Joint Optimization Framework For Maximization of Instantaneous Transmission Rate In Signal To Interference Noise Ratio Constrained UAVs-Supported Self-Organized Device-To-Device Network

Abhishek Mondal (abhishekkmondal532@gmail.com)
National Institute of Technology Silchar
https://orcid.org/0000-0001-6849-6207

Ashraf Hossain
National Institute of Technology Silchar

Research Article

Keywords: Unmanned aerial vehicle (UAV), device-to-device (D2D) communication, Markov decision process (MDP), reinforcement learning (RL), state-action-reward-state-action (SARSA), signal to interference noise ratio (SINR)

DOI: https://doi.org/10.21203/rs.3.rs-404001/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Joint Optimization Framework for Maximization of Instantaneous Transmission Rate in Signal to Interference Noise Ratio Constrained UAVs-Supported Self-Organized Device-to-Device Network

Abhishek Mondal and Ashraf Hossain
Department of Electronics and Communication Engineering
National Institute of Technology Silchar, India
*Corresponding Author E-mail: abhishekmondal532@gmail.com

Abstract
Due to their high maneuverability, flexible deployment, and line of sight (LoS) transmission, unmanned aerial vehicles (UAVs) could be an alternative option for reliable device-to-device (D2D) communication when a direct link is not available between source and destination devices due to obstacles in the signal propagation path. Therefore, in this paper, we have proposed a UAVs-supported self-organized device-to-device (USSD2D) network where multiple UAVs are employed as aerial relays. We have developed a novel optimization framework that maximizes the total instantaneous transmission rate of the network by jointly optimizing the deployed location of UAVs, device association, and UAVs’ channel selection while ensuring that every device should achieve a given signal to interference noise ratio (SINR) constraint. As this joint optimization problem is nonconvex and combinatorial, we adopt reinforcement learning (RL) based solution methodology that effectively decouples it into three individual optimization problems. The formulated problem is transformed into a Markov decision process (MDP) where UAVs learn the system parameters according to the current state and corresponding action aiming to maximize the generated reward under the current policy. Finally, we conceive SARSA, a low complexity iterative algorithm for updating the current policy in the case of randomly deployed device pairs which achieves a good computational complexity-optimality tradeoff. Numerical results validate the analysis and provide various insights on the optimal deployment of UAVs. The proposed methodology improves the total instantaneous transmission rate of the network by 75.37%, 52.08%, and 14.77% respectively as compared with RS-FORD, ES-FIRD, and AOIV schemes.

Keywords- Unmanned aerial vehicle (UAV), device-to-device (D2D) communication, Markov decision process (MDP), reinforcement learning (RL), state-action-reward-state-action (SARSA), signal to interference noise ratio (SINR).

1 Introduction
The fifth-generation mobile communication (5G) has obliged to enable new features including critical and disaster recovery operations in the modern communication systems. These operations offer consistent network performance by maintaining the required quality of service (QoS) which incorporated latency, transmission rate, and reliability throughout urban and rural environments [1]. Although, maintaining the required QoS is challenging because radio coverage varies due to obstacles in the signal propagation path and high user density significantly reduces network capabilities. These limitations can be combated by deploying unmanned aerial vehicles (UAVs) having the ability to establish line of sight (LoS) dominant air to ground channel in a controllable manner [2]. Specifically, the UAV-enabled wireless sensor network (WSN) utilizes UAVs as computing hub and mobile data collectors to perform several analyses on the data, collected from spatially separated sensor nodes (SNs) and communicating devices [3].

The use of UAVs in existing wireless communication systems makes them more efficient in terms of coverage, capacity, and energy consumption. Hence, the flying cell site technology employs UAVs as aerial base stations to serve a region that is beyond the coverage of the cellular network [4]. The European Commission has launched Project Loon [5] where UAVs are deployed as flying base stations (BSs) to provide broadband service for rural and remote areas. The ABSOLUTE project, based on Long Term Evolution-Advanced (LTE-A) aerial base station has been inaugurated to support the services for disaster relief activities by providing huge coverage and connectivity [6]. Moreover, due to their agility and mobility, the deployment of UAV relays can be beneficial in a zone, where constructing cellular and mobile infrastructure is very expensive [7]. Although, when UAVs are operated as an aerial relay, not only the link between ground source device to UAV but also the link between UAV to destination device influence the network performance. Therefore, the deployment of UAVs as relays for device-to-device communication (D2D) is a more practical concern.
The existing literature focuses on the deployment of UAV relays for D2D communication under a broad range of aspects [8]-[16]. In [8], the authors have estimated the optimal UAV relay position in a multi-rate communication system using theoretical and simulated analysis. The work in [9] investigates optimal deployment and movement of UAV relay to improve the connectivity of ground users. The authors of [10] and [11] maximize the lower bound of uplink transmission rate and enhance the quality of the link between UAV relay and ground devices using dynamic heading adjusting approaches. For throughput maximization of mobile relaying system, an iterative algorithm is developed [12],[13] which jointly optimizes trajectory of the relays, transmit power of sources and UAVs by satisfying the practical constraints. Chen et al. [14] suggested two relaying schemes under different channel models for deployment of multiple UAV, but no resource competition is considered among source-destination pairs in the network. Underlay spectrum sharing [15] between three-dimensional drone cells and the traditional cellular networks has been modeled using a two-dimensional Poisson point process where the coverage probability and achievable throughput of drone cells are maximized by satisfying cellular network efficiency constraint. However online spectrum allocation corresponding to various UAVs’ locations has not been considered. In [16], the authors maximize the throughput of the UAV relay network by jointly optimizing transmit power, bandwidth, transmission rate, and relay deployment. Although, in this work model based centralized approach considering only a single relay is used where all necessary system parameters are required. Additionally, the research gap still exist on the enhancement of network performance for source-destination device pair communication by optimal placement of UAV relays. In such existing work, new problems will occur as follows.

- Due to the limited number of UAV relays in the D2D network, relay competition occurs among the source-destination device pairs. Moreover, a device receives different QoS from different positions of multiple UAVs. Therefore, device association with UAV relay is a significant factor that influences network capacity, which is overlooked by current articles.

- In a practical scenario, the spectrum resources that can be used for UAV communication networks are limited. Whenever multiple UAV relays and multiple device pairs are assigned the same channel, mutual interference occurs among the communicating nodes. Especially, the effect of interferences increases when the number of communicating devices and UAVs increases in the network. Therefore, channel assignment needs to be considered for a UAV communication system.

- The relationship among device association with UAV relay, UAV’s channel selection, and optimal deployment of UAV relay are intractable where one variable change may impact other variables. Furthermore, the existing optimization approaches related to these variables are all model-based. Some initial information, e.g., channel model and location of ground users are utilized to solve the optimization process. But in reality, the communication environment is unknown which indicates that the exact channel model and location of the ground unit are hard to know. Therefore, a model-free approach must be designed where UAVs can estimate channel models by field measurement and statistics.

During recent decades, machine learning (ML) algorithms have gained remarkable attention in the communication and networking sector. Moreover, reinforcement learning (RL) being a part of ML can estimate the channel model and learns the system behavior with an unknown environment in UAV assisted wireless network [17],[18]. Liu et al. [19] proposed a Q-learning based deployment and movement algorithm that maximizes the achieved sum mean opinion score of ground users. A multi-agent RL framework has been adopted by Cui et al. [20] for investigating dynamic resource allocation where the aim is to maximize the long-term expected reward by automatically selecting power level and subchannel allocation when there is no information exchange among UAVs. The authors of [21] presented deep reinforcement learning (DRL) algorithm to control the trajectory of UAVs and minimize the energy consumption of all user equipment via jointly optimizing user association, resource allocation, and trajectory of UAVs.

It is observed from the existing literature that UAV relays are used for D2D pairs communication when a direct link cannot be established between them. Furthermore, UAVs help low transmitting power devices for conveying data to the intended receiver located at a long distance where UAVs’ channel selection and device association with a UAV are executed in a decentralized manner. However, the deployment of UAVs using conventional methodology [7] faces difficulties to satisfy network coverage and capacity requirements. Particularly, the interactable relationship among UAVs’ deployment, relay assignment, and resource allocation cannot be optimized jointly using a centralized model-based approach in an unknown communication environment. Therefore, in this paper, we have proposed a UAVs-supported self-organized device-to-device (USSD2D) network containing multiple source-destination device pairs and multiple UAVs where the aim is to find the optimal deployed location of UAVs, device association with UAV, and UAVs’ channel selection that simultaneously supports reliable data transmission between source and destination device pairs. To the best of our knowledge, this is the first work that considers signal to interference noise ratio (SINR) constrained maximization of the total instantaneous transmission rate of the USSD2D network by jointly optimizing device association, UAV’s channel selection, and their deployed location at every time slot while considering the random
deployment of static source-destination device pairs in an unknown communication environment. Based on the proposed framework, the major contributions are summarized as follows:

- Considering a realistic scenario, we have formulated an SINR constraint joint optimization problem for maximizing the total instantaneous transmission rate of the USSD2D network over the entire flight period of UAVs. Because of its time slot dependent, nonconvex and combinatorial nature, the formulated optimization problem is non-trivial. To solve this challenging problem, a model-free RL-based solution methodology is proposed that decouples the joint optimization problem into three individual optimization problems. UAVs acting as RL agents learn the adaptive strategy from their own experiences earned by interaction with an unknown environment.

- We have decomposed the formulated non-convex and combinatorial optimization problem into a multi-period decision-making process, having three individual optimization problems such as individual device association with UAV, UAVs’ channel selection, and their deployed location at every time slot. Specifically, we model the trajectory optimization problem as a Markov decision process (MDP) where the states and actions are represented as locations of UAVs at the current time slot and corresponding movement of UAVs respectively. As each device selects a single UAV according to its association probability, the total transmission rate achieved by all devices at a time slot over the allocated channel is treated as an instantaneous reward attained by the UAVs.

- State-action-reward-state-action (SARSA) algorithm and $\epsilon$-greedy action selection scheme are adopted for obtaining the optimal policy that allows UAVs to find their deployed location at every time slot without the need for system identification. The proposed methodology obtains the current policy by using Q-value for all state-action pairs. Furthermore, to enhance the learning process, we have analyzed both the convergence and complexity of the proposed algorithm.

- We have provided numerical results to validate the analysis and gain insights on the impact of various system parameters for optimal UAVs deployment versus total instantaneous transmission rate trade-off. Finally, to corroborate the importance of the proposed joint optimization framework, we have presented its performance comparison against the benchmark scheme to quantify the achievable improvement of the total instantaneous transmission rate of the USSD2D network.

The rest of the article is organized as follows. In section 2, the proposed USSD2D network including objective problem formulation is elaborated. Afterward, we discuss the key techniques of the SARSA algorithm and optimal decision-making policy in section 3 and section 4 respectively. Simulation results and corresponding inferences are focused in section 5. Finally, the conclusion is outlined in section 6 followed by references.

### 2 System model

Consider a UAV-supported self-organized device-to-device (USSD2D) network as depicted in Fig. 1 is operating under a discrete-time axis. There is a set of randomly deployed source and destination device pairs on the ground with a fixed location in the target area. Some of the device pairs can communicate via direct links because of short distances and good channel conditions between them which are identified as direct D2D pairs. Although in case of long-distance communication or obstructed by obstacles in signal propagation path, the direct link cannot be established for all communicating device pairs. In order to create links for the source-destination device pairs when a direct link is not
available, multiple UAV equipped with single and half-duplex antenna is deployed at a fixed altitude $H_u$ to operate as a relay for their communication. We consider those device pairs as UAV-assisted D2D pairs.

### 2.1 Channel model of USSD2D network

The set of source and destination devices of direct D2D pairs is represented as $\mathcal{K}_S = \{1,2, ..., K\}$ and $\mathcal{K}_D = \{K+1, K+2, ..., 2K\}$ respectively, where $K$ is the number of direct D2D pairs. For any device $k \in \{\mathcal{K}_S \cup \mathcal{K}_D\}$, its location is denoted as $(x_k, y_k)$ and its respective active and silent probability is $P_k$ and $1-P_k$. Similarly, the set of source and destination devices of UAV-assisted D2D pairs is represented as $\mathcal{K}_S = \{1,2, ..., K\}$ and $\mathcal{K}_D = \{K+1, K+2, ..., 2K\}$ respectively, where $\breve{K}$ is the number of UAV-assisted D2D pairs. For a device $\breve{k} \in \{\mathcal{K}_S \cup \mathcal{K}_D\}$, its location is denoted as $(x_k, y_k)$. The set of $M$ number of UAV relays represented by $\breve{M} = \{1,2, ..., M\}$ is deployed to support the D2D pair’s communication in absence of their direct link. The flight duration of each UAV ($T_f$) is discretized into equally spaced time slots represented by $T_f = \delta T$, where $\delta$ is slot duration and $T$ is total available time slots. The location of $m^{th}$ UAV at time slot $t$ is denoted as $U_m(t) = (x_m(t), y_m(t), H_u)$, $\forall m \in \mathcal{M}$, $t \in \mathcal{T} = \{1,2, ..., T\}$. The set of $J$ orthogonal channels denoted by $\breve{J} = \{1,2, ..., J\}$ is available in the target area. Any communication node selects its channel from this available channel set.

According to practical scenario [22], one source device can associate only one UAV at a time slot but one UAV can be shared by multiple source devices. Furthermore, each UAV can occupy only a single orthogonal channel for each time slot and direct D2D pairs are assigned fixed channels during the entire flight period of UAVs. The transmission link between source device to associated UAV and UAV to destination device follows round-robin protocol [22] which is characterized by an amplify-and-forward relaying model. To indicate whether device $k \in \mathcal{K} = \{\mathcal{K}_S \cup \mathcal{K}_D \cup \mathcal{K}_S \cup \mathcal{K}_D\}$ associates with UAV $m \in \mathcal{M}$ at time slot $t$, an indicator function $I_{k,m}(t)$ is proposed as

$$I_{k,m}(t) = \begin{cases} 1, & \text{if user } k \text{ associates with UAV } m \\ 0, & \text{Otherwise} \end{cases}$$

(1)

The number of devices associated with UAV $m \in \mathcal{M}$ at time slot $t$ is calculated as

$$I_m(t) = \sum_{k \in \mathcal{K}} I_{k,m}(t)$$

(2)

Similarly, an indicator function $I_{m,j}(t)$ is used to represent whether UAV $m \in \mathcal{M}$ selects the channel $j \in \mathcal{J}$ at time slot $t$ as

$$I_{m,j}(t) = \begin{cases} 1, & \text{if UAV } m \text{ selects channel } j \\ 0, & \text{Otherwise} \end{cases}$$

(3)

The path loss between device $k \in \mathcal{K}$ and UAV $m \in \mathcal{M}$ at time slot $t$ can be expressed as [23]

$$P_{L_{k,m}}(t) = \frac{\mu_{LoS} - \mu_{NLoS}}{1 + b_1 \exp \left[ -b_2 \frac{180}{\pi} \phi_{k,m}(t) - b_1 \right]} + 20 \log \left( \frac{4\pi f_c D_{k,m}(t)}{c} \right) + \mu_{NLoS}$$

(4)

where $c$ is the velocity of light, $f_c$ is the carrier frequency of the channel selected by UAV, $\mu_{LoS}$ and $\mu_{NLoS}$ are attenuation factors corresponding to the LoS and NLoS path respectively, $b_1$ and $b_2$ are the environment-dependent constant.

$$\phi_{k,m}(t) = \sin^{-1} \left( \frac{H_u}{D_{k,m}(t)} \right)$$

is the elevation angle between device $k$ and UAV $m$ where the instantaneous distance between device $k$ and UAV $m$ is calculated as,

$$D_{k,m}(t) = \sqrt{(x_m(t) - x_k)^2 + (y_m(t) - y_k)^2 + H_u^2}$$

(5)

The instantaneous channel gain between source device $k \in \mathcal{K}$ and relay UAV $m \in \mathcal{M}$ can be calculated as

$$G_{k,m}(t) = 10^{-P_{L_{k,m}}(t)/10}$$

(6)
2.2 Transmission model

At time slot $t$, the received signal to interference plus noise ratio (SINR) of UAV $m$ over channel $j \in J$ from source device $k \in K$ can be expressed as [22]

$$\Gamma_{k,m}^j(t) = \frac{P_k^{Tx} I_{k,m}^j(t)}{N_0 + \rho_{k,m}(t)}$$

where $P_k^{Tx}$ is transmit power of ground source device $k$, $N_0$ is the power of additive white Gaussian noise (AWGN) and $\rho_{k,m}(t)$ is instantaneous interference at the receiver of UAV $m$ due to receiving a signal from source device $k$ over channel $j$. The instantaneous interference received by UAV $m$ can be categorized into two parts. The first part is the interferences from the sources of the relays that occupy the same channel as UAV $m$ which is denoted as $\tilde{\psi}_{k,m}^j(t)$ and the second part is interferences from the sources of direct D2D pairs that occupy the same channel as UAV $m$, represented by $\bar{\psi}_{k,m}^j(t)$. The total interference received by UAV $m$ is calculated as $\psi_{k,m}^j(t) = \tilde{\psi}_{k,m}^j(t) + \bar{\psi}_{k,m}^j(t)$. Since the number of source devices that causes the interference $\tilde{\psi}_{k,m}^j(t)$ varies in each time slots and the sources of direct D2D pairs transmits probabilistically, hence the expected instantaneous SINR is useful in this case, which is approximated as

$$\tilde{\Gamma}_{k,m}^j(t) = \frac{P_k^{Tx} I_{k,m}^j(t)}{N_0 + \mathbb{E}[\tilde{\psi}_{k,m}^j(t)]}$$

where $\mathbb{E}[\cdot]$ is expectation operator and $\mathbb{E}[\psi_{k,m}^j(t)] = \mathbb{E}[\tilde{\psi}_{k,m}^j(t)] + \mathbb{E}[\bar{\psi}_{k,m}^j(t)]$. Let us consider, the set of UAV relays that selects channel $j$ as $\mathcal{M}_j \subseteq \mathcal{M}, j \in J$ and the associated source devices of UAV relay $m_{\tilde{\psi}} \in \{\mathcal{M}_j \setminus m\}$ create interference on UAV $m$, when it receives a signal from source device $k$. Therefore, at time slot $t$ the expected interference at UAV $m$, caused by the source devices of UAV-assisted D2D pair can be estimated as

$$\mathbb{E}[\tilde{\psi}_{k,m}^j(t)] = \sum_{m_{\tilde{\psi}} \in \mathcal{M}_j \setminus m} \sum_{k_{\tilde{\psi}} \in \mathcal{K}_{m_{\tilde{\psi}}}} \frac{P_k^{Tx} G_{k_{\tilde{\psi}},m_{\tilde{\psi}}} I_{k_{\tilde{\psi}},m_{\tilde{\psi}}}^j(t)}{|\mathcal{R}_{m_{\tilde{\psi}}}|}$$

where $\mathcal{R}_{m_{\tilde{\psi}}}$ is the set of the source devices associated with UAV $m_{\tilde{\psi}}$ and $|\mathcal{R}_{m_{\tilde{\psi}}}|$ is the number of devices in the set $\mathcal{R}_{m_{\tilde{\psi}}}$. Similarly, the expected instantaneous interference at UAV $m$, caused by source devices of direct D2D pairs can be estimated as

$$\mathbb{E}[\bar{\psi}_{k,m}^j(t)] = \sum_{k_{\bar{\psi}} \in \mathcal{R}_j} P_{k_{\bar{\psi}}}^{Tx} G_{k_{\bar{\psi}},m_{\bar{\psi}}} I_{k_{\bar{\psi}},m_{\bar{\psi}}}^j(t)$$

where $\mathcal{R}_j \subseteq \mathcal{R}_j$ is the set of the sources of direct D2D pairs which select channel $j$. The total expected interference at time slot $t$ can be calculated as

$$\mathbb{E}[\psi_{k,m}^j(t)] = \mathbb{E}[\tilde{\psi}_{k,m}^j(t)] + \mathbb{E}[\bar{\psi}_{k,m}^j(t)]$$

The expected instantaneous SINR received by destination device $k + K \in K$ from UAV $m \in \mathcal{M}$ over channel $j \in J$, can be expressed as
\[ \hat{I}_{m,k+K}^j(t) = \frac{P_{m}^{Tx}G_{m,k+K}(t)\hat{I}_{k+K,m}^j(t)\hat{I}_{m,j}^j(t)}{N_0 + \mathbb{E}[\psi_{m,k+K}^j(t)]} \] 

(12)

where \( P_{m}^{Tx} \) is transmit power of UAV \( m \), \( K \) is the total number of source and destination device pairs in the target area and \( \mathbb{E}[\psi_{m,k+K}^j(t)] \) is expected instantaneous interference at destination device \( k + K \) when it receives a signal from UAV \( m \) over channel \( j \). This interference occurs when the other UAVs and direct D2D pairs select the same channel as UAV \( m \). Therefore, the expected interference received by the destination device at time slot \( t \) is estimated as

\[ \mathbb{E}[\psi_{m,k+K}^j(t)] = \sum_{m \notin [M] \setminus m} P_{m}^{Tx}G_{m,k+K}(t)\hat{I}_{m,j}^j(t) + \sum_{k \notin K \setminus S_j} P_{k}^{Tx}G_{k,k+K}(t) \] 

(13)

Since source device \( k \) and destination device \( k + K \) are both on the ground, the amplify-and-forward channel model is not suitable for their communication. Hence, we consider the conventional channel model for ground user communication. According to this channel model, instantaneous channel gain between the device \( k \) and \( k + K \) can be expressed as

\[ G_{k,k+K}(t) = \beta_0 D_{k,k+K}^{-q}(t) \] 

(14)

where \( \beta_0 = \left( \frac{4\pi f_c c}{c} \right)^2 \) is free space path loss at distance 1 m and \( q \) is path loss exponent.

The expected instantaneous SINR of direct D2D communication from source device \( k \in K \) to destination \( k + K \) over channel \( j \in J \) can be expressed as

\[ \hat{\Gamma}_{k,k+K}^j(t) = \frac{P_{k}^{Tx}G_{k,k+K}(t)}{N_0 + \mathbb{E}[\psi_{k,k+K}^j(t)]} \] 

(15)

where \( \mathbb{E}[\psi_{k,k+K}^j] \) is expected interference at the receiver of device \( k + K \) when it receives a signal from device \( k \) over channel \( j \) at time slot \( t \), can be measured as

\[ \mathbb{E}[\psi_{k,k+K}^j(t)] = \sum_{m \notin [M] \setminus m} \sum_{\hat{m} \notin K \setminus \hat{m}} P_{k}^{Tx}G_{k,k+K}(t)\hat{I}_{m,j}^j(t) + \sum_{k \notin K \setminus S_j} P_{k}^{Tx}G_{k,k+K}(t) \] 

(16)

The instantaneous expected SINR received by destination device over channel \( j \) must be greater than the threshold \( \gamma \) for reliable communication and maintaining required QoS, then \( \mathbb{E}[\hat{\Gamma}_{k,k+K}^j(t)] > \gamma, \forall k \in K, j \in J \). The expected instantaneous transmission rate of a link between source and destination device of direct D2D pair over channel \( j \in J \) can be expressed as \[24\]

\[ \bar{R}_{k,k+K}^j(t) = B \log_2 \left[ 1 + \hat{\Gamma}_{k,k+K}^j(t) \right], \forall k \in \mathcal{K} \cup \mathcal{K}_D \] 

(17)

where, \( B \) is the bandwidth of the channel. The total instantaneous transmission rate of the links between direct D2D pairs can be calculated as

\[ \bar{R}_{\text{Sum}}(t) = \sum_{j=1}^{J} \sum_{k=1}^{K} \bar{R}_{k,k+K}^j(t) \] 

(18)

The expected instantaneous transmission rate of the link between the source and destination device of UAV-assisted D2D pair via UAV \( m \) and selected channel \( j \in J \), can be formulated as \[24\]
The total instantaneous transmission rate of the links between UAV-assisted D2D pairs can be calculated as:

\[
\hat{R}_{k,m,k+R}^j(t) = \frac{B \hat{I}_{k,m}^j(t)}{2 \gamma_m(t)} \log_2 \left[ 1 + \frac{\hat{I}_{k,m}^j(t) \hat{I}_{m,k+R}^j(t)}{1 + \hat{I}_{k,m}^j(t) + \hat{I}_{m,k+R}^j(t)} \right], \forall m \in \mathcal{M}, k \in \{\mathcal{K}_s \cup \mathcal{K}_d\}
\]  

(19)

The total instantaneous transmission rate of the links between UAV-assisted D2D pairs can be calculated as:

\[
\hat{R}_{\text{Sum}}^j(t) = \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{R}_{k,m,k+R}^j(t)
\]  

(20)

The overall instantaneous transmission rate of the USSD2D network is obtained as:

\[
R_{\text{Sum}}(t) = \hat{R}_{\text{Sum}}(t) + \bar{R}_{\text{Sum}}(t)
\]  

(21)

### 2.3 Problem formulation

When UAVs fly towards some devices to acquire a better channel environment, the remaining devices of the network cannot receive adequate service from the UAVs. Moreover, the transmission rate of the devices which are far away from UAVs becomes lower than those devices that are closer to the UAVs. We consider that one source device can associate only one UAV for aid but one UAV can be shared by multiple devices. Furthermore, each UAV can only occupy a single orthogonal channel for each time slot. In order to maximize the total instantaneous transmission rate of the USSD2D network, we aim to jointly optimize the deployed location of UAVs, device association, and UAVs’ channel selection at each slot, while assuring that the expected instantaneous SINR received by each destination device achieve a threshold \( \bar{\gamma} \) for maintaining the required QoS. The corresponding optimization problem is formulated as

Maximize

\[
P1: \left\{ \left( x_m(t), y_m(t), I_{k,m}(t), I_{m,j}(t) \right), \forall k \in \mathcal{K}, m \in \mathcal{M}, j \in \mathcal{J} \right\} R_{\text{Sum}}(t)
\]  

(22)

Subjected to constraints

\[
C1: \mathbb{E}[\gamma_{k,k+R}^j(t)] > \bar{\gamma}, \forall k \in \mathcal{K}, j \in \mathcal{J}
\]  

(23a)

\[
C2: I_{k,m}(t) \in \{0,1\}, I_{m,j}(t) \in \{0,1\}, \forall k \in \mathcal{K}, m \in \mathcal{M}, j \in \mathcal{J}
\]  

(23b)

\[
C3: \sum_{m \in \mathcal{M}} I_{k,m}(t) \leq 1, \forall k \in \mathcal{K}
\]  

(23c)

\[
C4: \sum_{j \in \mathcal{J}} I_{m,j}(t) \leq 1, \forall m \in \mathcal{M}
\]  

(23d)

Here, C1 assures that the USSD2D network maintains the required QoS if the received instantaneous SINR of each destination device is greater than the given threshold; C2 specifies the device association indicator and UAVs’ channel selection indicator constraint at every time slot. C3 defines that at every time slot each device associates with a single UAV; and C4 implies that each UAV selects an orthogonal channel from the available channel set in each time slot.

The formulated problem is complicated and intractable because the variables \( x_m(t), y_m(t), I_{k,m}(t) \) and \( I_{m,j}(t) \) are coupled where the deflection of one variable impacts the optimization of other variables and the value of the objective function. Moreover, due to the absence of a central controller, path loss and gain of the selected channel, as well as the location of the communication node are prior unknown. Hence this non-convex and combinatorial optimization problem cannot be solved by the classical non-convex optimization method. To tackle these challenges, an online model-free learning-based UAV deployment strategy is needed where the aim is to find the optimal deployed location of UAVs via obtaining the required system parameters using real-time measurements and statistics of collected information without previous information of the channel model and device location in an unknown communication environment.
3 Dynamics of optimal decision making policy

The advanced automatic systems are trained to make intelligent decisions using RL schemes where RL agent communicates with an unknown environment directly by receiving feedback as a reward or penalty corresponding to the quality of taken action to perform a specific job [25]. Being an RL agent, each UAV adopts the environment through a self-learning procedure and decides an action according to its learning experiences. Conventionally [26], the optimal deployment of UAVs, device association, and UAVs’ channel selection has been obtained by UAV assisted relaying system under the offline framework when the devices know the exact ground to air channel conditions. However, in our model, devices only know the causal information, such as its association probability, UAV’s instantaneous location, and channel selection probability at the current time slot. But the statistical information corresponding to those parameters cannot be obtained for all time slots. Therefore, we refer the objective function mentioned by (22) to a non-convex and combinatorial optimization problem, and the corresponding suboptimal solution can be regarded as a sequential decision-making policy which is modeled by Markov Decision Process (MDP). These suboptimal solutions obtained by MDP for each time slot indicate the throughput of the proposed USSD2D network in terms of the total instantaneous transmission rate.

3.1 Analysis of model-free MDP

In order to solve this MDP, we adopt the online RL method because it does not require statistical information. At each time slot, devices can only obtain causal information. The action selected by UAVs entirely depends on their instantaneous position. The current states of UAVs are only related to the state and action of the previous time slot. Thus, this system satisfies the Markov properties and can be composed of quintuple \((S,A,P_{tr},R,N)\), where \(S\) and \(A\) denote the state and action space respectively, \(P_{tr}\) is state transition probability, \(R\) is the reward, and \(T\) is a set of time slots in which certain decisions need to be made. Specifically, for our optimization problem, the definitions of each tuple are explained as follows.

3.1.1 State space

In our scenario, we can only get \(x_m(t)\) and \(y_m(t)\), \(\forall m \in M\). Hence, we define the state of UAV \(m\) at time slot \(t\) as a vector of two elements \(x_m(t)\) and \(y_m(t)\) which represents the current location of UAV \(m\) denoted by \(s_m(t) \in S\). The elements of state-space are independent and identically distributed random variables across the horizon of \(T\) time slots and \(S\) is the set of all state elements arranged by the combination of all possible values.

![Fig. 2: Illustration of UAVs’ action space corresponding to each possible state.](image)

3.1.2 Action space

To solve the optimization problem described in (22), UAVs estimate their deployed locations at every time slot. We define the instantaneous action of UAV \(m\), \(\forall m \in M\) as a change of its location denoted by \(a_m(t) \in A\). These changed locations are measured with respect to the X and Y coordinate of UAVs’, obtained at the current time slot. Each UAV explores the sensing region by selecting a moving direction and moves towards the selected direction for a distance traveled by slot duration time length in each time slot. For better understanding, the moving direction of each UAV is illustrated in Fig. 2. It is observed from this figure that, UAV has a maximum of eight possible moving directions at each state i.e. \{east, north-east, north, north-west, west, south-west, south, and south-east\}. After selecting a moving direction, the corresponding X and Y coordinate changes of UAV \(m\) at time slot \(t\) are denoted as \(\delta_x^m(t) \in \{0, \theta(t)\delta\}\) and \(\delta_y^m(t) \in \{0, \theta(t)\delta\}\) respectively, \(\forall a_m(t) = \{\delta_x^m(t), \delta_y^m(t)\} \in A, t \in T\), where \(\theta(t)\) is the velocity of each UAV at time slot \(t\).
and \( \mathcal{A} \) is the action set where all possible actions can store. Therefore, the obtained X and Y coordinate of UAV \( m \) at time slot \( t + 1 \) is measured as

\[
x_m(t + 1) = x_m(t) + \delta_x^m(t) \\
y_m(t + 1) = y_m(t) + \delta_y^m(t)
\]

\[ (24) \]

\[ (25) \]

### 3.1.3 Reward formulation

The generated reward depends on the device association, UAVs’ channel selection and their current state as well as taken action at each time slot. As a device can associate with a single UAV and a UAV is shared by multiple devices over a particular channel, therefore we need to find such locations of UAVs that maximize the total instantaneous transmission rate of the USSD2D network. The expected deployed location of UAV \( m \), \( \forall m \in \mathcal{M} \) is obtained more precisely when its generated reward at the current time slot is beneficial for the long term. To reflect this property, we model the instantaneous reward function of UAV \( m \) as

\[
\mathcal{R}(s_m(t), a_m(t)) = \sum_{j=1}^{J} \sum_{k=1}^{K} \tilde{R}_{k,m,k+R}^j(t) + \sum_{j=1}^{J} \sum_{k=1}^{K} \tilde{R}_{k,k+R}^j(t), \forall m \in \mathcal{M}
\]

\[ (26) \]

### 3.1.4 State transition probability

It is defined as the probability that is obtained by UAV \( m \), \( \forall m \in \mathcal{M} \) for the transition from state \( s_m(t) \) to \( s_m(t + 1) \) after selecting an action \( a_m(t) \in \mathcal{A} \) at the time slot \( t \) which is denoted as \( P_{tr}\{s_m(t + 1) \in S|s_m(t), a_m(t)\} \). Let us consider the probability vectors of device association and UAVs’ channel selection at time slot \( t \) as \( P_{k}^{DA}(t) = [\tilde{P}_{k,1}(t), \tilde{P}_{k,2}(t), ..., \tilde{P}_{k,m}(t)], \forall k \in \{\mathcal{K}_S \cup \mathcal{K}_D\} \) and \( P_{m}^{CS}(t) = [\tilde{P}_{m,1}(t), \tilde{P}_{m,2}(t), ..., \tilde{P}_{m,j}(t)], \forall m \in \mathcal{M} \) respectively where \( \tilde{P}_{k,m}(t) \) indicates the association probability of device \( k \) with UAV \( m \) at time slot \( t \) and \( \tilde{P}_{m,j}(t) \) is the probability that UAV \( m \) selects channel \( j \) at time slot \( t \). In each time slot, source and destination device associate with a single UAV, and that UAV selects a single orthogonal channel from the available channel set according to probability vectors \( P_{k}^{DA}(t) \) and \( P_{m}^{CS}(t) \) respectively. While each UAV changes its state by selecting an action, the probability vectors corresponding device association and UAVs’ channel selections are updated for the next time slot as follow

\[
\tilde{P}_{k,m}(t + 1) = \begin{cases} \\
\tilde{P}_{k,m}(t) + w_1 \tilde{r}_{k,m}(t) \left( 1 - \tilde{P}_{k,m}(t) \right), & m = U_{k}^{Max}(t) \\
\tilde{P}_{k,m}(t) - w_1 \tilde{r}_{k,m}(t) \tilde{P}_{k,m}(t), & m \neq U_{k}^{Max}(t)
\end{cases}
\]

\[ (27) \]

\[
\tilde{P}_{m,j}(t + 1) = \begin{cases} \\
\tilde{P}_{m,j}(t) + w_2 \tilde{r}_{m,j}(t) \left( 1 - \tilde{P}_{m,j}(t) \right), & j = C_{m}^{Max}(t) \\
\tilde{P}_{m,j}(t) - w_2 \tilde{r}_{m,j}(t) \tilde{P}_{m,j}(t), & j \neq C_{m}^{Max}(t)
\end{cases}
\]

\[ (28) \]

where \( w_1 \) and \( w_2 \) are the learning step sizes. \( U_{k}^{Max}(t) \) and \( C_{m}^{Max}(t) \) are current best UAV for device \( k \) when selected channel by that UAV is fixed for that time slot and the current best channel of UAV \( m \) for associated devices at that time slot respectively, which can be expressed as

\[
U_{k}^{Max}(t) = \arg\max_{m \in \mathcal{M}} \tilde{R}_{k,m,k+R}(t), \forall k \in \{\mathcal{K}_S \cup \mathcal{K}_D\}
\]

\[ (29) \]

\[
C_{m}^{Max}(t) = \arg\max_{j \in \mathcal{J}} \tilde{R}_{m,j}(t), \forall m \in \mathcal{M}
\]

\[ (30) \]

where \( \tilde{R}_{k,m}(t) \) and \( \tilde{r}_{m,j}(t) \) are the normalized reward achieved by the source device \( k \) and UAV \( m \) at time slot \( t \) respectively, which are defined as

\[
\tilde{R}_{k,m}(t) = \frac{\tilde{R}_{k,m,k+R}(t)}{\max_{m \in \mathcal{M}} \tilde{R}_{k,m,k+R}(t)}
\]

\[ (31) \]

\[
\tilde{r}_{m,j}(t) = \frac{\tilde{R}_{m,j}(t)}{\max_{j \in \mathcal{J}} \tilde{R}_{m,j}(t)}
\]

\[ (32) \]
From (27) and (28), it is observed that the update of selection probability vectors depends on instantaneous transmission rate which does not need any prior information. Thus, the process of device association and UAVs’ channel selection at each time slot is completely distributive and model-free.

3.1.5 Time steps

The optimal deployment of UAVs is referred to multi-period decision-making problem where UAVs can learn to take decisions via interaction with the environment through the episodic task. At time slot \( t \), UAV \( m \), \( \forall m \in \mathcal{M} \) makes a strategy to select \( a_m(t) \) according to \( \sigma_m(t) \) and obtains a new state \( \sigma_m(t + 1) \). We define the transition of UAV \( m \) from the state \( \sigma_m(t) \) to \( \sigma_m(t + 1) \) as one step and this transition is repeated until the end of \( T \) time slots. Thus, the sequential decision making policy corresponding to the location of UAVs acquires the suboptimal solution which maximizes the total instantaneous transmission rate of the USSD2D network over \( T \) time slots.

3.2 On policy approach

For obtaining a suboptimal decision-making policy, we use RL based method. The RL model is illustrated in Fig. 3 where each UAV acting as an RL agent to build up its knowledge about the surrounding environment by an agent-environment interaction process. In this formulated optimization problem, each UAV decides its location at every time slot. As UAV \( m \), \( \forall m \in \mathcal{M} \) does not know the communication environment, therefore it tentatively selects an action \( a_m(t) \) at state \( \sigma_m(t) \) and gets a corresponding \( Q(\sigma_m(t), a_m(t)) \) value of state-action pair and immediately receives the expected reward \( R(\sigma_m(t), a_m(t)) \). This reward is used to measure the effectiveness of the selected action for the current state. Then, the state \( \sigma_m(t) \) will update to the next state \( \sigma_m(t + 1) \). From Fig. 3, we can see that the agent continuously interacts with the environment and finally find a suboptimal decision-making policy by obtaining the characteristics of the environment. Since in our proposed scenario UAVs encounter a lack of plenary knowledge of the environment, therefore State-Action-Reward-State-Action (SARSA), an on-policy temporal difference (TD) learning algorithm is assigned where the initial conditions for updating the policy need to be assumed [27]. Being iterative dynamic programming, SARSA is less computationally complex and finds the suboptimal solution by the currently defined policy.

![Fig. 3: Basic block diagram of reinforcement learning.](image)

4 Reinforcement learning based on SARSA method

4.1 Mapping between state and action for optimal decision

The multi-period decision-making problem is formulated by the model-free TD method and it aims to approximate the state-action value function via RL. The model-free TD method consists of on-policy which is used to control MDP. The on-policy is a mapping function from state-space to the probability distribution of selecting each possible action. Therefore, UAVs’ policy indicates the probability distribution of changed location in each time slot for each possible state. UAV \( m \), \( \forall m \in \mathcal{M} \) experiences the environment by taking suitable action in a particular state following policy \( \pi_m \in \Pi \) where the expected mapping value between state and action can be expressed using Bellman equation [28] as

\[
V(\sigma_m(t)) = \max_{a_m(t)} \{ R(\sigma_m(t), a_m(t)) + \gamma \mathbb{E}[V(\sigma_m(t + 1)) | \sigma_m(t), a_m(t)] \}
\]

where \( \gamma \in [0,1] \) is discount factor and \( \mathbb{E}[.] \) is expectation operator. Using the value iteration method, the solution of the equation (33) is denoted as \( V^* \) which can obtain the optimal policy as
∀\(m \in \mathcal{M}\) maximize of the continuous interaction with the environment are recorded in a table for choosing the best option from the action space. Using the value iteration method, the state value function is expressed as

\[
V^\pi_m(s_m(t)) = \mathbb{E} \left[ \sum_{\tau=t}^{\infty} \gamma^{\tau-t} R(s_m(\tau), a_m(\tau)) \mid s_m(t), a_m(\tau) \in \pi_m(a_m(\tau) \mid s_m(\tau)) \right] = R(s_m(t), \pi_m(a_m(t) \mid s_m(t))) + \gamma \mathbb{E}[V^\pi_m(s_m(\tau + 1)) \mid s_m(t), \pi(a_m(t) \mid s_m(t))] \quad (35)
\]

The optimal decision-making policy can be achieved by the value iteration method where the state transition probability is required to solve the MDP. Thus, we develop an online model-free MDP-RL algorithm to find the suboptimal decision-making policy.

### 4.2 Updating the action value function using SARSA

The value iteration based on Q-learning algorithm is executed by exploration and exploitation stages. Exploration motivates UAVs to learn from the environment, opt for a specific action, observe reward and then update their action policy. More specifically, UAV \(m\), \(\forall m \in \mathcal{M}\) selects an action \(a_m(t)\) at the state \(s_m(t)\). Then \(s_m(t)\) will update to \(s_m(t + 1)\) and corresponding \(Q(s_m(t), a_m(t))\) value and immediate reward \(R(s_m(t), a_m(t))\) are generated. The results of the continuous interaction with the environment are recorded in a Q-table which indicates the state action value function corresponding a step of an episode. After approximating the Q-table by the above repeated process, this Q-learning algorithm goes into the exploitation phase. During this phase, each UAV takes the advantage of previously gained Q-table for choosing the best option from the action space. Using the value iteration method, Q-learning algorithm updates the \(Q(s_m(t), a_m(t))\) value corresponding UAV \(m\) at each episode as [29]

\[
Q(s_m(t), a_m(t)) \leftarrow (1 - \alpha)Q(s_m(t), a_m(t)) + \alpha \left[ R(s_m(t), a_m(t)) + \gamma \max_{a \in A} Q(s_m(t + 1), a) \right] \quad (36)
\]

where \(\alpha \in [0,1]\) is the learning rate. However, the update method of \(Q(s_m(t), a_m(t))\) value in Q-learning algorithm will lead to an incomplete exploration of the environment and the best action can not be distinguished among all the actions. Hence, we use the SARSA algorithm to mitigate this problem.

Similar to Q-learning, SARSA also follows the exploration and exploitation process and record \(Q(s_m(t), a_m(t))\) value corresponding \(s_m(t)\) and \(a_m(t)\) in Q-table. But there is some difference between Q-learning algorithm and SARSA algorithm which is explained as follows. According to the action selection scheme, UAV \(m\), \(\forall m \in \mathcal{M}\) selects an action \(a_m(t)\) at current state \(s_m(t)\). Then UAV \(m\) receives immediate reward \(R(s_m(t), a_m(t))\) and corresponding \(Q(s_m(t), a_m(t))\) value. Finally, \(s_m(t)\) is updated to the next state \(s_m(t + 1)\) and UAV \(m\) selects the next action \(a_m(t + 1)\). Consequently, we can get a quintuple \((s_m(t), a_m(t), R(s_m(t), a_m(t)), s_m(t + 1), a_m(t + 1))\) which is used to update the action-value function as [30]

\[
Q(s_m(t), a_m(t)) \leftarrow (1 - \alpha)Q(s_m(t), a_m(t)) + \alpha \left[ R(s_m(t), a_m(t)) + \gamma Q(s_m(t + 1), a_m(t + 1)) \right] \quad (37)
\]

Unlike Q-learning algorithm, SARSA updates the Q-value in an online manner. During the learning process, UAVs consider all the possible actions from the action space and select an action with a certain probability that provides maximum reward. Hence SARSA algorithm achieves enough experience about the environment to find the suboptimal solutions. Choosing suitable action at every state is the most important aspect of RL. \(\epsilon\)-greedy policy is applied to maximize Q-value which obtains better performance under the finite state and action space. The probability that UAV \(m\), \(\forall m \in \mathcal{M}\) takes action \(a_m(t) \in \mathcal{A}\) corresponding state \(s_m(t) \in \mathcal{S}\) at time slot \(t\) under \(\epsilon\)-greedy policy can be expressed as [29]

\[
\pi^*_m = \begin{cases} 
\arg \max_{a_m(t) \in \mathcal{A}(s_m(t), s_m(t))} Q(s_m(t), a_m(t)), & \text{with probability } 1 - \epsilon \\
\text{Random Selection}, & \text{with probability } \epsilon 
\end{cases} \quad (38)
\]
4.3 Optimizing UAVs strategy based on SARSA

The transition of a state corresponding to changes of successive time slots is regarded as an episode and the total number of episodes is denoted as \( T \). Each UAV tries state-action pairs repeatedly in every episode and consequently experiences the interactions with the environment. The interaction result is recorded in \( Q \)-table and updates the learning policy in each episode until it converges. We can get a decision-making policy corresponding to a suboptimal solution when UAVs change their location. Algorithm 1 summarizes the optimal deployment learning procedure using the adaptive SARSA technique.

**Algorithm 1:** Optimal UAV deployment strategy using adaptive SARSA technique

**Input:** \( N_0, B, \mu_{\text{los}}, \mu_{\text{Nlos}}, \bar{f}_c, b_1, b_2, H_u, \theta_0, \bar{K}, \bar{R}_S, \bar{R}_D, \bar{R}_S, \bar{R}_D, \bar{R}_{\text{tx}}, M, M, p_{m}^{\text{tx}}, J, J, w_1, w_2, \gamma, \alpha, \epsilon, \zeta \)

\( \forall s_m(t) \in S, a_m(t) \in A, k \in K = \{ \bar{R}_S \cup \bar{R}_D \cup \bar{R}_S \cup \bar{R}_D \}, m \in M \)

**Output:** Instantaneous reward generated by all UAVs as \( R(t) \)

1. Initialize \( Q(s, a) = 0, \forall s \in S, a \in A \)
2. Set initial device association probability as \( \bar{P}_k,m(1) = \frac{1}{M}, \forall k \in \{ \bar{R}_S \cup \bar{R}_D \}, m \in M \)
3. Set initial channel selection probability of UAVs as \( \bar{P}_m,j(1) = \frac{1}{J}, \forall m \in M, j \in J \)
4. Initially deploy UAV \( m, \forall m \in M \) at the random position as \( s_m(1) = (x_m(1), y_m(1), H_u) \)
5. for \( t = 1, 2, \ldots, T \) do
6. for \( k = 1, 2, \ldots, K \) do
7. for \( m = 1, 2, \ldots, M \) do
8. Obtain the association probability of device \( k \) with UAV \( m \) as \( \bar{P}_k,m(t) \)
9. Calculate \( \bar{f}_k,m(t) \) and \( \bar{f}_{m,k+k}(t) \) according to (8) and (12) respectively for a fixed assigned channel
10. if \( \bar{f}_k,m(t), \bar{f}_{m,k+k}(t) \geq \zeta \) then
11. Calculate \( \bar{R}_k,m,k+k(t) \) using (19) for a fixed assigned channel
12. else
13. \( \bar{R}_k,m,k+k(t) = 0 \)
14. end if
15. end for
16. end for
17. for \( k = 1, 2, \ldots, K \) do
18. for \( m = 1, 2, \ldots, M \) do
19. Set \( \bar{f}_k,m(t) = 1 \) when \( m = \arg \max_{m \in M} \bar{P}_k,m(t) \), otherwise \( \bar{f}_k,m(t) = 0 \)
20. According to (27), update the association probability as \( \bar{P}_k,m(t) \leftarrow \bar{P}_k,m(t + 1) \)
21. end for
22. end for
23. for \( m = 1, 2, \ldots, M \) do
24. for \( j = 1, 2, \ldots, J \) do
25. UAV \( m \) obtains the \( j^{th} \) channel selection probability as \( \bar{P}_m,j(t) \)
26. Calculate \( \bar{f}_m,j(t) \) according to (8) for the fixed associated devices
27. if \( \bar{f}_m,j(t) \geq \zeta \) then
28. \( \bar{R}_m,j(t) = \sum_{k=1}^{K} B \log_2 \left[ 1 + \bar{f}_m,j_k(t) \right] \)
29. else
30. \( \bar{R}_m,j(t) = 0 \)
31. end if
32. end for
33. end for
34. for \( m = 1, 2, \ldots, M \) do
35. for \( j = 1, 2, \ldots, J \) do
36. Set \( \bar{I}_{m,j}(t) = 1 \) when \( j = \arg \max_{j \in J} \bar{P}_{m,j}(t) \), otherwise \( \bar{I}_{m,j}(t) = 0 \)
37. According to (28), update channel selection probability as \( \bar{P}_{m,j}(t) \leftarrow \bar{P}_{m,j}(t + 1) \)
38. end for
39. end for
for $m = 1, 2, \ldots, M$ do

38: Choose the action values $a_m(t) = \{\delta_x^m(t), \delta_y^m(t)\}$ by (38)

39: Find next state as $s_m(t + 1) = (x_m(t + 1), y_m(t + 1), H_u)$ where $x_m(t + 1)$ and $y_m(t + 1)$ are calculated by (24) and (25)

40: Calculate the immediate reward $R(s_m(t), a_m(t))$ of UAV $m$ by (26)

41: Choose the action $a_m(t + 1) = \{\delta_x^m (t + 1), \delta_y^m (t + 1)\}$ by (38) and obtain corresponding $Q(s_m(t + 1), a_m(t + 1))$ value from $Q$-table

42: Update $Q(s_m(t), a_m(t))$ value according to (37) and store it in $Q$-table

43: Update the state and action for the next time slot as $s_m(t) ← s_m(t + 1)$ and $a_m(t) ← a_m(t + 1)$ respectively

44: Calculate the instantaneous reward generated by all UAVs as $R(t) = \sum_{m=1}^{M} R(s_m(t), a_m(t))$

end for

It is clear from the above analysis that SARSA is a table-based algorithm that needs a $Q$-table to record the value of each state-action pair. In this algorithm, UAVs select the action according to the $\epsilon$-greedy action selection scheme at each time slot. As, the total number of episodes is $T$ and each episode estimates the contribution of $M$ UAVs, thus the computational complexity depends on the total steps $TM$ including the dimension of the state and action space of the proposed RL. In our scenario, there are two-dimensional state locations and eight possible actions for each time slot. Therefore, the computational complexity of Algorithm 1 is $O(16TM)$ and it includes the complexity of $\epsilon$-greedy policy in each step. Since SARSA estimates the optimal $Q$-value by iterative manner, hence its convergence relies on the learning scheme. The $Q$-table will converge if learning rate $\alpha$ satisfies the conditions $\sum_{t=0}^{\infty} \alpha = \infty$ and $\sum_{t=0}^{\infty} \alpha^2 < \infty$ ($0 \leq \alpha \leq 1$) regardless of the initial settings. Moreover, the velocity of each UAV at time slot $t$ represented by $\vartheta(t)$ influences the action selection probability. When $\vartheta(t) \to 0$, UAVs select random action with equal probability which indicates an inefficient search technique. On the other way if $\vartheta(t) \to \infty$, UAVs always select the action with maximum reward. To find the optimal deployed location of each UAV efficiently, its velocity should be a decreasing function of the time slot index. Therefore, the instantaneous velocity of each UAV is expressed as $\vartheta(t) = \frac{\vartheta_0}{t}$ where $\vartheta_0$ is the initial velocity of UAVs.

5 Simulation Results and Discussions

In this section, we perform the required simulation experiments and corresponding numerical results are presented to validate the performance of the proposed optimization framework. To do so first, we investigate the converging outcomes of the proposed SARSA algorithm for optimal deployment and subsequently evaluate the performance of reliable communication between source and destination device pairs. Finally, the effectiveness and superiority of the proposed design are compared with the existing baseline [22] i.e., random selection with fixed optimal relay deployment (RS-FORD), an exhaustive search for relay assignment and channel allocation with fixed initial relay deployment (ES-FIRD), and alternative optimization for the individual variable (AOIV) by considering randomly deployed direct D2D pair and UAV-assisted D2D pair devices within a squared area of size $4 \text{km} \times 4 \text{km}$. Furthermore, we adopt the primary simulation parameters from [22] and [30] which are summarized in Table 1.

Table 1: Simulation parameters

| Parameters | Values |
|------------|--------|
| $R$        | 10     |
| $T$        | 1000   |
| $\mu_{\text{LoS}}$ | 1 dB |
| $\mu_{\text{LoS}}$ | 20 dB |
| $b_1$      | 9.61   |
| $b_2$      | 0.16   |
| $p_{m,Tx}$ | 10 mW |
| $\gamma$  | 0.9    |
| $\alpha$  | 0.1    |
| $w_1$      | 0.004  |
| $w_2$      | 0.002  |

| Parameters | Values |
|------------|--------|
| $\delta$  | 1 sec  |
| $J$        | 7      |
| $f_c$      | 2 GHz  |
| $\varsigma$ | 1 dB |
| $N_0$      | -174 dBm/Hz |
| $\vartheta_0$ | 100 m/s |
| $\epsilon$ | 0.1   |
| $M$        | 5      |
| $H_u$      | 500 m  |
| $B$        | 1 MHz  |
| $N$        | -174 dBm/Hz |

It is clear from the analysis that SARSA is a table-based algorithm that needs a $Q$-table to record the value of each state-action pair. In this algorithm, UAVs select the action according to the $\epsilon$-greedy action selection scheme at each time slot. As, the total number of episodes is $T$ and each episode estimates the contribution of $M$ UAVs, thus the computational complexity depends on the total steps $TM$ including the dimension of the state and action space of the proposed RL. In our scenario, there are two-dimensional state locations and eight possible actions for each time slot. Therefore, the computational complexity of Algorithm 1 is $O(16TM)$ and it includes the complexity of $\epsilon$-greedy policy in each step. Since SARSA estimates the optimal $Q$-value by iterative manner, hence its convergence relies on the learning scheme. The $Q$-table will converge if learning rate $\alpha$ satisfies the conditions $\sum_{t=0}^{\infty} \alpha = \infty$ and $\sum_{t=0}^{\infty} \alpha^2 < \infty$ ($0 \leq \alpha \leq 1$) regardless of the initial settings. Moreover, the velocity of each UAV at time slot $t$ represented by $\vartheta(t)$ influences the action selection probability. When $\vartheta(t) \to 0$, UAVs select random action with equal probability which indicates an inefficient search technique. On the other way if $\vartheta(t) \to \infty$, UAVs always select the action with maximum reward. To find the optimal deployed location of each UAV efficiently, its velocity should be a decreasing function of the time slot index. Therefore, the instantaneous velocity of each UAV is expressed as $\vartheta(t) = \frac{\vartheta_0}{t}$ where $\vartheta_0$ is the initial velocity of UAVs.
5.1 Convergence analysis

For the considered case of randomly deployed static device pairs in the USSD2D network, we perform the training operations and corresponding results are illustrated in Fig. 4. At the beginning of the simulation, we assign two fixed channels among the direct D2D device pairs. It is observed in this figure that UAVs are capable of carrying out their actions in an iterative manner and learn from mistakes to improve the instantaneous transmission rate of the USSD2D network. According to the results of these simulations, it is obvious that the total instantaneous transmission rate of the USSD2D network corresponding to the proposed SARSA algorithm outperforms that of the RS-FORD, ES-FIRD, and AOIV scheme. The reason is that the proposed SARSA algorithm using $\epsilon$-greedy policy obtains large state space which helps UAVs to explore the target region more efficiently and consequently mutual interferences among UAVs and devices can be reduced by jointly optimizing device association, UAVs’ channel selection, and their optimal deployment. From Fig. 4, we can say that RL based on SARSA methodology improves the overall instantaneous transmission rate of the USSD2D network by 71.39%, 48.55%, and 14.45% respectively compared with RS-FORD, ES-FIRD, and AOIV schemes.

![Fig. 4: The variation of the total transmission rate of the USSD2D network corresponding to each episode](image)

5.2 Performance analysis of the proposed USSD2D network

Fig. 5(a) represents the variations of the total instantaneous transmission rate of the network with the different number of UAVs corresponding to proposed and benchmark schemes when the UAV-assisted D2D pairs, direct D2D pairs, and available orthogonal channels are set as 10, 2, and 7 respectively. It is observed from the figure that the proposed SARSA algorithm outperforms the benchmark schemes and the total instantaneous transmission rate of the network increases while the number of UAVs increases because UAVs utilize all the available orthogonal channels efficiently at their optimal deployed location. We can see from Fig. 5(a) that the total instantaneous transmission rate of the network does not change significantly when the number of UAVs exceeds 7 because due to limited numbers of orthogonal channels, more UAV reuse the allocated spectrum which enhances mutual interferences among UAVs and source-destination device pairs. RL based on SARSA methodology improves the overall instantaneous transmission rate of the USSD2D network by 73.53%, 50.05%, and 13.94% respectively compared with RS-FORD, ES-FIRD, and AOIV schemes.

Fig. 5(b) plots the total instantaneous transmission rate of the network under different numbers of channels corresponding to proposed and benchmark schemes when UAV-assisted D2D pairs, direct D2D pairs, and number of UAVs are fixed as 10, 2, and 5 respectively. It is clear from the figure that the proposed SARSA algorithm improves the instantaneous transmission rate of the network compared to benchmark schemes due to efficient estimation of the channel selection probability vector. Apart from that, the instantaneous transmission rate of the network does not increase any more when the numbers of available orthogonal channels are greater than 7 because these are sufficient resources to mitigate the mutual interferences completely. From Fig. 5(b), we can say that RL based on SARSA methodology improves the overall instantaneous transmission rate of the USSD2D network by 77.33%, 55.45%, and 13.93% respectively compared with RS-FORD, ES-FIRD, and AOIV schemes.

Fig. 5(c) shows the total instantaneous transmission rate of the network versus the number of UAV-assisted D2D pairs corresponding to proposed and benchmark schemes while the number of direct D2D pairs, available orthogonal channels, and number of UAVs are constant i.e., 2, 7, and 5 respectively. From Fig. 5(c) it can be confirmed that the instantaneous
transmission rate of the network almost remains unchanged with the increase of the number of UAV-assisted device pairs. Since all the communication nodes in the network utilize a fixed amount of resources, therefore no such variation is found in this figure and we can say that RL based on SARSA methodology improves the overall instantaneous transmission rate of the USSD2D network by 71.85%, 48.89%, and 14.66% respectively compared with RS-FORD, ES-FIRD, and AOIV schemes.

The variations of the total instantaneous transmission rate of the network corresponding to the proposed and benchmark schemes are illustrated in Fig. 5(d) when the number of UAV-assisted D2D pairs, available orthogonal channels, and number of UAVs are set as 10, 7, and 5 respectively. It is observed from the figure that the total instantaneous transmission rate of the network decreases with the increase of the number of direct D2D pairs. When the number of direct D2D pairs is increased, the fixed number of orthogonal channels is allocated to the greater number of communication nodes. As a result, mutual interferences among devices and UAVs increase which leads to the reduction of the network throughput due to limited resources. From Fig. 5(d), we can say that RL based on SARSA methodology improves the overall instantaneous transmission rate of the USSD2D network by 78.46%, 54.2%, and 13.52% respectively compared with RS-FORD, ES-FIRD, and AOIV schemes.

Fig. 6 presents the impact of various UAVs’ altitudes on the total instantaneous transmission rate of the USSD2D network corresponding to the proposed and conventional schemes. We can see from this figure that the total instantaneous transmission rate of the network increases at a certain limit then it decreases gradually with the increasing value of the UAVs’ altitude. Since LoS probability increases when altitudes of UAVs increase and consequently instantaneous transmission rate also increase. But when the altitude exceeds a certain limit, the transmission distance between UAV to
its associated device increases which leads to a reduction of the instantaneous transmission rate. Moreover, there is still a gap between the proposed and conventional schemes. The reason is that the proposed optimization method efficiently estimates the location of the devices and accordingly it deploys the UAVs at a given altitude, whereas the conventional schemes deploy UAVs using successive learning techniques without proper approximation of devices’ location. From Fig. 6, we can say that RL based on SARSA methodology improves the overall instantaneous transmission rate of the USSD2D network by 79.67%, 55.36%, and 18.11% respectively compared with RS-FORD, ES-FIRD, and AOIV schemes.

![Fig. 6: Impact of various UAVs’ altitudes on network throughput.](image)

5.3 Impact of learning parameters on deployment strategy

According to (27) and (28), it is observed that learning step sizes and initial velocity of UAVs influence the evolution of instantaneous transmission rate of the network and convergence properties corresponding to the proposed SARSA algorithm. Fig. 7(a) plots the total instantaneous transmission rate of the USSD2D network for each episode with different values of learning step sizes when \( \theta_0 = 100 \) m/s. We can see that, with decreasing values of \( w_1 \) and \( w_2 \), the convergence rate of the proposed SARSA algorithm decreases but the total instantaneous transmission rate of the network increase because small values of learning step size encourage SARSA to update the device association probability and UAVs’ channel selection probability more efficiently.

![Fig. 7: Total instantaneous transmission rate of the USSD2D network with different learning parameters.](image)
Fig. 7(b) plots the total instantaneous transmission rate of the USSD2D network for each episode with the different initial velocities of UAVs’ when the learning step sizes are fixed as $w_1 = 0.004$ and $w_2 = 0.002$. We can see that the too small and too fast initial velocity of UAVs achieves a low convergence rate and poor performance while the moderate initial velocities of the UAVs achieve a fast convergence rate and good performance. This is because too small and too large initial velocity motivates UAVs for random movement which leads to an incomplete exploration of the environment and consequently UAVs do not have a chance to choose other actions. Therefore, the probability of the best-chosen action is equally matched as the worst action.

6 Conclusion
In this article, we have presented a UAV-supported self-organized device-to-device (USSD2D) network where multiple UAVs are deployed to support the reliable communication between source-destination device pairs. The above analysis shows that the performance of the USSD2D network is impacted by the UAVs’ position, in which the goal is to maximize the total instantaneous transmission rate of the network by jointly optimizing the deployed location of UAVs, device association, and UAVs channel selection under SINR constraint. The formulated objective function is verified as a multi-period decision-making problem that is solved by RL based on the SARSA algorithm. UAVs pretending as RL agents select the optimal policy that maximizes the long-term reward by converging towards maximum values. Numerical results indicate that the proposed methodology provides much better performance than existing baselines and improves the total instantaneous transmission rate of the network by 75.37%, 52.08%, and 14.77% respectively as compared with RS-FORD, ES-FIRD, and AOIV schemes. We hope that this work would be extended for meaningful guidance and inspiration of developing future multi-UAV networks, especially the enhancement of network throughput in 5G and beyond communication systems.

7 Declaration

7.1 Funding
Not applicable

7.2 Conflicts of interest/Competing interests
On behalf of all authors, the corresponding author states that there is no conflict of interest

7.3 Availability of data and material
Not applicable

7.4 Code availability
Not applicable

References
1. Bekkouche, O., Samdanis, K., Bagaa, M., & Taleb, T. (2020). A service-based architecture for enabling UAV enhanced network services. *IEEE Network, 34*(4), 328–335. https://doi.org/10.1109/MNET.001.1900556
2. Al-Shehri, S. M., Loskot, P., & Hirsch, M., J. (2018). Enabling connectivity for tactical networks in mountainous areas by aerial relays. *Telecommunication Systems, 71*, 561–575. https://doi.org/10.1007/s11235-018-0532-3
3. You, C. & Zhang, R. (2019). 3D trajectory optimization in Rician fading for UAV-enabled data harvesting. *IEEE Transactions on Wireless Communications, 18*(6), 3192–3207, https://doi.org/10.1109/TWC.2019.2911939
4. Zhang, X. & Duan, L. (2019). Fast deployment of UAV networks for optimal wireless coverage. *IEEE Transactions on Mobile Computing, 18*(3), 588–601. https://doi.org/10.1109/TMC.2018.2840143
5. Grace, D., Capstick, M. H., Mohoric, M., Horwath, J., Pallavicini, M. B., & Fitch, M. (2005). Integrating users into the wider broadband network via high altitude platforms. *IEEE Wireless Communication, 12*(5), 98–105. https://doi.org/10.1109/MWC.2005.1522112
6. Chandrasekharan, S., Gomez, K., Al-Hourani, A., Kandeepan, S., Rasheed, T., Goratti, L., Reynaud, L., Grace, D., et al. (2016). Designing and implementing future aerial communication networks. *IEEE Communications Magazine, 54*(5), 26–34. https://doi.org/10.1109/MCOM.2016.7470932
7. Mondal, A. & Hossain, A. (2020). Channel characterization and performance analysis of unmanned aerial vehicle-operated communication system with multiphop radio frequency-free space optical link in dynamic environment. *International Journal of communication systems, 33*(16), e4568. https://doi.org/10.1002/dac.4568

8. Larsen, E., Landmark, L. & Kure, O. (2017). Optimal UAV relay positions in multi-rate networks. *In Proceeding of Wireless Days, Porto, Portugal*, 8–14. https://doi.org/10.1109/WD.2017.7918107

9. Han, Z., Swindlehurst, A. L., & Liu, K. J. R. (2009). Optimization of MANET connectivity via smart deployment/movement of unmanned air vehicles. *IEEE Transactions on Vehicular Technology, 58*(7), 3533–3546. https://doi.org/10.1109/TVT.2009.2015953

10. Jiang, F. & Swindlehurst, A. L. (2010). Dynamic UAV relay positioning for the ground-to-air uplink. *In Proceeding IEEE Globecom Workshop*, 1766-1770. Miami, FL, USA. https://doi.org/10.1109/GLOCOMW.2010.5700245

11. Zhan, P., Yu, K., & Swindlehurst, A. L. (2011). Wireless relay communications with unmanned aerial vehicles: performance and optimization. *IEEE Transactions on Aerospace and Electronic Systems, 47*(3), 2068–2085. https://doi.org/10.1109/TAES.2011.5937283

12. Zeng, Y., Zhang, R., & Lim, T. J. (2016). Throughput maximization for UAV-enabled mobile relaying systems. *IEEE Transactions on Communications, 64*(12), 4983–4996. https://doi.org/10.1109/TCOMM.2016.2611512

13. Ono, F., Ochiai, H. & Miura, R. (2016). A wireless relay network based on unmanned aircraft system with rate optimization. *IEEE Transactions on Wireless Communications, 15*(11), 7699–7708. https://doi.org/10.1109/TWC.2016.2606388

14. Chen, Y., Zhao, N., Ding, Z. & Alouini, M. S. (2018). Multiple UAVs as relays: multi-hop single link versus multiple dual-hop links. *IEEE Transactions on Wireless Communications, 17*(9), 6348–6359. https://doi.org/10.1109/TWC.2018.2859394

15. Zhang, C. & Zhang, W. (2017). Spectrum sharing for drone networks. *IEEE Journal on Selected Areas in Communications, 35*(1), 136–144. https://doi.org/10.1109/JSAC.2016.2633040

16. Fan, R., Cui, J., Jin, S., Yang, K., & An, J. (2018). Optimal node placement and resource allocation for UAV relaying network. *IEEE Communications Letters, 22*(4), 808–811. https://doi.org/10.1109/LCOMM.2018.2800737

17. Zhan, C., Zeng, Y., & Zhang, R. (2018). Energy-efficient data collection in UAV enabled wireless sensor network. *IEEE Wireless Communications Letters, 7*(3), 328–331. https://doi.org/10.1109/LWC.2017.2776922

18. Sallouha, H., Azari, M. M., Chiumento, A., & Pollin, S. (2018). Aerial anchors positioning for reliable RSS-based outdoor localization in urban environments. *IEEE Wireless Communications Letters, 7*(3), 376–379. https://doi.org/10.1109/LWC.2017.2778723

19. Liu, X., Liu, Y., & Chen, Y. (2019). Reinforcement learning in multiple UAV networks: deployment and movement design. *IEEE Transactions on Vehicular Technology, 68*(8), 8036–8049. https://doi.org/10.1109/TVT.2019.2922849

20. Cui, J., Liu, Y., & Nallanathan, A. (2020). Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Transactions on Wireless Communications, 19*(2), 729–743. https://doi.org/10.1109/TWC.2019.2935201

21. Wang, L., Wang, K., Pan, C., Xu, W., Aslam, N., & Nallanathan, A. (2021). Deep reinforcement learning based dynamic trajectory control for UAV-assisted mobile edge computing. *IEEE Transactions on Mobile Computing (Early Access)*. https://doi.org/10.1109/TMC.2021.3059691

22. Zhong, X., Guo, Y., Li, N., & Chen, Y. (2020). Joint optimization of relay deployment, channel allocation, and relay assignment for UAVs-aided D2D networks. *IEEE/ACM Transactions on Networking, 28*(2), 804–817. https://doi.org/10.1109/TNET.2020.2970744

23. Al-Hourani, A., Kandeepan, S., & Lardner, S. (2014). Optimal LAP altitude for maximum coverage. *IEEE Wireless Communications Letters, 3*(6), 569–572. https://doi.org/10.1109/LWC.2014.2342736

24. Chen, Z., Lin, T. & Wu, C. (2016). Decentralized learning-based relay assignment for cooperative communications. *IEEE Transactions on Vehicular Technology, 65*(2), 813–826. https://doi.org/10.1109/TVT.2015.2399653

25. Chincoli, M. & Liotta, A. (2018). Self-learning power control in wireless sensor networks. *Sensors 18*(2), 375. https://doi.org/10.3390/s18020375

26. Chen, Y., Zhao, N., Ding, Z. & Alouini, M. S. (2018). Multiple UAVs as relays: multi-hop single link versus multiple dual-hop links. *IEEE Transactions on Wireless Communications, 17*(9), 6348–6359. https://doi.org/10.1109/TWC.2018.2859394.
27. Aslam, N., Xia, K. & Hadi, M. U. (2019). Optimal wireless charging inclusive of intellectual routing based on SARSA learning in renewable wireless sensor networks. *IEEE Sensors Journal, 19*(18), 8340–8351. https://doi.org/10.1109/JSEN.2019.2918865

28. Wu, K., Tellambura, C. & Jiang, H. (2017). Optimal transmission policy in energy harvesting wireless communications: a learning approach. In *Proceeding of IEEE International Conference on Communications (ICC)* Paris, 1–6. https://doi.org/10.1109/ICC.2017.7997233

29. Ortiz, A., Al-Shatri, H., Li, X., Weber, T. & Klein, A. (2016). Reinforcement learning for energy harvesting point-to-point communications. In *Proceeding of IEEE International Conference on Communications (ICC)*, 1–6. https://doi.org/10.1109/ICC.2016.7511405

30. Mu, X., Zhao, X., Liang, H. (2020). Power allocation based on reinforcement learning for MIMO system with energy harvesting,” *IEEE Transactions on Vehicular Technology, 69*(7), 7622–7633. https://doi.org/10.1109/TVT.2020.2993275
Figures

Figure 1

System model of UAVs-supported self-organized device-to-device network.

Figure 2

Illustration of UAVs’ action space corresponding to each possible state.
Figure 3

Basic block diagram of reinforcement learning.
Figure 4

The variation of the total transmission rate of the USSD2D network corresponding to each episode

(a) Variations of the network throughput with the different number of UAVs.
(b) Variations of the network throughput with the different number of available orthogonal channels.
(c) Variations of the network throughput with the different number of UAV-assisted D2D pairs.
(d) Variations of the network throughput with the different number of direct D2D pairs.

Figure 5

Total instantaneous transmission rate of the USSD2D network with different network parameters.
Figure 6

Impact of various UAVs’ altitudes on network throughput.
Figure 7

Total instantaneous transmission rate of the USSD2D network with different learning parameters.