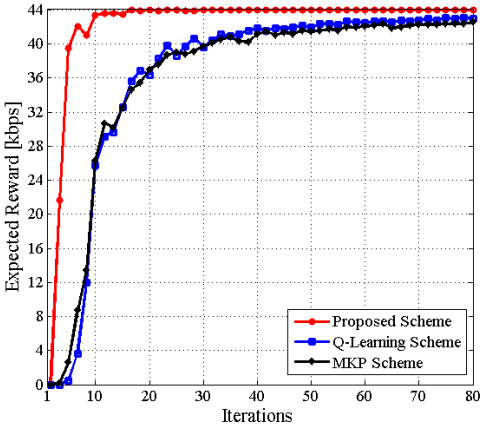

Fig. 4. The CC comparison with the proposed scheme, Q-learning scheme and MKP scheme.

D. The Computational Complexity of the Proposed Scheme

Fig. 4 shows the expected reward versus the number of iterations. Here, $\mathcal{O}(N(K^2+K+P))$ presents the CC of MKP where the amount of M2M pairs, length of MKP and the total iteration numbers are denoted by K, P, and N, respectively. This figure considers 80 and 30 as the values of N and K. The RL based Q-learning scheme for resource allocation, which is used in [8]. From Fig. 4, it can be seen that if the number of M2M pairs is small, then the Q-learning scheme has a close performance to the MKP scheme. As the number gets larger, the Q-learning scheme has achieved better reward performance and convergence than that of the MKP scheme as well as lower than that of the proposed scheme. In the case of large state-action space, Q-learning cannot provide good efficiency due to poor reward performance and convergence. On the other hand, by introducing PN architecture in the training process, larger reward performance, more stable convergence, lower CC, and adequate network performance are attained using the proposed scheme than that of MKP and Q-learning schemes.

V. Conclusions

This paper investigated the optimal resource distribution for M2M communications coexisting with cellular networks using the nonlinear programming mixed-integer optimization and low-complexity process which is based on the PN. Monte Carlo simulations have been performed to evaluate the performance of RL-based resource utilization policy. It has been found that the number of M2M pairs and iterations have substantial impact on the results. Simulation results showed that the proposed mechanism has achieved a better system performance in terms of average throughput compared to traditional approaches while significantly mitigating the interference. For further upgrading the spectrum allocation policy, federated learning framework can be used since it provides a global solution for complex network optimization problems without sharing information between eNBs that makes it an interesting topic for further investigation.

References

[1] F. Hussain, Internet of Things: Building Blocks and Business Models, 1st ed. Springer, Cham, 2017, pp. 73.

[2] S. K. Das and M. F. Hossain, "A Location-Aware Power Control Mechanism for Interference Mitigation in M2M Communications over Cellular Networks," *Comput. & Elect. Eng.*, vol. 88, pp. 1–23, Dec. 2020.

[3] X. Li, L. Ma, Y. Xu, and R. Shankaran, "Resource Allocation for D2D-Based V2X Communication With Imperfect CSI," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3545-–3558, Apr. 2020.

[4] H. Xu et al., "Robust transmission design for multicell D2D underlaid cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5922—5936, Jul. 2018.

[5] S. K. Das and R. Mudi, "A Location-Aware Resource Efficiency and Energy Efficiency Optimization for M2M Communications over Downlink Cellular Systems," *IEEE* 31^{st} *Annu. Int. Symp. on Pers., Indoor and Mobile Radio Commun.,* 2020, pp. 1–6.

[6] Y. Hao, Q. Ni, H. Li, and S. Hou, "Robust Multi-Objective Optimization for EE-SE Tradeoff in D2D Communications Underlaying Heterogeneous Networks," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4936—4949, Oct. 2018.

[7] W. Wu, R. Liu, Q. Yang, and T. Q. S. Quek, "Learning-Based Robust Resource allocation for D2D Underlaying Cellular Network," May 2021, Accessed: Jul. 03, 2021. [Online]. Available: *http://arxiv.org/abs/*2105.08324.

[8] D. Wang et al., "Joint resource allocation and power control for D2D communication with deep reinforcement learning in MCC," *Physical Commun.*, vol. 45, Apr. 2021.

[9] L. Zhu, C. Liu, J. Yuan, and G. Yu. "Machine Learning-Based Resource Optimization for D2D Communication Underlaying Networks," *IEEE* 92^{nd} *Veh. Technol. Conf. (VTC*2020*-Fall)*, 2020, pp. 1–6.

[10] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio, "Neural Combinatorial Optimization with Reinforcement Learning," 5^{th} *Int. Conf. Learn. Represent. ICLR* 2017 - *Work.* Track Proc., Nov. 2016, Accessed: Jul. 08, 2021. [Online]. Available: https://arxiv.org/abs/1611 : 09940v3.