EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

User selection and dynamic power allocation in the SWIPT-NOMA relay system

Xianzhong Xie[1], Min Li[1*], Zhaoyuan Shi[1], Hong Tang[2] and Qian Huang[1]

*Correspondence: lxmsxqs@163.com
[1] Chongqing Key Laboratory of Computer Network and Communication Technology, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
Full list of author information is available at the end of the article

Abstract

Non-orthogonal multiple access (NOMA) technology provides an effective solution to massive access with a high data rate demand in new-generation mobile networks. The paper combinations with NOMA and simultaneous wireless information and power transfer (SWIPT) relay to maximize the sum rate in the downlink system. To that end, it is critical how to select effectively users access system and power allocation for the access user. This paper proposes a user selection and dynamic power allocation (USDPA) scheme in the NOMA-SWIPT relay system based on neural network because traditional optimization methods have difficulty solving nonlinear and non-convex problems. We establish a user selection network utilizing a deep neural network (DNN) and propose a power allocation network using deep reinforcement learning. The simulation results show that the proposed scheme achieves better performance than other related schemes, especially for high quality of service requirements.

Keywords: Non-orthogonal multiple access (NOMA), Simultaneous wireless information and power transfer (SWIPT), User selection and power allocation, Deep neural network (DNN), Deep reinforcement learning (DRL)

1 Introduction

Due to the massive amount of wireless equipment accessed via the internet, researchers have focused on the high demand for charging wireless mobile terminals (MTs). Thus, as a promising green communication solution, simultaneous wireless information and power transfer (SWIPT) was introduced to increase the battery life. SWIPT can achieve significant gains in energy consumption and spectrum efficiency (SE), improve interference management, and reduce transmission delays by enabling the simultaneous transmission of power and information [1]. Two practical receiving methods exist for the SWIPT strategy, i.e., time switching (TS) and power splitting (PS), to harvest energy and decode information. In addition, a cooperative relaying (CoR) method combined with SWIPT was proposed to increase network reliability and expand the signal coverage area [2, 3]. The non-orthogonal multiple access (NOMA) scheme has been regarded as a promising technique to improve the SE for 5G and future communication systems because the signals of different MTs in NOMA can be multiplexed on the same resource elements [4]. Therefore, some researchers combined NOMA with SWIPT relay technology to improve the SE and achieve green communication [4, 5].

Numerous studies were conducted on wireless resource management to improve the performance of the SWIPT-NOMA relay system [6–11]. Reference [6] utilized an average power allocation scheme in the downlink and fixed power control in the uplink to evaluate the ergodic rate; however, the strategy does not guarantee that all signals are successfully decoded in the downlink and uplink. Reference [7] compared a cognitive radio-inspired power allocation scheme with a fixed power allocation scheme to ensure the fairness of the data rate. Reference [8] analyzed the error probability of the SWIPT-NOMA system by using a fixed allocation power scheme. In [9], the outage probability was regarded as the optimization function to obtain the power allocation factors. The analysis in [10] was in line with realistic scenarios regarding the impact of imperfect channel state information (ICSI) and residual hardware impairments (RHIs). Reference [11] evaluated the performance of a complex SWIPT scenario that allocated fixed power in the downlink. However, most papers focused on fixed power allocation, whereas artificial intelligence-based (AI) schemes for wireless resource allocation have not been well researched.

Many studies investigated access schemes in SWIPT-NOMA relay systems [12–14]. Reference [12] analyzed a SWIPT-NOMA relay system that considered channel estimation errors (CEEs) and RHIs. When all users accessed the system, more interference occurred at the receivers. Reference [13] investigated the outage probability choosing an optimum near destination node and an optimum far destination node, and the near node was used as the relay. Nevertheless, neither [12] nor [13] considered the channel gain from the relay to the MT, causing performance degradation. [7–11] considered access to all users, which prevented the decoding of all signals and generated more interference at the receivers. Furthermore, the performance of an all-access users' scheme is not high [15, 16] despite high model complexity. In contrast to [7–14] proposed access to users who have fed back channel state information (CSI), the algorithm had difficulty converging.

Therefore, it is imperative to develop a scheme that provides user access and allocates power to qualified users. AI techniques can extract valuable information from data to learn and support different functions for optimization, prediction, and decision-making in mobile edge computing, mobility prediction, optimal handover solutions, and spectrum management [17]. Deep reinforcement learning (DRL) can solve real-time and dynamic decision-making problems for power allocation [18–20]. Reference [18] proposed a deep Q network (DQN) for each MT to obtain the optimal power allocation scheme. The objective was to reduce the size of the state space; however, this distributed power allocation method has no information interaction between the MTs, resulting in power allocation conflicts. Reference [19] proposed a two-step model-free DRL-based power control scheme to maximize the long-term sum energy efficiency (EE). Based on a multi-carrier NOMA network with SWIPT, reference [20] proposed to use a deep belief network (DBN) to approximate the optimal power allocation.

1.1 Contribution

To deal with problems of traditional methods [7–16], inspired by the above studies [17–20], we propose a combined user selection and dynamic power allocation (USDPA) scheme that chooses the best users access the system and decides optimal power allocation to maximize the sum rate. The main advantages and contributions of this paper are summarized as follows.

- The USDPA scheme is proposed in the SWIPT-NOMA relay system to optimize the user access and power allocation simultaneously to maximize the sum rate because traditional optimization methods have difficulty solving nonlinear and non-convex problems. More importantly, the results show that our algorithm can successfully access more users than comparable algorithms.
- We use a deep neural network (DNN) for the user selection network to generate the access decision. Subsequently, the access decision is mapped to several candidate access actions, whose number changes adaptively. In addition, the result displays that the model converges quickly without adding additional computational complexity.
- We utilize a DQN to generate the optimal power allocation for each candidate access action. Afterwards, we use the optimal pair of access action and power allocation action with the maximum sum rate in the system. The best power allocation action is stored in the replay memory to train this network.
- Finally, we compare the performance of the USDPA with other schemes. The simulations under different scenarios show that the proposed algorithm improves quality of service (QoS) and can achieve better performance than other related schemes.

1.2 Organization

The remainder of this paper is organized as follows. Section 2 describes the system model and the problem formulation of the user selection and power allocation model for the SWIPT-NOMA relay system. Section 3 presents the USDPA scheme, including the user selection network and power allocation network. Section 4 presents the experimental results and analysis, including the convergence, the sum rate, and the number of successful communication users (NSCUs). Finally, the conclusions are summarized in Sect. 5.

2 System model and problem formulation

2.1 Problem formulation

We consider a system model that includes a base station (BS), a relay employing a decode-and-forward (DF) protocol, and N destinations, as illustrated in Fig. 1. Hereafter, subscripts S, R and D_i will be used for the BS, relay, and destination i, respectively. The radius of sector S_1 are Υ_{S_1} with the BS at the center and an angle ϕ. The radius of sector S_2 are Υ_{S_2} with same center and angle as sector S_1. The relay is located on a circular arc of radius Υ_{S_1}, and the N destinations are randomly and uniformly distributed in the region between Υ_{S_1} and Υ_{S_2}. Each destination node and the relay have a single antenna operating in half-duplex (HD) mode. We assume that all small-scale fading in the system is independent and identically distributed Rayleigh fading occurs. The channel coefficients of the links from the BS to the relay and the relay to D_i are $h_{S,R} \sim \mathcal{CN}\left(0, d_{S,R}^{-\tau}\right)$ and $h_{R,D_i} \sim \mathcal{CN}\left(0, d_{R,D_i}^{-\tau}\right)$, respectively, where $d_{i,j}$ denotes the distance between node i and node j, and τ is the path-loss exponent.

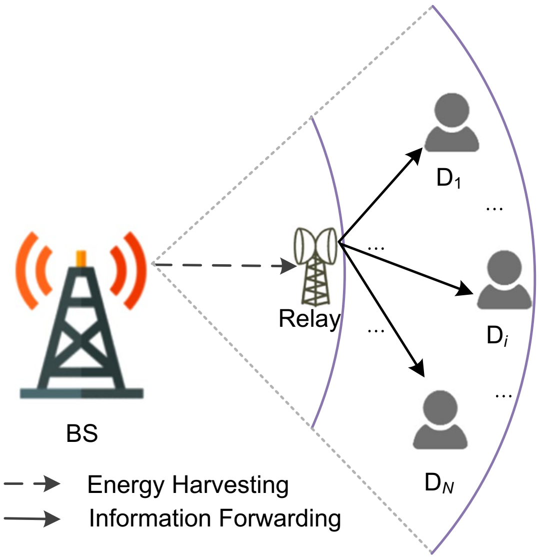


Fig. 1 The SWIPT-NOMA relay model. The BS sends signal to the relay, and then, the relay harvests energy to forward signals to the qualified destinations with NOMA

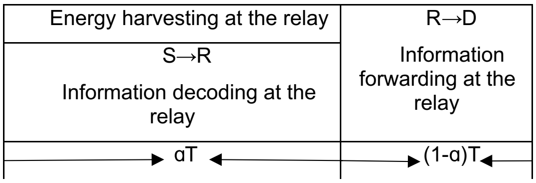

Energy harvesting at the relay
S→R
Information decoding at the relay
R→D
Information forwarding at the relay
αT
(1-α)T

Fig. 2 The PSR strategy. The relay uses the PSR strategy to harvest energy and forward the superimposed signals to the qualified users

For simplicity, we assume that the transmission time $T = 1$ and the bandwidth $B = 1$. The power splitting relay (PSR) strategy is used (Fig. 2). Within the duration of each αT, the relay performs energy harvesting (EH) and information decoding (ID); within each $(1 - \alpha)T$ period, the relay performs information forwarding (IF) in the NOMA mode. ρ is the power splitting factor for harvesting energy, and $(1 - \rho)$ is for decoding information. At the end of αT of each slot, the relay receives the signal from the BS, which can be expressed as:

$$y_R = \sqrt{P_S} h_{S,R} x_S + n_R, \tag{1}$$

where x_S is the signal transmitted by the BS to the relay. $n_R \sim \mathcal{CN}\left(0, \sigma_R^2\right)$ is the additive white Gaussian noise.

The energy harvested by the relay is defined as follows:

$$E_{EH} = \eta \rho P_S |h_{S,R}|^2 \alpha T, \tag{2}$$

where η is the energy conversion efficiency factor. The remaining battery power of the relay is $B_r(t)$ at the beginning of each slot and $B_r(0){=}0$ in the first time slot. We assume that the harvested energy is much less than the maximum storage capacity of the relay. After αT of each slot, the total energy of the battery is:

$$B(t) = B_r(t-1) + E_{EH}(t), \tag{3}$$

where $B_r(t-1)$ is the remaining energy of the previous time slot. The relay decodes the received signal and forwards the superimposed signal through NOMA. In each $(1-\alpha)T$, the maximum transmitting power of the relay can be expressed as:

$$P_R = \frac{B(t)}{(1-\alpha)T}. \tag{4}$$

We assume that D_j belongs to a set $\mathbb{C}$ that includes the qualified access users, where $|\mathbb{C}| = \varpi$. The received signals from the relay can be defined as:

$$y_{D_j} = h_{R,D_j}\left(\sum_{j=1}^{\varpi}\sqrt{P_R\lambda_j x_j}\right) + n_D, \tag{5}$$

where λ_j is the power factor allocated by the relay to signal x_j, and the power allocation factor λ_j, $n_D \sim \mathcal{CN}\left(0, \sigma_D^2\right)$ is the additive white Gaussian noise.

The expression of the remaining energy of the battery at the relay after each time slot can be expressed as:

$$B_r(t) = (1-\alpha)TP_R\left(1 - \sum_{j=1}^{\varpi}\lambda_j\right). \tag{6}$$

We implement successive interference cancellation (SIC) based on the power ranking from strong to weak. If the j-th user is able to eliminate the signals of weaker users, the signal-to-interference-plus-noise ratio (SINR) for decoding its own signal is:

$$SINR_{R,D_j} = \begin{cases} \frac{\rho_R|h_{R,D_j}|^2\lambda_j}{\sum_{j=j+1}^{\varpi}\rho_R|h_{R,D_j}|^2\lambda_j+1}, & j = 1, \ldots, \varpi - 1 \\ \rho_R|h_{R,D_j}|^2\lambda_j, j = \varpi, \end{cases} \tag{7}$$

where $\rho_R = P_R/\sigma_D^2$. The achievable data rate at the D_j is defined as follows:

$$C_{D_j} = (1-\alpha)\log_2\left(1 + SINR_{R,D_j}\right). \tag{8}$$

The sum rate of this system is as follows:

$$C_{sum} = \sum_{j=1}^{\varpi} C_{D_j}. \tag{9}$$

2.2 Problem formulation

We consider the maximum sum rate of the SWIPT-NOMA relay system; thus, the optimization problem is expressed as:

$$\begin{aligned} &\max \sum_{j=1}^{\varpi} C_{D_j}, \\ s.t.\ &C1 : D_j \in \mathbb{C}, j = 1, \ldots, \varpi, \\ &C2 : C_{D_j} \geq R_{th}, \\ &C3 : \sum_{j=1}^{\varpi} \lambda_j \leq 1, \lambda_j \in \Lambda, \end{aligned} \tag{10}$$

where the set $\mathbb{C}$ includes the qualified access users; R_{th} and Λ are data rate threshold for the D_j to decode the signal successfully and the set of power allocation factors, respectively.

Constraint $C1$ represents that each access user belongs to the qualified set $\mathbb{C}$; constraint $C2$ represents the minimum quality of service (QoS) requirements for selected access users, where the data rate of each qualified access user needs to be larger than the rate threshold; constraint $C3$ states that the power cannot be larger than the transmission power of the relay.

A user selection network is established to reduce the interference caused by the access of all users. In addition, since power allocation adjustment is inefficient, we propose a DQN algorithm to solve this problem.

3 USDPA scheme

In this section, we describe the USDPA scheme to determine user access and power allocation. The USDPA algorithm for the downlink SWIPT-NOMA relay system is presented in Fig. 3. We first determine the user access based on the user selection network and subsequently derive the power allocation based on the DQN.

The relay forwards the signals to the users with actions of the user selection network and the power allocation network. By obtaining optimized the user access and power allocation of the system, we maximize the sum rate. The USDPA algorithm is shown in Algorithm 1.

3.1 User selection network

In this part, we design an access policy that rapidly generates an access decision $Y(t)$.

$$\pi_x(t) : h_{R,D}(t) \rightarrow Y(t), \tag{11}$$

where $Y(t)$ represents the output of the user selection network.

3.1.1 User selection algorithm

The user selection network has an embedded parameter $\omega_1(t)$ that connects the hidden neurons. At the beginning of each slot, the user selection network uses $h_{R,D}(t)$ as the input and outputs a relaxed user access action $Y(t)$ with N dimensions according to the access policy $\pi_x(t)$ and the parameterized $\omega_1(t)$. Since each value in $Y(t)$ is between 0 and 1, it is difficult to determine who should access the system; thus, we design a mapping rule to quantize the output $Y(t)$. According to this rule, $Y(t)$ is mapped into W access vectors, and the one with the maximum sum rate is the best access vector $q^*(t)$.

A four-layer DNN is designed with one input layer, two hidden layers, and one output layer. The dimensions of the input $h_{R,D}(t)$ and output $Y(t)$ are N, which denotes the number of destination nodes. The two hidden layers' activation function is a Relu function, and

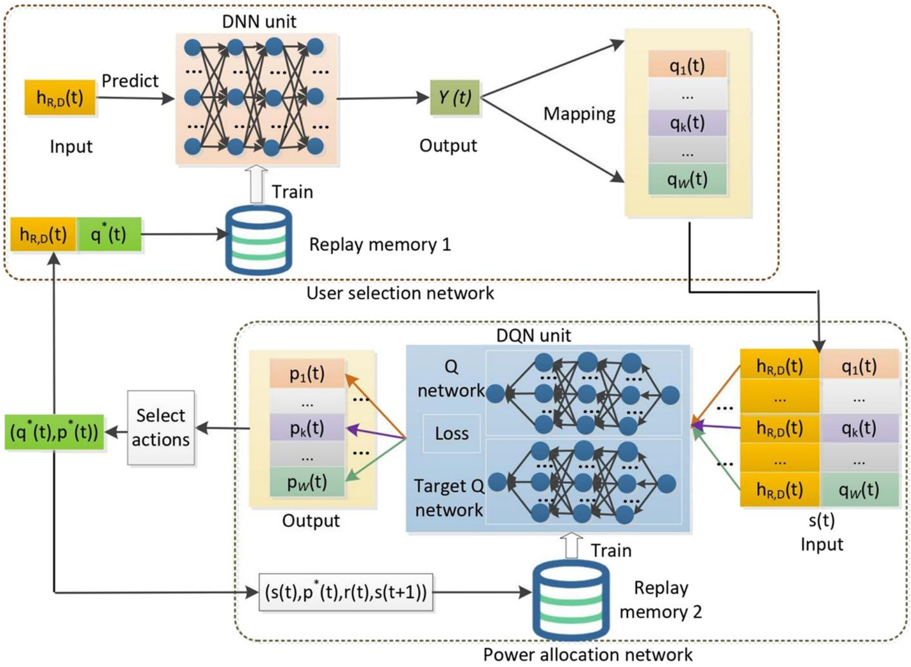


Fig. 3 The USDPA algorithm. The proposed scheme includes user selection network to generate the access decision and power allocation network to generate power allocation decision effectively according to the access decision

the output layer uses a Sigmoid activation function. In the t-th slot. The output of the user selection network can be expressed as $Y(t) = f_{\omega_1}\left(h_{R,D}(t)\right)$. The user selection algorithm is shown in Algorithm 2.

3.1.2 The mapping rule

The output $Y(t)$ of the user selection network is mapped to W vectors. Each value of the vector is either 0 or 1, which 0 means the user is not accessed and otherwise. It should be noted that there are 2^N cases for vectors; consequently, $W \in \left[1, 2^N\right]$, where its initial value is the same as N. Reference [21] proved the effectiveness of this method using the same binary representation in edge computing to evaluate the output of the DNN. The detailed mapping rules are as follows:

(1) q_1 accounts for the first mapping vector of $Y(t)$ and is obtained by comparing $Y(t)$ with 0.5.

$$q_1[j] = \begin{cases} 1, & Y[j] > 0.5 \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where $j = 1, \ldots, N$.

(2) The new sequence $Y^*(t)$ is obtained by sorting $Y(t)$ according to the absolute value of the difference between $Y(t)$ and 0.5.

(3) The values of the remaining W-1 mapping vectors are related to $Y^*(t)$, and the vector of the i-th mapping is as follows:

$$q_i[j] = \begin{cases} 1, Y[j] > Y^*[l] \\ 0, Y[j] < Y^*[l], \end{cases} \tag{13}$$

where $l = 2, \ldots, W - 1$ and $j = 1, \ldots, N$. Specifically, when $Y[j] = Y^*[l]$,

$$q_i[j] = \begin{cases} 1, \ 0.5 > Y^*[l] \\ 0, \ \text{otherwise.} \end{cases} \tag{14}$$

Algorithm 1: USDPA algorithm

```
Initialize the parameters of the user selection network and empty replay memory 1;
Initialize the parameters of the power allocation network and empty replay memory 2;
Set the training interval $\Delta_1$ and $\Delta_2$ of the user selection network and the power allocation network, respectively;
for $t = 1, \ldots, L$ do
  Obtain the selected user actions $Y(t) = f_{\omega_1}(h_{S,R}(t))$;
  Map $Y(t)$ into the $W$ user access vectors;
  for $k = 1, \ldots, W$ do
    Initialize the environment and obtain initial observation;
    if random possibility $\leq \varepsilon$ then
      Select the power allocation factor randomly;
    else
      According to $a(t) = \arg\max Q(s(t), a(t); \omega_2)$ obtain the power allocation $p_k(t)$;
    end if
  Obtain the immediate reward according to Eq. (16);
  Obtain the next state of the power allocation network;
  Compute the sum rate by using the action $(p_k(t), q_k(t))$;
  Select the best access actions pair $(q^*(t), p^*(t))$ with the max sum rate;
  end for
  Update replay memory 1 by adding $(h_{S,R}(t), q^*(t))$;
  if $t\%\Delta_1 = 0$ then
    Uniformly sample a batch of the data set from the replay memory 1;
    Train the user selection network with the batch data;
    Update the parameters of each layer of the user selection network;
  end if
  Store the pair $(s(t), p^*(t), r(t), s(t+1))$ corresponding to $p^*(t)$ in the replay memory 2;
  Sample a random batch from replay memory 2;
  Update the parameters of the power allocation network by minimizing the $Loss(\omega_2)$ according to Eq. (25);
  if $\zeta \% 32 = 0$ then
    Update $W^*$ according to min(max($W(t-1), \cdots, W(t - \zeta)$) + 1, $N$);
  end if
end for
```

After each ζ slot, $W^* = \min(\max(W(t-1), \ldots, W(t-\zeta)) + 1, N)$, where $W(t-\zeta)$ is the position of the best user selection vector corresponding to slot $(t-\zeta)$ of W vectors.

Algorithm 2: The user selection algorithm to determine the users access decision

```
Initialize the parameters of each layer of this network and empty replay memory 1;
Set the training interval $\Delta_1$;
Obtain selected user actions $Y(t) = f_{\omega_1}(h_{S,R}(t))$;
Map $Y(t)$ into $W$ user access vectors;
for $k = 1, \ldots, W$ do
    Use the pair $(h_{S,R}(t), q_k(t))$ in the power allocation network which outputs the power allocation action $p_k(t)$;
    Compute the sum rate by using the action $(p_k(t), q_k(t))$;
end for
Select the best access action $q^*(t)$ with the max sum rate;
Update replay memory 1 by adding $(h_{S,R}(t), q^*(t))$;
if $t\%\Delta_1 = 0$ then
    Uniformly sample a batch of the data set from the replay memory 1;
    Train the user selection network with the batch data;
    Update the parameters of the user selection network;
end if
```

3.1.3 The training of the user selection network

To maximize the sum rate, $h_{S,R}(t)$ and each access vector $q_k(t)$ are the inputs of the power allocation network, which outputs the power allocation $p_k(t)$. Then, the sum rate is calculated using each action pair $(q_k(t), p_k(t))$ where $k = 1, \ldots, W$. The system selects the best access action $q^*(t)$ and adds the newly obtained pair $(h_{S,R}(t), q^*(t))$ to the replay memory 1 for training, and $q^*(t)$ is used as labels. Subsequently, a batch of training samples $\Omega_1(t)$ are from the replay memory 1 to train the user selection network and the parameters $\omega_1(t)$ and the policy $\pi_x(t)$ are updated. The $\omega_1(t)$ is updated by reducing the loss function of the user selection network every Δ_1 slots as follows:

$$\begin{aligned} Loss(\omega_1) = &-\frac{1}{|\Omega_1(t)|}\sum_{u\in\Omega_1(t)} ((q^*(u)) \log f_{\omega1(u)}(h_{R,D}(u)) \\ &+ (1-q^*(u)) \log(1-f_{\omega1(u)}(h_{S,D}(u))) \end{aligned} \tag{15}$$

The Adam optimizer is utilized in the training process with learning rate θ_1. After training, the user selection policy $\pi_x(t)^*$ can be updated.

3.2 Power allocation algorithm

Next, we obtain the appropriate allocation action using the DQN; the algorithm is shown in Algorithm 3. We first provide some background information on reinforcement learning (RL) to clarify the algorithm. The key elements of RL are defined as follows:

State space: The state space is defined as $s = \{[h_{R,D}(t), q_1(t)], \ldots, [h_{R,D}(t), q_w(t)]\}$.

Action space: $a = \{a^1, \ldots, a^z\}$ is defined for its power allocation action space where $z = A_M^N$. There are M power allocation factors, and the action space for N destinations has A_M^N actions.

Reward: We use the NSCUs, whose data rate is no less than the QoS threshold to obtain an immediate reward, which is defined as follows:

$$r(t) = \begin{cases} 10,\ \text{NSCUs} = \varpi_k(t), \\ 9,\ \text{NSCUs} = \varpi_k(t) - 1 \text{ and } \varpi_k(t) > 1, \\ 8,\ \text{NSCUs} = \varpi_k(t) - 2 \text{ and } \varpi_k(t) > 2, \\ 7,\ \text{NSCUs} = \varpi_k(t) - 3 \text{ and } \varpi_k(t) > 3, \\ 0, \text{otherwise,} \end{cases} \tag{16}$$

where $\varpi_k(t)$ is the number of qualified users accessing the system in the k-th access vector. Moreover, the cumulative reward function of the power allocation network is defined as follows:

$$R(t) = \sum_{t=1}^{L} r(t)\gamma^{t-1}, \tag{17}$$

where γ is the discount factor of the reward during L slots.

Transition probability: $\mathcal{P}$ represents the transition probability, i.e., the probability to transition from state $s(t)$ to the next state $s(t+1)$, given the action $a(t)$ executed in the state $s(t)$.

The Q value function is instrumental in solving RL problems [22]. The function describes the expected cumulative reward $R(t)$ of initial $s(t)$, performing action $a(t)$, and following policy $\pi_r(t)$. To obtain the appropriate power allocation action, the Q value function is defined as:

$$Q^{\pi_r(t)}(s(t), a(t)) = E^{\pi_r(t)}[r(t) + \gamma Q^{\pi_r(t)}(s(t+1), a(t+1))|s(t), a(t)] \tag{18}$$

The optimal action-value function in Eq. (18) is equal to the Bellman optimality equation [22], which is expressed as follows:

$$Q^*(s(t), a(t)) = E[r(t) + \gamma \max Q^*(s(t+1), a(t+1))|s(t), a(t)]. \tag{19}$$

After the optimal Q-function $Q^*(s(t), a(t))$ is obtained, the Q-learning policy is determined by:

$$\pi_r(t)(s(t), a(t)) = \arg\max Q^*(s(t), a(t)) \tag{20}$$

The state-value function is obtained as follows:

$$V(s(t)) = \max Q(s(t), a(t)). \tag{21}$$

The Q-value is defined as follows:

$$Q_{t+1}(s(t), a(t)) = (1 - \theta_2(t))Q_t(s(t), a(t)) + \theta_2(t)(r(t) + \gamma V_t(s(t+1)) \tag{22}$$

where $\theta_2(t)$ is the learning rate of the power allocation network.

In general, the Q learning algorithm adopts the $\varepsilon - greedy$ policy to select the power allocation action $a(t)$ with probability $1 - \varepsilon$, whereas a random action has a probability of $\varepsilon = 0.8$. The power allocation action is generated by:

$$a_1(t) = \arg\max Q(s(t), a(t); \omega_2), \tag{23}$$

where ω_2 is the parameter of the power allocation network.

3.2.1 Power allocation algorithm based on the DQN

Nevertheless, the Bellman equation is difficult to obtain because it is nonlinear and does not have a closed-form solution. The solution to this problem is to utilize neural networks to estimate the Q value. Therefore, we adopt a DQN to establish the power allocation network with a DNN to output the estimated Q value.

We design a power allocation policy $\pi_r(t)$ that quickly generates a power allocation decision corresponding to each access vector of the user selection network. The power allocation is implemented by the DQN, which is characterized by the embedded parameter $\omega_2(t)$ that connects the hidden neurons. After the output of the user selection network has been mapped to W access vectors $q(t)$, $h_{R,D}(t)$ combined with each access vector $q_k(t)$ is used as the input of the power allocation network. The output of this algorithm is $p_k(t)$ corresponding to each access vector $q_k(t)$. Then, we choose the actions $(q^*(t), p^*(t))$ with the maximum sum rate as the best actions and add the newly obtained pair $(h_{R,D}(t), p^*(t))$ to the replay memory 2. Subsequently, a batch of training samples Ω_2 from the replay memory 2 is used to train the power allocation network, and the parameters $\omega_2(t)$ and $\pi_r(t)$ are updated.

A five-layer power allocation network is designed, with one input layer, three hidden layers, and one output layer. The Relu function is used as the activation function in the first two hidden layers, and the tanh function is used in the last hidden layer. The output of the power allocation network can be expressed as $p_k(t) = f_{\omega_2}\left(h_{R,D}(t), q_k(t)\right)$.

After allocating power according to access vectors, the relay executes the optimal actions $(q^*(t), p^*(t))$ with the maximum sum rate of the system and receives the immediate reward $r(t)$. Subsequently, the system moves to the next state, and the replay memory 2 is used to store the tuple $(s(t), p^*(t), r(t), s(t+1))$ of each slot. When the replay memory 2 is full, the oldest record is removed, and the newest record is stored.

3.2.2 The training of power allocation algorithm

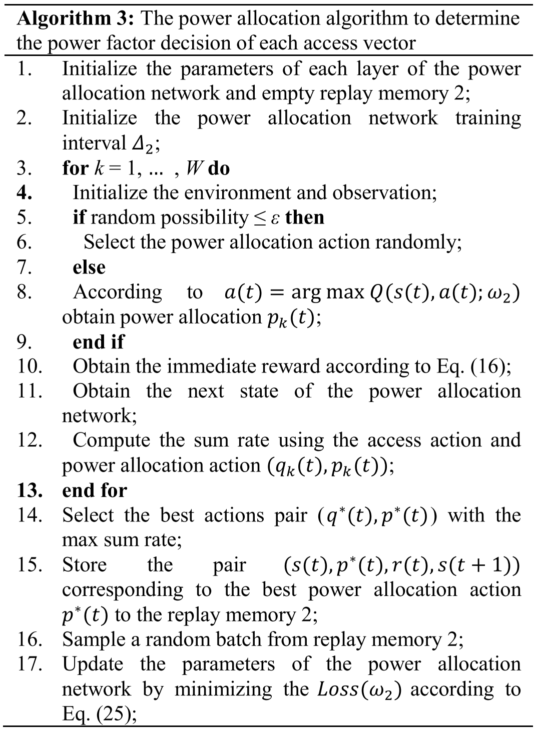
A batch of training samples $\Omega_2(t)$ from replay memory 2 is used to train the power allocation network. Then, the target Q value is obtained according to the target Q network as follows:

$$y_i = r(i) + \max Q(s(i+1), a_1(i+1); \overline{\omega}_2), \tag{24}$$

where $\overline{\omega}_2$ is the parameter of target Q network. The Q network is trained with $\Omega_2(t)$ by minimizing the loss function of power allocation network which is defined as:

$$Loss(\omega_2) = (y_i - Q(s(i), a_1(i); \omega_2))^2. \quad (25)$$

Meanwhile, the Adam optimizer is utilized in the training process with learning rate θ_2. We update the parameters $\overline{\omega}_2$ of the target Q network by copying the parameters of the Q network to each slot.



Algorithm 3: The power allocation algorithm to determine the power factor decision of each access vector

```
Initialize the parameters of each layer of the power allocation network and empty replay memory 2;
Initialize the power allocation network training interval $\Delta_2$;
for $k = 1, \ldots, W$ do
  Initialize the environment and observation;
  if random possibility $\leq \varepsilon$ then
    Select the power allocation action randomly;
  else
    According to $a(t) = \arg\max Q(s(t), a(t); \omega_2)$ obtain power allocation $p_k(t)$;
  end if
  Obtain the immediate reward according to Eq. (16);
  Obtain the next state of the power allocation network;
  Compute the sum rate using the access action and power allocation action $(q_k(t), p_k(t))$;
end for
Select the best actions pair $(q^*(t), p^*(t))$ with the max sum rate;
Store the pair $(s(t), p^*(t), r(t), s(t+1))$ corresponding to the best power allocation action $p^*(t)$ to the replay memory 2;
Sample a random batch from replay memory 2;
Update the parameters of the power allocation network by minimizing the $Loss(\omega_2)$ according to Eq. (25);
```

3.3 Complexity analysis

The complexity of the USDPA algorithm depends on the number of layers of the neural network and the number of neurons in each layer. The complexity of the user selection network is $M_1 \triangleq Nf_1 + f_1f_2 + f_2N$, where f_1 and f_2 are the numbers of neurons in the first and second hidden layers, respectively. The complexity of the power allocation network is $M_2 \triangleq Nf_{Q1} + f_{Q1}f_{Q2} + f_{Q2}f_{Q3} + f_{Q3}N$, where f_{Q1},f_{Q2} and f_{Q3} are the numbers of neurons in the first, second, and third hidden layers, respectively. In the USDPA algorithm, the output of the user selection network is mapped to W user access vectors; thus, the algorithm complexity is $O(M_1 + WM_2)$.

4 Results and discussion

In this section, the effectiveness of the proposed user selection and power allocation optimization scheme of the SWIPT-NOMA relay system is verified using the simulation. The effects of R_{th} and various levels of transmitting power at the BS on the performance

Table 1 Parameters values of the system

Parameter	Value
N	5
Υ_{S_1}	12 m
ϕ	90°
Υ_{S_2}	26 m
$d_{S,R}$	12 m
τ	2
α	0.5
ρ	0.7
σ_R^2, σ_D^2	-30 dbm
η	0.9
Λ	{0.1,0.15,0.2,0.25,0.3,0.35}
Δ_1	10
Δ_2	300
ζ	32
$\lvert\Omega_1\rvert$	100
θ_1	0.01
θ_2	10^{-5}
$\lvert\Omega_2\rvert$	32
f_1	120
f_2	80
f_{Q1}	256
f_{Q2}	256
f_{Q3}	512

of the SWIPT-NOMA relay system are analyzed to illustrate the superiority of the proposed scheme in increasing the sum rate.

In this paper, Tensorflow 2.0 is used for simulation. The simulation parameters are set as follows [23] (Table 1).

The sizes of replay memory 1 and replay memory 2 are 1000 and 400, respectively. The initial number of mapping vectors $W = N$.

4.1 Validation of training effects

In this part, we assess the performance of the proposed USDPA algorithm using simulations with different requirements for the successful decoding of the signals. Figure 4 shows the W of the USDPA algorithm versus the training slots when $P_S = 40$ dbm. The value of W converges quickly after 4000 slots, and the value of W is nearly stable within 2, indicating that the mapping scheme does not increase the computational complexity. It can be seen that the higher the QoS, the lower the value of W converges. The reason is that it is easier to satisfy the lower QoS. Figure 5 shows the loss of the USDPA algorithm with P_S=40 dBm and R_{th}=0.3 bits/s/Hz. It can be seen that the loss functions of the user selection network and the power allocation network converge quickly.

Figure 6 shows the average reward of the USDPA versus the training time slots with P_S=40 dBm for different QoS requirements. It can be observed that different QoS requirements have different effects on the performance of the USDPA algorithm. Specifically, the algorithm takes longer to reach convergence when the QoS requirements

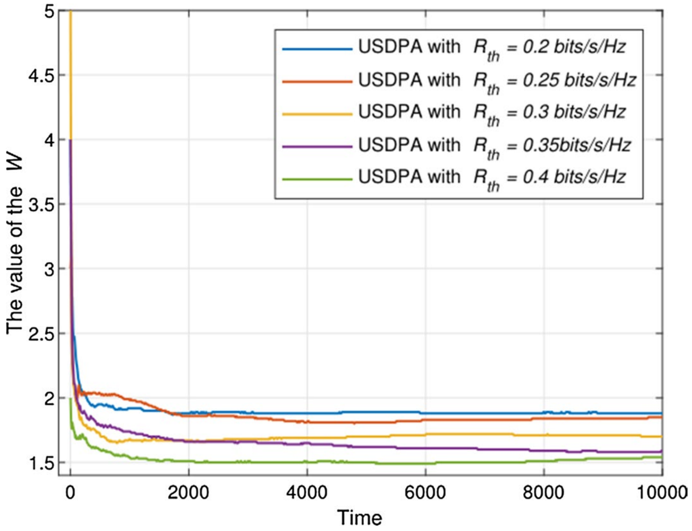


Fig. 4 The *W* comparison under various QoS requirements versus time when $P_S = 40$ dbm. The decreasing trend of *W* indicates that the complexity of the mapping rule is not increasing and can be converged quickly. The x-axis is the time slot, and the y-axis is the value of *W*. The illustrations are the USDPA with R_{th}=0.2 bits/s/Hz, R_{th}=0.25 bits/s/Hz, R_{th}=0.3 bits/s/Hz, R_{th}=0.35 bits/s/Hz, and R_{th}=0.4 bits/s/Hz

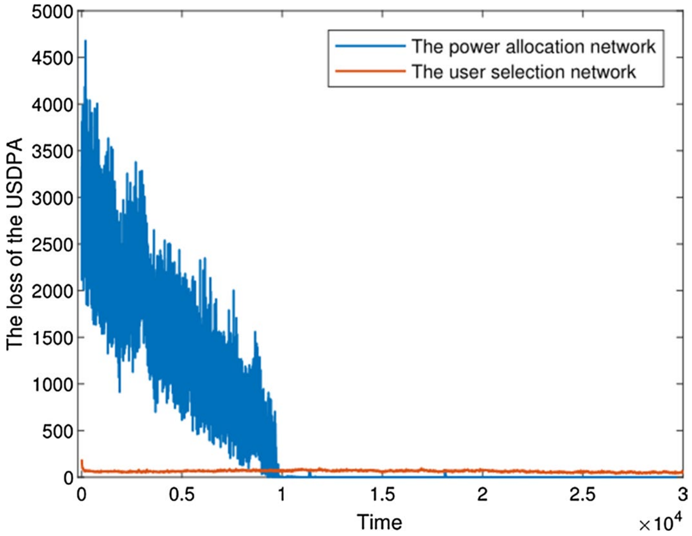


Fig. 5 The loss of USDPA algorithm with R_{th}=0.3 bits/s/Hz and $P_S = 40$ dBm versus time. After 1000 time slots, the user selection network can converge quickly. Therefore, the USDPA scheme converges quickly. The x-axis is the time slot, and the y-axis is the loss of the USDPA. The illustrations are the power allocation network and the user selection network

are high. In addition, when the QoS requirement is 0.3 bits/s/Hz, the loss of the allocation network converges rapidly after about 10,000 slots (Fig. 5), and the average reward converges rapidly to 10 (Fig. 6). Furthermore, the average reward after 2000 time slots is higher than 0.4 bits/s/Hz at a QoS requirement of 0.5 bit/s/Hz. The reason is that when the QoS requirement is 0.5 bits/s/Hz, the USDPA algorithm selects fewer users, causing less interference, and they can access the system more easily and successfully to meet

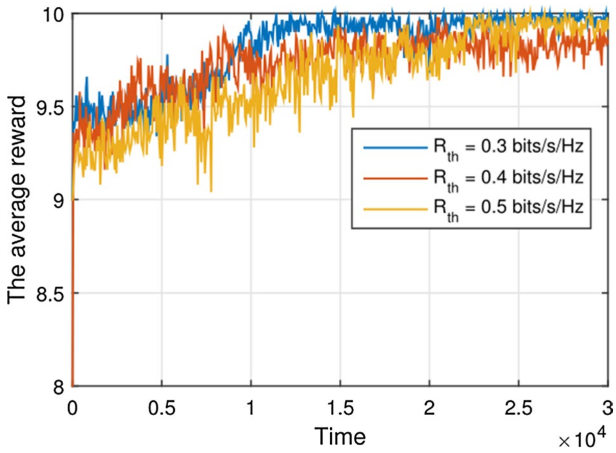


Fig. 6 The average reward of USDPA algorithm with $P_S=40$ dBm under different QoS requirements versus time. The reward function determines the speed and degree of convergence of the reinforcement learning algorithm. At low QoS, the reward function is able to converge quickly and vice versa. The x-axis is the time slot, and the y-axis is the average reward. The illustrations are $R_{th}=0.3$ bits/s/Hz, $R_{th}=0.4$ bits/s/Hz, and $R_{th}=0.5$ bits/s/Hz

the QoS requirement and allocate the appropriate power using the DQN. Therefore, the average rewards are relatively high. In general, the results indicate that the USDPA algorithm exhibits excellent learning performance for different QoS requirements.

4.2 Experimental results and discussion

The goal of this paper is to maximize the sum rate of the SWIPT-NOMA relay system. Consequently, the sum rate and the NSCUs are used to evaluate the algorithm's performance. Four algorithms are compared with the proposed algorithm: (1) All users access (AU) + DQN: all users access the system, and the power allocation scheme uses the DQN, which is the same as the power allocation scheme in [18]. (2) All users access + average power allocation (AU + AP): all users access the system, and the power of each user's signal is the average power factor. The algorithm decodes the signals using the order of channels from strong to weak. (3) The user selection scheme average power allocation (US + AP): the users that access the system are determined by the proposed user selection network, and the power of each user's signal is the average power factor. (4) Random user access (RU) + DQN: the users that access the system are determined randomly, and the DQN is used to allocate the power.

Figure 7 shows the NCMUs for different data thresholds. The NCMUs exhibits a decreasing trend for the USDPA, AU + DQN, US + DQN, AU + AP, and US + AP, when $P_S = 40$ dBm. The reason is that it is difficult for the system to allocate the appropriate power factor to enable the users to decode the signal successfully. AU + DQN shows the best NCMUs performance when the data thresholds are $R_{th}=0.2$ bits/s/Hz and $R_{th}=0.25$ bits/s/Hz. The reason is that AU + DQN is easier to satisfy the lower QoS requirements. The USDPA algorithm exhibits the optimum performance when $R_{th}=0.3$ bits/s/Hz, $R_{th}=0.35$ bits/s/Hz, and $R_{th}=0.4$ bits/s/Hz because the user selection network choose some users to access the system. The fewer the users accessing the system, the less interference there is. However, according to

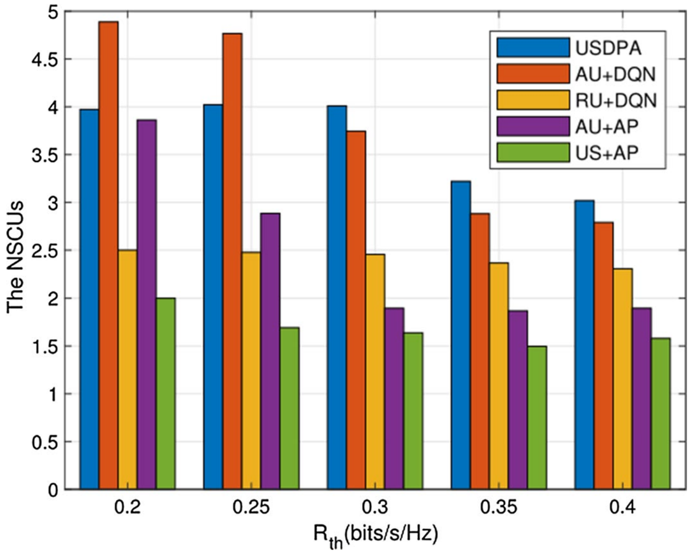


Fig. 7 The NSCUs comparison versus R_{th}. Under different QoS conditions, the more users are successfully decoded, the more effective users are accessed. Then the algorithm performance is better. The x-axis is the R_{th}, and the y-axis is the NSCUs. The illustrations are the USDPA, AU + DQN, RU + DQN, AU + AP, US + AP

the USDPA, it is not possible for the system to access only one user. The reason is that the power allocation factor is less than 1 and does not take 0 and 1, which means that the sum rate of one user access will not be the maximum. Therefore, the system always selects multiple users to access the system. What's more, it can be seen that the performance of AU + AP and US + AP schemes both converge when the QoS requirement is high. The reason is that both of them use the average power allocation factor for access users, which makes it difficult to guarantee that all qualified users can successfully decode the signal under the high QoS requirements. When R_{th} = 0.3 bits/s/Hz, the performance of the USDPA algorithm is 63.3%, 144.7%, 115%, and 7% higher than that of the RU + DQN, US + AP, AU + AP, and AU + DQN, respectively.

Figure 8 shows the average sum rate of the five schemes with $P_S = 40$ dBm. We can see that the performance of the USDPA is the best for all QoS requirements. When R_{th} = 0.3 bit/s/Hz, the average sum rate of the USPDA algorithm is 47.8%, 38.2%, 178%, and 63.1% higher than that of the AU + DQN, RU + DQN, AU + AP, and US + AP, respectively. The reason is that when the average power allocation is utilized for user access, there is no dynamic adjustment of the power allocation factor. More importantly, we observe that the average sum rate of the USDPA scheme is higher for a QoS requirement of 0.4 bits/s/Hz than a QoS requirement of 0.35 bits/s/Hz. The reason is that when the QoS requirements are higher, the USDPA algorithm selects fewer users to access the system. Thus, there is less interference at the receivers, and the achieved sum rate is higher. In addition, if all users access the system, there is more interference at the receivers, although the DQN is used to allocate power. The US + AP algorithm maintains stable performance as R_{th} increases because the user selection network chooses appropriate users to access the system. Although the RU + DQN algorithm chooses the users randomly, it still maintains a steady average sum rate because it uses the DQN algorithm to allocate power.

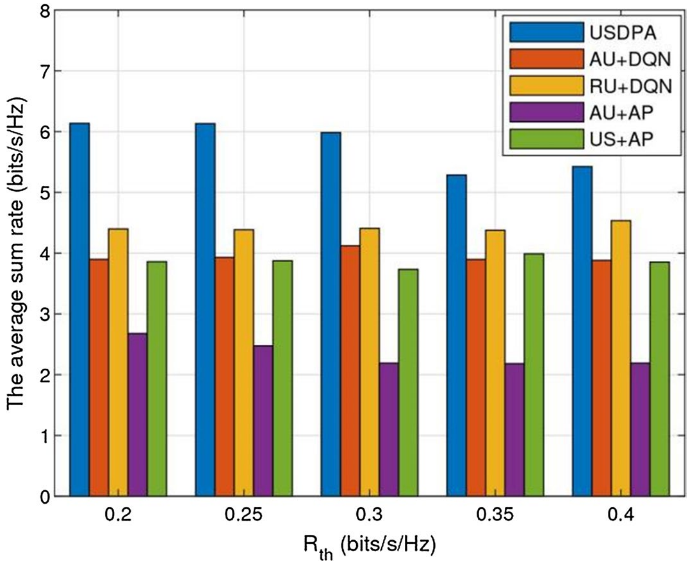


Fig. 8 The average sum rate comparison versus R_{th}. Different QoS can detect the stability of the algorithm. When a certain value is reached, the algorithm starts to converge. The x-axis is R_{th}, and the y-axis is the average sum rate. The illustrations are the USDPA, AU + DQN, RU + DQN, AU + AP, US + AP

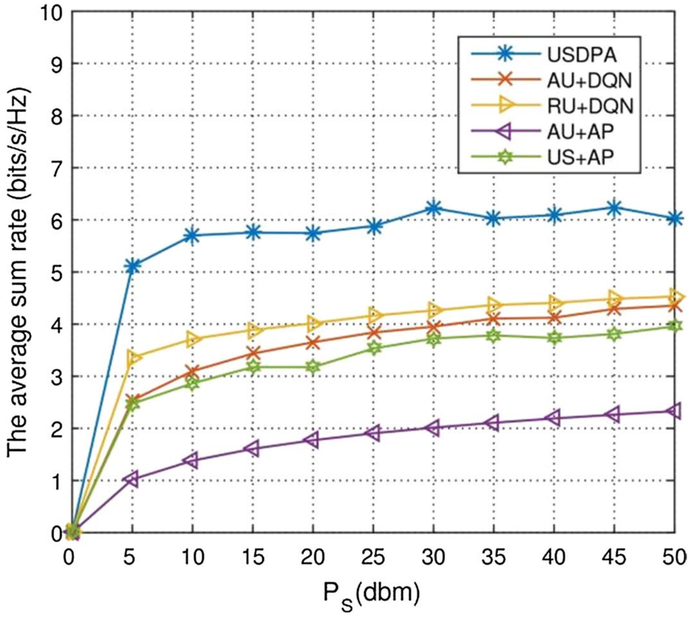


Fig. 9 The average sum rate comparison versus P_S. The x-axis is P_S, and the y-axis is the average sum rate. The illustrations are the USDPA, AU + DQN, RU + DQN, AU + AP, US + AP

Figure 9 displays the trend of the average sum rate of the five schemes with different levels of transmitting power at the BS, when R_{th} = 0.3 bits/s/Hz at the receivers. The average sum rate increases with increasing P_S. As P_S increases, the SINR at the accessed receivers improves, leading to a performance improvement. In addition, we find that the USDPA scheme outperforms the other four schemes. The proposed scheme jointly optimizes user access and power allocation, and the algorithm exhibits efficient learning ability by utilizing the user selection network and the power allocation network in the dynamic environment.

5 Conclusion

We propose a USDPA scheme in the SWIPT-NOMA relay system to maximize the sum rate in the downlink. A model of the SWIPT-NOMA relay system was established with a PSR scheme to harvest energy and forward signals. The USDPA was used to optimize the user access action and power allocation action simultaneously. The simulation results showed that the proposed scheme provided the best performance for increasing the sum rate. Due to the complexity of the problem, practical scenarios of multi-antenna configuration and a bidirectional relay will be analyzed in a future study.

Abbreviations

NOMA: Non-orthogonal multiple access; DF: Decode-and-forward; SWIPT: Simultaneous wireless information and power transfer; USDPA: User selection and dynamic power allocation; DNN: Deep neural network; DRL: Deep reinforcement learning; QoS: Quality of service; MTs: Mobile terminals; SE: Spectrum efficiency; TS: Time switching; PS: Power splitting; CoR: Cooperative relaying; ICSI: Imperfect channel state information; RHIs: Residual hardware impairments; AI: Artificial intelligence; CEEs: Channel estimation errors; CSI: Channel state information; NSCUs: Number of successful communication users; DQN: Deep Q network; EE: Energy efficiency; DBN: Deep belief network; BS: Base station; HD: Half-duplex; PSR: Power splitting relay; EH: Energy harvesting; ID: Information decoding; IF: Information forwarding; SIC: Successive interference cancellation; SINR: Signal-to-interference-plus-noise ratio; RL: Reinforcement learning; AU + DQN: All users + deep Q learning; AU + AP: All users + average power allocation; US + AP: User selection + average power allocation; RU + DQN: Random users + deep Q learning.

Acknowledgements

Not applicable.

Authors' contributions

XZX contributed to basic idea of the paper and provided suggestions for the experimental simulation. ML was responsible for the theoretical analysis, experimental simulation, and manuscript writing of this research. ZYS contributed to the model construction and algorithm design. QH and HT provided suggestions for theoretical analysis and English writing. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Nature Science Foundation of China under Grant No. 61601070, the Key Research Project of Chongqing Education Commission under Grant No. KJZD-K201800603, and the Doctoral Candidate Innovative Talent Project of CQUPT under Grant No. BYJS201912 and No. BYJS2017003.

Availability of data and materials

The author keeps the analysis and simulation datasets, but the datasets are not public.

Declarations

Competing interest

The authors declare that they have no competing interests.

Author details

[1]Chongqing Key Laboratory of Computer Network and Communication Technology, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. [2]Chongqing Key Laboratory of Computer Network and Communication Technology, School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

Received: 31 December 2020 Accepted: 26 April 2021
Published online: 12 May 2021

References

1. T.D.P. Perera, D.N.K. Jayakody, S. Chatzinotas et al., Simultaneous wireless information and power transfer (SWIPT): recent advances and future challenges. IEEE Commun. Surv. Tut. **20**(1), 264–302 (2018)
2. M.A. Hossain, M. Noorr, K.A. Yau et al., A survey on simultaneous wireless information and power transfer with cooperative relay and future challenges. IEEE Access **7**, 19166–19198 (2019)

3. T.M. Hoang, X.N. Tran, B.C. Nguyen et al., On the performance of MIMO full-duplex relaying system with SWIPT under outdated CSI. IEEE Trans. Veh. Technol. **69**(12), 15580–15593 (2020)
4. H.Q. Tran, C.V. Phan, Q.T. Vien, Power splitting versus time switching based cooperative relaying protocols for SWIPT in NOMA systems. Phys. Commun-Amst. **41**, 101098 (2020)
5. M. Hedayati, I. Kim, On the performance of NOMA in the two-user SWIPT system. IEEE Trans. Veh. Technol. **67**(11), 11258–11263 (2018)
6. S.K. Zaidi, S.F. Hasan, X. Gui, Evaluating the ergodic rate in SWIPT-aided hybrid NOMA. IEEE Commun. Lett. **22**(9), 1870–1873 (2018)
7. Z. Yang, Z.G. Ding, P.Z. Fan et al., The impact of power allocation on cooperative non-orthogonal multiple access networks with SWIPT. IEEE Trans. Wirel. Commun. **16**(7), 4332–4343 (2017)
8. L. Bariah, S. Muhaidat, A. Al-Dweik, Error probability analysis of NOMA-based relay networks with SWIPT. IEEE Commun. Lett. **23**(7), 1223–1226 (2019)
9. J.S. Zhou, Y.J. Sun, Q. Cao et al., QoS-based robust power optimization for SWIPT NOMA system with statistical CSI. IEEE Trans. Green Commun. Netw. **3**(3), 765–773 (2019)
10. X.W. Li, J.J. Li, L.H. Li, Performance analysis of impaired SWIPT NOMA relaying networks over imperfect Weibull channels. IEEE Syst. J. **14**(1), 669–672 (2020)
11. G.X. Li, D. Mishra, Y.H.S. Atapattu, Optimal designs for relay-assisted NOMA networks with hybrid SWIPT scheme. IEEE Trans. Commun. **68**(6), 3588–3590 (2020)
12. X.W. Li, Q.S. Wang, J.J. Liu et al., Cooperative wireless-powered NOMA relaying for B5G IoT networks with hardware impairments and channel estimation errors. IEEE Internet Things J. (2020). https://doi.org/10.1109/JIOT.2020.3029754
13. T.N. Do, D.B. Costa, T.O. Duoong et al., A BNBF user selection scheme for NOMA-based cooperative relaying systems with SWIPT. IEEE Commun. Lett. **21**(3), 664–667 (2017)
14. I.H. Lee, H. Jung, User selection and power allocation for downlink NOMA systems with quality-based feedback in rayleigh fading channels. IEEE Wirel. Commun. Lett. **9**(11), 1924–1927 (2020)
15. J.L. Ou, H.H. Yu, H.W. Wu et al., Security transmission scheme for two-way untrusted relay networks based on simultaneous wireless information and power transfer. J. Electr. Inf. Technol. **42**(12), 2908–2914 (2020)
16. T.S. Li, Q.L. Ning, Z. Wang, Optimization scheme for the SWIPT-NOMA opportunity cooperative system. J. Commun. **41**(8), 141–154 (2020)
17. H.L. Yang, A. Alphones, Z.H. Xiong et al., Artificial-intelligence-enabled intelligent 6G networks. IEEE Netw. **34**(6), 272–280 (2020)
18. X.M. Wang, Y.H. Zhang, R.J. Shen et al., DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems. IEEE Internet Things J. **7**(8), 7279–7294 (2020)
19. Y.H. Zhang, X.M. Wang, Y.Y. Xu, et al., Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning, in *IEEE 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, Xi'an, China, 1–6 (2019)
20. J. Tang, C. Luo Ji, J.H. Ou et al., Decoupling or learning: joint power splitting and allocation in MC-NOMA with SWIPT. IEEE Trans. Commun. **68**(9), 5834–5848 (2020)
21. L. Huang, S.Z. Bi, Y.J. Zhang, A deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks. IEEE Trans. Mobile Comput. **19**(11), 2581–2593 (2020)
22. H.L. Yang, Z.H. Xiong, J. Zhao et al., Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications. IEEE Trans. Wirel. Commun. **20**(1), 375–388 (2021)
23. O. Abbasi, A. Ebrahimi, N. Mokari, NOMA inspired cooperative relaying system using an AF relay. IEEE Wirel. Commun. Lett. **8**(1), 261–264 (2019)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.