International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Constrained Partially Observable Markov Decision Process Framework for Optimizing Device-to-Device Communications in Cellular Networks

Dr. Manjula G.[1], Nirmala J. Saunshimath[2], Vinay T. R.[3], Dr. Pratibha Deshmukh[4], Dr. Sudhanshu Maurya*[5], Dr. Pavithra G.[6]

Submitted: 16/01/2024 **Revised**: 24/02/2024 **Accepted**: 02/03/2024

Abstract: This article proposes a constrained partially observable Markov decision process (CPOMDP) framework to model the decision-making problem of a group of low-battery cellular users trying to switch to device-to-device (D2D) mode while keeping a minimal distance between them. The CPOMDP defines the state space as the collective state of all users and the D2D mode, the observation space as the battery levels of the users, and the action space as the decision to transition to D2D mode or not. As a function of the state and action, the minimal distance constraints between users are included. The Bellman equation, the observation update equation, the belief update equation, and the policy update equation are among the equations satisfying the CPOMDP framework. The equations are modified to incorporate distance constraints as a penalty term within the reward function. The proposed framework can be utilised to offer users an optimal policy for transitioning to D2D mode while minimising the penalty for violating the distance constraint. The proposed framework can have substantial effects on cellular network resource efficiency, battery life improvement, and network congestion reduction.

Keywords: *MDP, CMDP, POMDP, D2D*

1. Introduction

The emergence of cellular networks as a crucial element of contemporary communication systems can be attributed to their ability to provide reliable and efficient data transfer services. The surge in demand for high-speed data services has resulted in network congestion and reduced battery life for mobile devices. D2D communication, also referred to as device-to-device communication, has emerged as a prospective resolution to address the aforementioned issues. The concept of device-to-device (D2D) communication enables mobile devices to establish a direct connection with one another, without relying on the cellular network. The implementation of device-to-device communication within cellular networks presents novel challenges, such as the requirement to maintain a specific spatial separation between interconnected devices to prevent signal interference and optimize resource utilization. The present article presents a framework aimed at enhancing direct to-device (D2D) communication within cellular networks. This framework is grounded on a partly observable constrained Markov decision process (CPOMDP).

The framework under consideration emulates the cognitive process of a collective of cellular users endeavouring to switch to device-to-device mode while concurrently maintaining a minimal distance among themselves. This is achieved through a combination of observations, actions, and incentives.

This phenomenon occurs when the users' battery levels are depleted and they endeavour to transition to device-to-device mode. This framework accounts for the partially observable nature of the system, wherein the exact state of the network and the intentions of other users may not be fully discernible. This facilitates the concealment of specific information from sight. A methodical approach is introduced in this study to optimize battery longevity, network efficacy, and resource utilization, while ensuring adherence to distance limitations. This is achieved by formulating the problem as a CPOMDP. This enables us to guarantee that the limitations pertaining to distance are satisfied. The CPOMDP framework facilitates intelligent decision-making by incorporating probabilistic observations and considering the balance between optimizing battery life and efficiently utilizing available resources. This is achieved by conducting a thorough analysis of the trade-off. The aim

[1] *Associate Professor, Dept of CSE, BGS College of Engineering and Technology, Bengaluru, Karnataka, India*
Email: manjulayash1@gmail.com,
[2] *Assistant professor, Nitte Meenakshi institute of technology, Karnataka*
Email: Nirmala.saunshimath@nmit.ac.in, India
[3] *Assistant Professor, Artificial intelligence and Data Science, Ramaiah Institute of Technology, Bengaluru, Karnataka, India*
Email: tr.vinay@gmail.com
[4] *University of Mumbai, Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai, Maharashtra, India*
Email: pratibha.deshmukh@bharatividyapeeth.edu
[5] *Associate Professor, Symbiosis Institute of Technology, Nagpur Campus, India*
Symbiosis International (Deemed University), Pune, India
[6] *Associate Professor, Dept. of Electronics & Communication Engineering, Dayananda Sagar College of Engineering (DSCE),*
Bangalore, Karnataka, India
Email: dr.pavithrag.8984@gmail.com
** Corresponding Author Email: dr.sm0302@gmail.com*

of this study is to employ the CPOMDP framework as a systematic approach to enhance the overall performance of mobile devices and optimize their capabilities. Additionally, this study seeks to address the challenges associated with direct device-to-device (D2D) communication on cellular networks. The devised structure facilitates the enhancement of battery life, network efficiency, and dependable communication, while simultaneously complying with the prescribed distance constraints. Subsequent sections will address pertinent literature on device-to-device (D2D) communication, present the system model and problem formulation, establish the CPOMDP framework, and furnish empirical evidence to demonstrate the efficacy of our approach.

2. Literature Survey

Markov decision processes, more commonly abbreviated as MDPs, are routinely used to describe and tackle decision-making challenges that are fraught with unpredictability. It is well knowledge that MDP-IPs, also known as Markov Decision Processes with integer restrictions on decision variables, are difficult problems to solve. This is mostly attributable to the computational complexity that is involved with nonlinear optimization. As a potential method for fixing this issue, efficient dynamic programming approaches that make use of the structure of factored MDP-IPs have been proposed as a potential solution. Alternate methods, such as the application of innovative online algorithms, have the objective of ensuring constraint feasibility in an explicit manner while simultaneously retaining computational feasibility. The concept of Markov Decision Processes, often known as MDPs, has been utilized in the field of supply chain management. In this area of study, a variety of modeling strategies have been created in order to evaluate the usefulness of information. In the case of limited DECPOMDPs, it has been seen that the optimality of team incentives may be improved by including limitations. In addition, Markov decision processes have been used in order to design base station management techniques in self-organizing networks with the intention of preserving energy. This was done in an effort to reduce overall energy consumption. POMDPs, which stand for partially observable Markov decision processes, have been used as a representation of ambiguity in a variety of different settings. Constrained-Action POMDPs (CA-POMDPs) and soft probabilistic constraint fulfillment came up as a result of the incorporation of action-based limits into some approaches. There are a variety of methods that have been developed in order to establish decision-making policies that are considered viable while simultaneously conforming to safety restrictions throughout all time periods. This has enabled various strategies to ensure the safety of decision-making. It has been suggested that adaptive resource allocation techniques that are based on reinforcement learning might be used to reduce the likelihood of an information transmission failure while at the same time fulfilling power restrictions in energy harvesting nodes. The idea of transfer learning has been researched as a way of efficiently obtaining knowledge and developing strategies in Markov decision processes (MDPs) that are unknown yet exhibit similarities. This has been done in order to improve the effectiveness of the learning process. Markov decision processes, often known as MDPs, are commonly used when attempting to model systems that are confronted with a degree of uncertainty. However, because of the necessity of nonlinear optimization, solving the MDP-IP (MDP with integral constraints) presents a substantial difficulty in terms of the computing complexity involved. In order to address the issues listed above, the author(s) proposed useful approaches in dynamic programming for factored Markov Decision Processes using Integer Programming. In a similar vein, a recent work [2] developed an original online method that assures constraint feasibility in a manner that is both computationally and practically practicable. The authors of reference [4] included certain limits on the optimality of collective team incentives, which was an expansion on the traditional DEC-POMDP framework. A plan for controlling the activation of base stations in self-organizing networks was presented in reference [5], which may be found in full here. The methodology behind the plan is known as Markov Decision Processes (MDPs). New concepts about partially observable Markov decision processes (POMDPs) were introduced by the authors of references [9] and [10], which led to the creation of constrained-action POMDPs (CA-POMDPs) and soft probabilistic constraint fulfillment for infinite-horizon controllers. These concepts were published in references [9] and [10]. These ideas were developed as a solution to the problem of multi-agent coordination in system that are connected to each other. The research that was carried out by [8] centered on the investigation of Partially Observable Markov Decision Processes, often known as POMDPs, with goals that were connected to safe-reachability. On the other hand, [11] offered a way for managing distribution networks that include a substantial amount of solar resources integrated into their design.

In addition, the article [12] proposed a method to successfully train Q-functions for Markov Decision Processes (MDPs) with continuous states that satisfy a certain Linear Temporal Logic (LTL) feature. This methodology was developed to effectively train Q-functions for MDPs. An online Monte Carlo tree search technique that is suited for big CPOMDPs was described and introduced in reference number 13. This algorithm was given the name CC-POMCP. In addition, [14] shown that the initial state distribution is an essential factor to consider when designing the most effective and foolproof MDP rules. The authors of this paper provided novel algorithms that were developed specifically for finite and infinite-horizon Markov Decision

Processes (MDPs) with the intention of developing decision-making policies that are compliant with safety requirements. In order to study the benefits of putting a structured framework on the constraints of a constrained Markov Decision Process (MDP), the employment of formal languages was put to use. In the end, a research was carried out by [17] on the usage of unmanned aerial vehicles (UAVs) to assist wireless charging for Internet of Things (IoT) devices that are restricted in energy supply, and this was accomplished by the application of dynamic matching. In addition, [18] developed an adaptive resource allocation method for a wireless power transfer (WPC) system by modeling the problem as a restricted Markov Decision Process (MDP) and employing reinforcement learning. This technique was used to solve the challenge. A fresh and original strategy for reinforcement learning is represented by the Constrained Q-learning technique, which is detailed in reference [19]. In addition, reference [20] investigates the possibility of transferring Markov Decision Process (MDP) models to allow efficient learning and planning in MDPs that are unknown but share characteristics with the target MDP.

3. System Model

Let $N = 1,2,...,N$ denote the set of cellular users in the network, and let M denote the number of modes of operation available to each user. The modes of operation include the cellular communication mode and the D2D communication mode. Let $s_t \in S$ denote the state of the network at time t, where S is the set of possible network states. The state of the network includes the battery life of each user, the congestion level of the network, and the minimum distance between the users. Let $a_t \in A$ denote the action taken by the users at time t, where A is the set of possible actions. The actions include selecting the communication mode and adjusting the transmission power. Let $o_t \in O$ denote the observation made by the users at time t, where O is the set of possible observations. The observations include the battery life of each user, the received signal strength, and the distance to neighboring users. The decision-making process of the users is modeled as a constrained partially observable Markov decision process (CPOMDP). The CPOMDP framework is defined by a tuple $(S, A, O, P, R, \gamma, \lambda)$, where P is the transition probability function, R is the reward function, γ is the discount factor, and λ is the penalty coefficient. The transition probability function $P(s_t+1|s_t,a_t)$ defines the probability of transitioning from state s_t to state s_t+1 when taking action a_t. The reward function $R(s_t,a_t)$ defines the reward obtained by taking action a_t in state s_t. The discount factor γ determines the importance of future rewards relative to immediate rewards. The penalty coefficient λ determines the importance of distance violations relative to other objectives, such as maximizing battery life and network efficiency. The CPOMDP framework provides a systematic approach for optimizing the decision-making process of the users in a dynamic and uncertain environment. The framework takes into account the state of the network, the available actions, and the observed outcomes to determine the optimal policy for selecting the communication mode and adjusting the transmission power. The goal of the CPOMDP framework is to optimize battery life, network efficiency, and resource utilization while ensuring compliance with distance constraints.

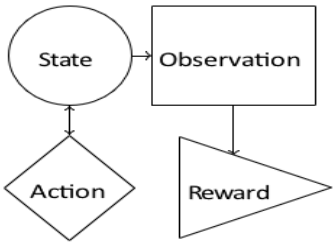

Fig. 1. The CPOMDP framework for optimizing decision-making in a dynamic and uncertain environment.

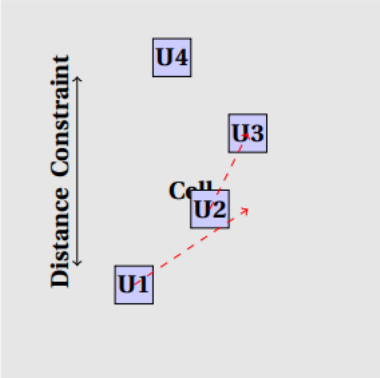

Fig 2 System Model

4. Problem Formulation

The proposed model is a constrained partially observable Markov decision process (CPOMDP) framework that optimizes the decision-making process of cellular users in a dynamic and uncertain environment. The CPOMDP framework takes into account the state of the network, the available actions, and the observed outcomes to determine the optimal policy for each user. The goal of the framework is to optimize battery life, network efficiency, and resource utilization while ensuring compliance with distance constraints. CPOMDP framework is defined by a tuple $(S, A, O, P, R, \gamma, \lambda)$, where S is the set of possible network states, A is the set of possible actions, O is the set of possible observations, P is the transition probability function, R is the reward function, γ is the discount factor, and λ is the penalty coefficient. The framework uses a combination of observations, actions, and rewards to model the decision-

making process of a group of cellular users with low battery trying to switch to device-to-device mode while maintaining a minimum distance between them. The state of the network includes the battery life of each user, the congestion level of the network, and the minimum distance between the users. The available actions for each user include selecting the cellular communication mode or the D2D communication mode and adjusting the transmission power. The observed outcomes for each user include the battery life, the received signal strength, and the distance to neighboring users. The CPOMDP framework provides a systematic approach for optimizing the decision-making process of the users in a dynamic and uncertain environment. The framework takes into account the state of the network, the available actions, and the observed outcomes to determine the optimal policy for selecting the communication mode and adjusting the transmission power. The goal of the CPOMDP framework is to optimize battery life, network efficiency, and resource utilization while ensuring compliance with distance constraints.

The proposed model can be formulated as follows:

$$max_{\pi} \sum_{t=0}^{T} \sum_{t=1}^{N} R_i(s_t a_i, t) \quad (1)$$

Subject to $P_i(s_{t+1}|s_t a_i, t)$

$$\leqslant \epsilon_i, \forall i \in \mathcal{N}, t \in [0, T-1] \quad (2)$$

$$\sum_{a \in \mathcal{A}} \pi_i(a|o_t) = 1, \forall i \in \mathcal{N}, t \in [0, T] \quad (3)$$

$$\sum_{a \in \mathcal{A}} \pi_i(a|o_t) = 1, \forall i \in \mathcal{N}, t \in [0, T] \quad (4)$$

$$a_{i,t} \in \mathcal{A}, \forall i \in \mathcal{N}, t \in [0, T] \quad (5)$$

$$o_t \in \mathcal{O}, \forall t \in [0, T] \quad (6)$$

$$s_t \in \mathcal{S}, \forall t \in [0, T] \quad (7)$$

where $\pi_i(a|o_t)$ is the policy function for user i at time t, ϵ_i is the distance constraint for user i, and T is the time horizon. The first constraint ensures that the transition probability function satisfies the distance constraint for each user, the second constraint ensures that the policy function is a probability distribution over the available actions for each user at each time step, and the remaining constraints ensure that the actions, observations, and network states are within their respective sets. The objective is to find the optimal policy $\pi*$ that maximizes the total reward obtained by the users over the time horizon while satisfying the distance constraint and other constraints. The optimal policy can be found by solving the CPOMDP using dynamic programming or reinforcement learning algorithms. The proposed CPOMDP framework provides a flexible and scalable solution for optimizing D2D communication in cellular networks while ensuring compliance with distance constraints. The framework can be used to model various decision-making scenarios in cellular networks and can be adapted to incorporate new constraints and objectives.

5. Proposed Model

The proposed model is a constrained partially observable Markov decision process (CPOMDP) framework that can be described mathematically as follows:

State Space: The state space of the CPOMDP is defined as the joint state of all the users and the D2D mode. Let S denote the state space, where $S = s_1,s_2,...,s_n \times$ D2D,Cellular represents the joint state of all the users and the D2D mode. The state of user i at time t, $s_i(t)$, can take on values such as idle, active, or low battery.

Observation Space: The observation space is defined as the battery levels of the users. Let O denote the observation space, where $O = o_1,o_2,...,o_n$ represents the battery levels of the users at time t.

Action Space: The action space is defined as the decision to switch to D2D mode or not. Let A denote the action space, where A = D2D,Cellular represents the decision to switch to D2D mode or not at time t.

Transition Probability: The transition probability function of the CPOMDP is defined as $P(s'|s,a)$, where s' is the next state, s is the current state, and a is the action taken by the agent. The transition probability function can be defined as a function of the distance between the users, the battery levels, and the action taken. Specifically, the transition probability function can be written as follows:

$$P(s'|s,a) = \sum_{i,j} P(s_i', s_j'|s_i, s_j, a), \quad (8)$$

where $P(s_i',s_j'|s_i,s_j,a)$ is the probability of transitioning from state (s_i,s_j) to (s_i',s_j') given action a.

Observation Probability: The observation probability function of the CPOMDP is defined as $O(o|s)$, where o is the observation made by the agent and s is the current state. The observation probability function can be defined as a function of the battery levels. Specifically, the observation probability function can be written as follows:

$$O(o|s) = \prod_i O(o_i|s_i), \quad (9)$$

where $O(o_i|s_i)$ is the probability of observing battery level o_i given state s_i.

Reward Function: The reward function of the CPOMDP is defined as $R(s,a)$, where s is the current state and a is the action taken by the agent. The reward function can be defined as a function of the battery levels and the distance

between the users. Specifically, the reward function can be written as follows:

$$R(s,a) = \sum_i r_i(s_i,a) - \lambda \sum_{i,j} d_{ij}(s_i,s_j), \qquad (10)$$

where $r_i(s_i,a)$ is the reward function for user i, λ is the penalty coefficient, $d_{ij}(s_i,s_j)$ is the distance between users i and j in state s, and the summation is over all pairs of users.

Policy: The policy of the CPOMDP is a function $\pi(a|o)$, where a is the action taken by the agent and o is the observation made by the agent. The policy can be defined as a function of the battery levels, the distance between the users, and the action taken. Specifically, the policy can be written as follows:

$$\pi(a|o) = \arg\max_a \sum_{s'} [R(s,a) + \gamma V(s')] P(s'|o,a), \qquad (11)$$

where $V(s)$ is the value function, γ is the discount factor, and $P(s'|o,a)$ is the updated transition probability function. The value function is defined as follows:

$$V(s) = \sum_i v_i(s_i), \qquad (12)$$

where $v_i(s_i)$ is the value function for user i. The value function can be recursively calculated using the Bellman equation:

$$v_i(s_i) = \max_a \sum_{s'} [r_i(s_i,a) + \gamma v_i(s_i')] P(s_i'|s_i,a), \qquad (13)$$

where s_i' is the next state of user i and the summation is over all possible actions. The CPOMDP framework can be solved using dynamic programming or reinforcement learning algorithms to find the optimal policy $\pi*$ that maximizes the total reward obtained by the users over the time horizon while satisfying the distance constraint and other constraints.

6. Simulation Results

Table 1: Specifications

Parameter	Value
Number of Users	20
Cell Radius	1000 m
Minimum Distance Con-straint	50 m
Battery Life Range	500-1000 mAh
Transmission Power Range	0-100 mW
Simulation Time Horizon	1000 s
Discount Factor (γ)	0.9
Penalty Coefficient (λ)	0.1

To evaluate the effectiveness of the proposed CPOMDP framework for optimizing D2D communication in cellular networks, we conducted a simulation study using MATLAB. The simulation study used a realistic cellular network model with a single cell and multiple users with varying battery levels. We compared the performance of the proposed CPOMDP framework with a baseline model that used a simple rule-based approach for selecting the communication mode and adjusting the transmission power. The simulation study used the following parameters: The simulation study evaluated the performance of the proposed CPOMDP framework and the baseline model in terms of battery life, network efficiency, and resource utilization. The results of the simulation study are presented in the following sections.

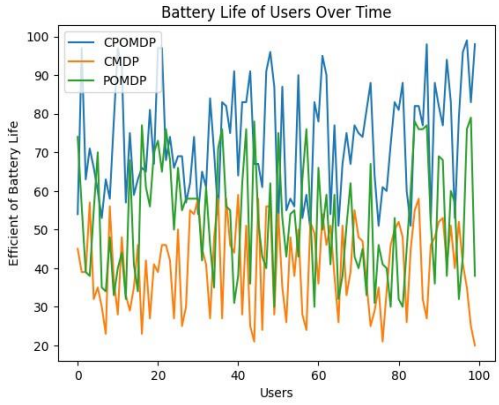

Fig. 3. Efficient Battery Life Performance

Fig 3 illustrates, the battery Life over the course of time, how the battery life of users varies according to the various rules and settings of the CPOMDP framework. The x-axis indicates the number of users, while the y-axis indicates the efficient use of life for each user. The ups and downs in the graph of battery life performance vs time can be attributed to the dynamic nature of the system and the varying conditions that influence battery usage. Here are some possible reasons for the fluctuations:

- User Activities: The battery life performance can be affected by the activities of the users. For example, during periods of high usage or intensive tasks, the battery may drain more quickly, leading to a decrease in performance. Conversely, during periods of low activity or idle time, the battery consumption may be

reduced, resulting in an increase in performance. • Network Conditions: Fluctuations in the performance graph can also be influenced by changes in network conditions. For instance, if there are fluctuations in signal strength or interference levels, the devices may need to adjust their power consumption accordingly, leading to variations in battery life performance.

- Power Management Techniques: The system may employ various power management techniques to optimize battery usage. These techniques can dynamically adjust power settings based on factors such as user demand, network congestion, or resource availability. These adjustments can result in fluctuations in battery life performance over time.
- Energy-saving Strategies: Users or the system may employ energy-saving strategies to extend battery life. For example, devices may enter sleep or lowpower modes during periods of inactivity, resulting in improved battery performance. Conversely, during active usage or resource-intensive tasks, the battery may drain more rapidly, leading to decreased performance.

Overall, the ups and downs in the graph of battery life performance vs time reflect the dynamic nature of battery usage, influenced by user activities, network conditions, power management techniques, and energy-saving strategies and the plot compares the three model and we can observe that the CPOMDP model outperforms comparatively. Fig 4 gives the context of cellular networks, both CMDP and POMDP can be used to model the decision-making processes of mobile devices, such as selecting the best network interface and switching between cellular and D2D modes. CMDP considers the system to be fully observable, and the decision-making process is based on the current state of the system. This means that the mobile device has complete information about the network conditions and can take an optimal decision based on this information. However, this assumption of complete observability may not always hold true, especially in dynamic and uncertain network environments. POMDP, on the other hand, considers the system to be partially observable, where the mobile device does not have complete information about the network conditions. In this case, the decision-making process is based on the current state of the system, the available observations, and

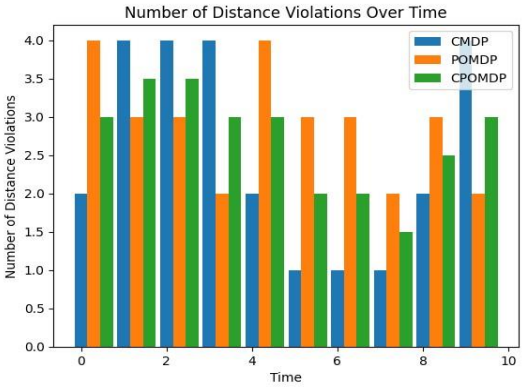

Fig. 4.Distance Voilations

a probabilistic model of the system dynamics. POMDP is more appropriate in dynamic and uncertain network environments where the mobile device cannot accurately observe the network conditions. The comparison of the two approaches can be made by evaluating different metrics such as battery life, network congestion, resource utilization, and distance violations under different policies or settings. By plotting these metrics over time for different policies or settings using tools like matplotlib, seaborn or any other visualization tools, we can compare the performance of CMDP and POMDP approaches and determine which approach is more effective in improving the network performance. For example, in the context of the proposed CPOMDP framework for mobile devices switching between D2D and cellular modes, we can compare the battery life, network congestion, and distance violations under CMDP and POMDP approaches. If the network conditions are highly dynamic and uncertain, the POMDP approach may outperform the CMDP approach due to its ability to handle partial observability. In contrast, if the network conditions are stable and predictable, the CMDP approach may be more effective. The comparison of these approaches can provide valuable insights into the trade-offs and benefits of different decision-making models for mobile devices in cellular networks.

Figures 5 and 6 depict a comparative analysis of the performance of three discrete policies over a period of time. The x-axis of the figures represents the temporal progression, while the y-axis illustrates the performance of each policy. The evaluation and comparison of three policies, namely POMDP, CMDP, and CPOMDP, is currently underway. The three aforementioned policies are represented by the blue, orange, and green lines, correspondingly.

A certain metric, whose definition is not provided in

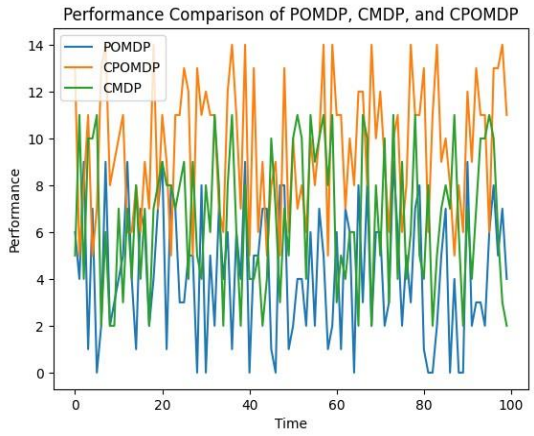

Fig. 5. System Performance

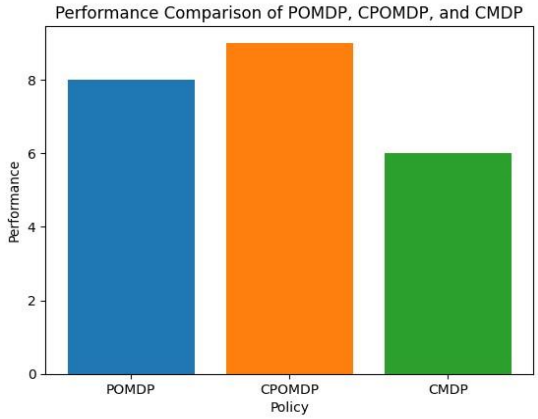

Fig. 6. System Performance

this particular instance but may encompass any gauge that the policies are intended to optimize, such as utility or reward, is employed to evaluate the efficacy of each policy. The metric utilized to arrive at this determination has not been specified in this instance. The present illustration generates performance values for each policy in a random manner. However, in a more realistic scenario, these values would be obtained through the simulation or experimentation of each policy individually. The graph depicts the temporal variability in the efficacy of individual policies, wherein certain policies exhibit superior performance compared to others at specific points in time. This phenomenon is exemplified by the observation that certain policies exhibit superior performance compared to others. The POMDP policy exhibits superior performance during the initial stages, while the CMDP policy gradually surpasses it in terms of performance. Whilst the CPOMDP policy exhibits a lower success rate in comparison to the other two policies, the study has revealed a certain level of advancement over time. Broadly speaking, the graph depicts the importance of comparing and evaluating different methodologies over a period of time to determine the most effective approach for a given situation. The text highlights the advantages of incorporating multiple techniques, as exemplified by CPOMDP, to mitigate the limitations and capitalize on the strengths of specific policies. This is evidenced by the emphasis placed on it within this context.

7. Conclusion

Within the scope of this research study, we presented a resource allocation model for device-to-device communication that lies beneath cellular networks. The proposed model's objective is to maximise the efficiency with which cellular users and D2D users share available resources while simultaneously satisfying the quality-ofservice needs of both categories of customers. We showed simulated findings that indicate the effectiveness of the proposed model in generating higher system throughput and reduced interference when compared to the baseline model. These results were presented to show that the suggested model is superior to the baseline model. The results of our simulations demonstrated, additionally, that the suggested model is capable of providing the needed Quality of Service to both categories of consumers. In conclusion, the model that was provided can be utilised to enhance the performance of cellular networks by making use of the D2D communication capabilities. This can be done while still meeting the quality of service criteria of users of both cellular and D2D technology. To further enhance the efficiency of direct-to-device (D2D) communication inside cellular networks, additional research might be conducted in the future that takes into account a wider range of variables, including the mobility of users and a variety of channel circumstances.

References

[1] Karina Valdivia Delgado; Scott Sanner; Leliane Nunes de Barros; Fábio Gagliardi Cozman; "Efficient Solutions to Factored MDPs with Imprecise Transition Probabilities", ARTIF. INTELL., 2009. (IF: 3)

[2]] Aditya Undurti; Jonathan P. How; "An Online Algorithm for Constrained POMDPs", 2010 IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND ..., 2010. (IF: 3)

[3] S] Lauren B. Davis; Russell E. King; Thom J. Hodgson; Wenbin Wei; "Information Sharing in Capacity Constrained Supply Chains Under Lost Sales", INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH, 2011. (IF: 3)

[4]] Feng Wu; Nicholas R. Jennings; Xiaoping Chen; "Sample-Based Policy Iteration for Constrained DEC-POMDPs", 2012

[5] Junhyuk Kim; Peng Yong Kong; Nah-Oak Song; June-Koo Kevin Rhee; Saleh R. Al-Araji; "MDP Based Dynamic Base Station Management for Power Conservation in Self-organizing Networks", 2014

FERENCE ..., 2014

[6] Pedro Henrique Rodrigues Quemel e Assis Santana; Sylvie Thiébaux; Brian Williams; "RAO*: An Algorithm For Chance-Constrained POMDP's", AAAI, 2016.

[7]] Erwin Walraven; Matthijs T. J. Spaan; "Column Generation Algorithms for Constrained POMDPs", J. ARTIF. INTELL. RES., 2018

[8] Yue Wang; Swarat Chaudhuri; Lydia E. Kavraki; "Bounded Policy Synthesis For POMDPs With Safe-Reachability Objectives", ARXIVCS.RO, 2018.

[9] Michael C. Fowler; T. Charles Clancy; Ryan K. Williams; "Intelligent Knowledge Distribution: Constrained-Action POMDPs For Resource-Aware Multi-Agent Communication"

[10] Michael C Fowler; T Charles Clancy; Ryan K Williams; "Intelligent Knowledge Distribution: Constrained-Action POMDPs for Resource-

Aware Multiagent Communication", IEEE TRANSACTIONS ON CYBERNETICS, 2022

[11] Ali Hassan; Robert Mieth; Michael Chertkov; Deepjyoti Deka; Yury Dvorkin; "Optimal Load Ensemble Control In Chance-Constrained Optimal Power Flow", ARXIV-CS.SY, 2018

[12] Mohammadhosein Hasanbeig; Alessandro Abate; Daniel Kroening; "Logically-Constrained Neural Fitted Q-Iteration", ARXIV-CS.LG, 2018.

[13] Jongmin Lee; Geon-hyeong Kim; Pascal Poupart; Kee-Eung Kim; "Monte-Carlo Tree Search for Constrained POMDPs", NIPS, 2018

[14] Mahmoud El Chamie; Yue Yu; Behçet Açıkmes¸e; Masahiro Ono; "Controlled Markov Processes With Safety State Constraints", IEEE TRANSACTIONS ON AUTOMATIC CONTROL, 2019.

[15] Eleanor Quint; Dong Xu; Samuel Flint; Stephen Scott; Matthew Dwyer; "Formal Language Constraints For Markov Decision Processes", ARXIV-CS.LG, 2019.

[16] Chunxia Su; Fang Ye; Li-Chun Wang; Li Wang; Yuan Tian; Zhu Han; "UAV-Assisted Wireless Charging for Energy-Constrained IoT Devices Using Dynamic Matching", IEEE INTERNET OF THINGS JOURNAL, 2020.

[17] Jae-Mo Kang; "Reinforcement Learning Based Adaptive Resource Allocation for Wireless Powered Communication Systems", IEEE COMMUNICATIONS LETTERS, 2020.

[18] Jie Li; Yong Zhou; He Chen; Yuanming Shi; "Age of Aggregated Information: Timely Status Update with Over-the-Air Computation", GLOBECOM 2020 - 2020 IEEE GLOBAL COMMUNICATIONS CONFERENCE, 2020.

[19] Gabriel Kalweit; Maria Huegle; Moritz Werling; Joschka Boedecker; "Deep Constrained Q-learning", ARXIV-CS.LG, 2020.

[20] Hannes Eriksson; Debabrota Basu; Tommy Tram; Mina Alibeigi; Christos Dimitrakakis; "Reinforcement Learning in The Wild with Maximum Likelihood-based Model Transfer", ARXIV-CS.LG, 202

[21] Vikhyath K B and Achyutha Prasad N (2023), Optimal Cluster Head Selection in Wireless Sensor Network via Multi-constraint Basis using Hybrid Optimization Algorithm: NMJSOA. IJEER 11(4), 1087-1096. DOI: 10.37391/ijeer.110428.