Approximate SARSA Algorithm for Dimensionality-Challenged Resource Allocation Optimization in MIMO Communication Systems

Xinyan Huang
School of Computing and Artificial Intelligence, Shandong University of Finance and Economics Jinan, 250014, China
E-mail: 15953130256@163.com

Keywords: SARSA algorithm, communication system, resource allocation, multiple-input and multiple-output, dimensional disaster

Received: September 30, 2024

An approximate sate-action-reward-state-action (ASARSA) algorithm is proposed to solve the resource allocation optimization in multiple-input multiple-output (MIMO) communication systems, especially in the context of energy harvesting (EH) wireless communication systems. ASARSA algorithm aims to overcome the dimensional disaster problem faced by traditional SARSA algorithm in high-dimensional state space. By transforming the resource allocation optimization problem into a Markov decision-making problem and applying reinforcement learning, this study realizes the resource allocation optimization of EH-MIMO system. The experimental results showed that the system throughput of ASARSA algorithm reached 15.0×105 bits under the condition of 100 slots, which was 0.2×105 bits and 3.6×105 bits higher than that of SARSA and Q-Learning (QL) algorithms, respectively. In terms of convergence speed, ASARSA algorithm was close to the target accuracy after 76 iterations, which was 25 iterations and 77 iterations less than SARSA and QL algorithms, respectively. In addition, the average absolute error and root mean square error of ASARSA algorithm were 3.54% and 3.10%, which were 1.27% and 0.58%, 2.01% and 1.12% lower than those of SARSA and QL algorithms, respectively. These results show that ASARSA algorithm has higher efficiency and better optimization effect in resource allocation optimization. It is also found that ASARSA algorithm can maintain high computational efficiency and low approximate error, which proves its effectiveness and reliability in practical applications. Therefore, ASARSA algorithm can effectively optimize the allocation of EH-MIMO resources, solve the shortage of spectrum resources to some extent, and promote the development of EH-MIMO technology.

Povzetek: Predstavljen je nov SARSA-algoritem za optimizacijo razporejanja virov v MIMO komunikacijskih sistemih, ki učinkovito rešuje problem dimenzionalnosti in izboljšuje razporejanje virov.

1 Introduction

Nowadays, with the rapid development of society, people are no longer satisfied with a single communication mode, encouraging them to pursue more efficient communication systems. More demand has made space spectrum resources appear to be stretched, which has led the government to strictly manage and unified planning of wireless spectrum usage. Based on this background, a variety of communication technologies with high spectral efficiency have been developed constantly, among which multiple-input multiple-output (MIMO) systems have attracted wide attention [1]. In response to the initiative of developing and applying green communication technology, some research try to introduce energy harvesting device into MIMO wireless communication system to achieve energy saving and emission reduction and increase the service life of the system. However, MIMO is equipped with multiple antennas at both the transmitting and receiving ends. Therefore, the channel of MIMO is usually presented in the form of a matrix, which is more difficult to estimate and process [2]. In addition, the current energy harvesting-MIMO (EH-MIMO) wireless communication system resource allocation optimization algorithm has insufficient prior information and high algorithm complexity, which cannot effectively realize the resource allocation optimization of communication system [3]. Due to the complexity and dynamics of EH-MIMO environment, the resource allocation optimization is still a challenge. To this end, this study transforms the resource allocation optimization problem of EH-MIMO communication system into Markov decision-making problem. A novel method based on reinforcement learning (RL) approximate state-action-reward-state-action (SARSA) algorithm is proposed to obtain the suboptimal transmission strategy, so as to maximize the system throughput and finally complete the resource allocation optimization of

EH-MIMO communication system. By achieving these goals, it aims to contribute to the development of green communication technology and improve the overall performance of the EH-MIMO system. The innovation of the research mainly includes two points. The first point is to extract the characteristics of the EH-MIMO communication system resource allocation optimization problem, transform it into a Markov decision-making problem, and use SARSA algorithm to obtain suboptimal transmission strategies. The second point is to propose an approximate sate-action-reward-state-action (ASARSA) algorithm for dimensional disaster to improve the optimization effect of EH-MIMO communication system resource allocation. The main structure of the study is divided into four sections. The first section is a comprehensive organization and analysis of current relevant research literature. The second section proposes a resource allocation optimization strategy for MIMO communication systems based on SARSA algorithm. The third section analyzes the effectiveness of the resource allocation optimization strategy proposed in the study for MIMO communication systems. The final section is a summary of the entire research content.

2 Related works

MIMO technology has high spectral efficiency and can guarantee the data transmission rate and quality in the communication process. It has been concerned by relevant researchers. Liu et al. [4] put forward a joint transmit beamforming model for dual function MIMO radar and multi-user MIMO communication transmitter. A complexity reduction design was proposed based on zero forced inter user and radar interference. Ma et al. [5] designed a random model based on three-dimensional broadband non-stationary geometry for the MIMO channel of unmanned aerial vehicles. Both line of sight and non-line of sight conditions were considered to explore the rotation effect of unmanned aerial vehicles. Dang et al. [6] proposed a joint message passing detection and decoding algorithm to improve the information and data transmission efficiency. The findings denoted that the algorithm had good performance. Wang et al. [7] proposed a three-dimensional spatiotemporal frequency non-stationary geometric random model and applied it to capture channel characteristics of 6G terahertz ultra large-scale MIMO. Chang et al. [8] proposed a capacity optimization algorithm for MIMO communication systems by combining augmented Lagrangian method, intelligent reflector, and Broyden Fletcher Goldfarb Shano methods, which effectively improved the efficiency of MIMO communication systems. Grossi et al. [9] designed a spectrum sharing architecture that simultaneously existed in MIMO communication systems and surveillance radars. The coexistence and synchronization design of the two systems in the architecture under clutter environment were discussed, providing reference opinions for the practical application of MIMO communication systems and surveillance radar. Temiz et al. [10] aimed to optimize the dual function radar and communication system with the optimization goals of speed and energy efficiency. To achieve the above goals, an optimized pre-encoder for MIMO orthogonal frequency division multiplexing dual radar communication system was proposed. The experiments were designed to analyze the pre-encoder. Zhang [11] designed a signal propagation improvement method for MIMO communication systems by combining intelligent reflective surfaces and passive reflective units, thereby increasing the capacity, reducing the operating costs, and improving the energy efficiency of the MIMO communication system.

RL is one of the most widely used and frequent paradigms and methods in machine learning. Many scholars have paid more attention to the SARSA algorithm. Hassanien et al. [12] proposed an autonomous driving path planning model that combined the Dyna framework based on RL with the SARSA algorithm to address the hidden dangers in computational efficiency and safety in current autonomous driving path planning. This model could effectively ensure the efficiency and safety of path planning. Alfakih et al. [13] proposed a SARSA-based system resource management optimization algorithm for the task unloading and resource allocation of mobile edge computing in the current network physical social system. Chen et al. [14]. combined genetic network programming with evolutionary algorithm of SARSA algorithm to design an artificial financial market, which facilitated solving increasingly complex financial research problems. Rais et al. [15] extended the SARSA algorithm and proposed a Harmonic SK Deep SARSA algorithm to improve its stability. Then, the new algorithm was applied to the decision-making of autonomous vehicle in the expressway scene. Mohamed et al. [16] explored the usage of deep RL technology in network attack detection and classification. An anomaly network intrusion detection model based on the deep SARSA algorithm was designed. This model combined the advantages of SARSA algorithm and deep neural network. Ren et al. [17] constructed an optimization model that combined a neural network model with an RL-based SARSA algorithm. Through this model, the flow shop scheduling problem was solved, thereby improving the production efficiency of the flow shop. Shi et al. [18] proposed a delay aware routing strategy based on SARSA to optimize the network configuration and management of the distribution internet of things (IoT). The limited access range and signal attenuation caused by communication distance and obstacles in existing communication methods were addressed. Aljohani et al. [19] designed an optimization framework based on SARSA algorithm to optimize real-time energy consumption of electric vehicles.

In the above content, SARSA algorithm has important applications in various fields, and there are also certain research results in the optimization of system resource allocation. However, there are few studies in the literature applying SARSA to RA optimization in MIMO wireless communication systems. To solve this problem, a resource allocation optimization based on SARSA is proposed, which improves the performance of MIMO wireless communication system and provides theoretical guidance and new ideas for the development of MIMO wireless communication system. The suboptimal transmission strategy is obtained by Markov decision-making process and SARSA algorithm. Finally, the results and limitations of the existing research and the proposed method are further summarized and compared, as shown in Table 1.

Table 1: Summary table in related works

References	Research method	Limitations
Liu et al [4]	A joint transmission beamforming model is proposed	Requires precise user interference elimination
Ma et al. [5]	A stochastic model based on 3D broadband nonstationary geometry is designed	Model complexity makes it difficult to handle real-world changes
Dang et al. [6]	Improve the efficiency of information and data transmission	Algorithm is inefficient in high-dimensional state spaces
Wang et al. [7]	A three-dimensional space-time frequency non-stationary geometric stochastic model is proposed	Limited adaptability to actual environmental changes
Chang et al. [8]	Combined with augmented Lagrangian method, the efficiency of MIMO communication system is improved	Sensitive to initial conditions
Grossi et al. [9]	The spectrum sharing architecture of MIMO communication system and surveillance radar is designed	Challenges in synchronization design in complex environments
Temiz et al. [10]	An optimized pre-encoder for MIMO orthogonal frequency division multiplexing dual radar communication system is proposed.	The stability of the algorithm in non-ideal environments needs to be verified
Zhang [11]	Combining smart reflector and passive reflector improves the capacity of MIMO communication system	Sensitive to environmental changes
Hassanien et al. [12]	Combining Dyna framework and SARSA algorithm, an autonomous driving path planning model is proposed	Challenges in handling uncertainties in actual driving
Alfakih et al. [13]	An optimization algorithm of system resource management based on SARSA is proposed	Inefficiency in handling high-dimensional problems
Chen et al. [14]	The artificial financial market is designed by combining genetic network programming with evolutionary algorithm of SARSA algorithm	Inefficiency in dealing with complex financial issues
Rais et al. [15]	Harmonic SK Deep SARSA algorithm is proposed	Challenges in decision-making in high-speed scenarios
Mohamed et al. [16]	An abnormal network intrusion detection model is designed based on deep SARSA algorithm	Poor efficiency in handling large-scale network attacks
Ren et al. [17]	Combining neural network model and RL algorithm based on SARSA, the optimization model is constructed	Low efficiency in dealing with complex scheduling problems
Shi et al. [18]	A delay-aware routing strategy based on SARSA is proposed	Low efficiency in handling communication distance and obstacle issues in IoT
Aljohani et al. [19]	A real-time energy consumption minimization framework for electric vehicles based on SARSA algorithm is designed	Low efficiency in addressing real-time energy consumption optimization issues
This paper	A real-time energy consumption minimization framework for electric vehicles based on SARSA algorithm is designed	-

3 Resource allocation strategy for EH-MIMO system based on SARSA algorithm

MIMO technology is a combination of digital modulation, multi-carrier, digital signal processing, and space-time multiplexing technologies, which can effectively improve the anti-interference ability and transmission ability of the system. In this section, an EH-MIMO resource allocation mathematical model is constructed based on MIMO model, and the mathematical model is transformed into a Markov decision-making process, which is solved by RL. After that, the SARSA algorithm is introduced to alleviate the dimensional disaster problem of the model, so as to improve the resource allocation optimization effect of EH-MIMO communication system.

3.1 Construction of EH-MIMO resource allocation mathematical model

MIMO wireless communication system is a comprehensive technology combining digital modulation, multi-carrier transmission, digital signal processing, and space-time multiplexing technologies, which can effectively improve the robustness and transmission capacity of wireless communication system [20]. The integration of EH technology and MIMO technology can not only achieve energy-saving of wireless communication, but also alleviate the shortage of spectrum resources, which is an important direction for the development of green communication in the future [21]. In order to achieve green communication, reduce resource consumption and increase system life, an energy harvesting device is installed at the transmitter of the MIMO wireless communication system, and an EH-MIMO model is constructed. This allows the system to capture and store energy from wind and solar power, where energy storage is achieved through batteries of limited capacity. The EH-MIMO model is shown in Figure 1.

In Figure 1, there are a total of N_T antennas at the transmitting end. There is a total of N_R antennas at the receiving end. In the energy model, it is assumed that there is a total of time slots T within the operating time range. The interval between adjacent time slots τ is a constant. In a time period of $t = 1, 2, \ldots, T$, the collected energy of the energy model is E_t, and the maximum collected energy limit is $E_{\max}$. All collected energy is stored in a battery with a capacity of $B_{\max}$. Assuming that all the energy collected by the transmitting end is applied to the signal transmission work, and no other types of energy loss occur. In addition, during the storage or recycling of the battery, it has no energy loss. Before use, the battery stores a portion of energy B_0. In actual situations, the battery cannot be charged instantaneously. Therefore, during the time slot t, the stored energy of the battery is E_{t-1}. The energy reaching process mentioned above is shown in Figure 2 (a). In addition, in the EH-MIMO model, assuming that wireless channel is a block attenuation flat fading and the change in channel gain H_t over time τ can be ignored, the channel changes are shown in Figure 2 (b).

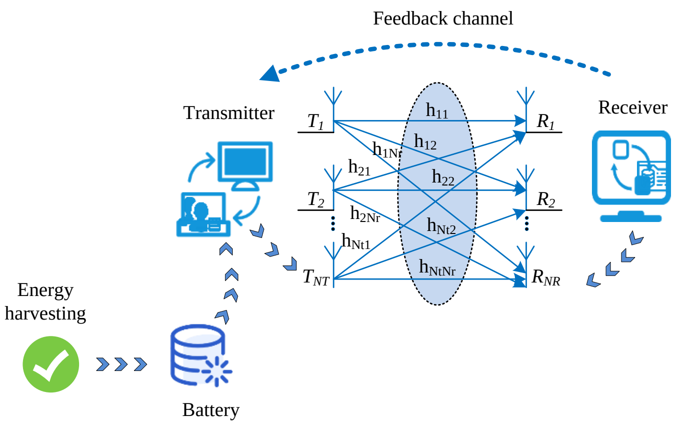

Figure 1: EH-MIMO model

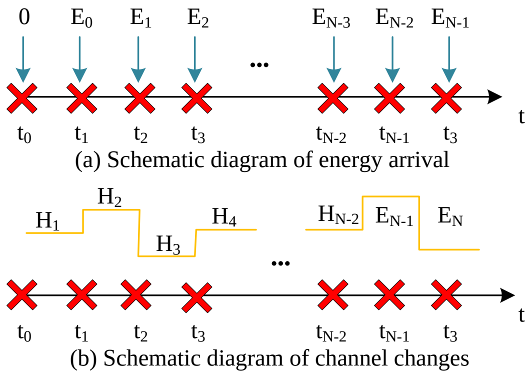

Figure 2: Schematic diagram of energy arrival and channel changes

In Figure 2(a), the energy collected by the system is stored in the battery. When the transmitting end uses the energy to transmit it to the transmitting end, the transmission power is P_t. Therefore, during the battery energy transfer, the update of battery energy follows Formula (1).

$$B_{t+1} = \min\left\{B_t + E_{t-1} - \tau P_t, B_{\max}\right\}, \forall t = 1,2,\ldots,T \qquad (1)$$

In Figure 2(b), the received signal Y_t at the receiving end can be represented by Formula (2).

$$Y_t = \sqrt{P_t} H_t X_t + n_t \qquad (2)$$

In Formula (2), $\sqrt{P_t} = \left\|H_t\right\|^2$ is the power gain of the channel. H_t is the channel gain. X_t is the modulation format vector of all transmitting antenna transmission symbols. n_t is the vector of additive Gaussian white noise, and obeys the mean of 0 and the variance of σ^2. In MIMO systems, due to the fact that the transmitting and receiving ends are equipped with multiple antennas, the channel gain is a matrix with a scale of $N_T \times N_R$. The biggest advantage of MIMO technology is its ability to gain space and capacity. After obtaining channel information, the channel matrix H_t of MIMO can be subjected to singular value decomposition (SVD) to obtain the eigenvalues of H_t, and all eigenvalues are not zero. r is the rank of H_t. In MIMO, there is $N_T \geq N_R$, as shown in Formula (3).

$$r = \min\left\{N_T, N_R\right\} = N_R \qquad (3)$$

Based on the above content, the channel matrix of the MIMO system is subjected to SVD processing to obtain independent parallel single-input single-output (SISO) channels r, as illustrated in Formula (4).

$$H_t = USV^H \qquad (4)$$

In Formula (4), U is a receive shaping filtering matrix with dimension, and $N_R \times N_R$. V is a transmission pre-wave filtering matrix with a dimension of $N_T \times N_T$. S is a diagonal matrix with elements $\lambda_t^1, \lambda_t^2, \ldots, \lambda_t^r$ on the diagonal and dimensions $N_R \times N_T$. At this point, the characteristic value $\lambda_t^1, \lambda_t^2, \ldots, \lambda_t^r$ of H_t can be utilized to stand for the state of each SISO channel at the time slot t. After SVD processing, the EH-MIMO is shown in Figure 3.

When using the transmitter, the number of bits it sends in the time slot t is the system throughput. When the transmitting end only knows causal information, the information that the transmitting end can know includes the current state of B_t, E_t, and H_t, while the future information is in an unknown state. Therefore, a mathematical model for EH-MIMO resource allocation problem can be constructed based on constraint conditions and objective functions, as shown in Formula (5).

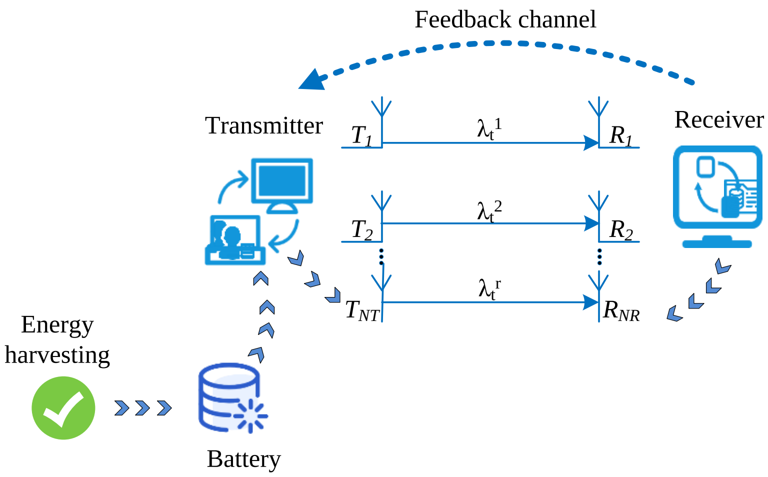

Figure 3: EH-MIMO model after SVD processing

$$\max_{p_t^i} \sum_{t=1}^{T}\sum_{i=1}^{r} B\log_2\left(1+\frac{p_t^i}{\sigma^2}\lambda_t^i\right)$$
$$s.t. \sum_{i=1}^{r} p_t^i \le P_t$$
$$0 \le \tau P_t \le B_t, \forall t = 1,2,\ldots,T$$
$$B_{t+1} = \min\left\{B_t + E_{t-1} - \tau P_t, B_{\max}\right\}, \forall t = 1,2,\ldots,T \quad (5)$$

In Formula (5), B represents the received signal bandwidth. p_t^i represents the transmission power allocated by the CC time slot to the i th SISO channel. σ^2 is the noise power of the SISO channel. Based on the above content, a mathematical model for the EH-MIMO resource allocation problem can be constructed. By solving the problem, EH-MIMO resource allocation optimization can be achieved.

3.2 EH-MIMO mathematical model solution based on RL

Formula (5) is a convex optimization model. However, in solving using convex optimization, it is necessary for the transmitting end to obtain the states of all time slots, but this is difficult to achieve in practice. Therefore, the convex optimization solution method is not applicable to the model shown in Formula (5). Therefore, the study transforms the mathematical model shown in Formula (5) into a Markov decision-making process, and then applies RL to solve it. Markov decision-making process is a process to find the optimal strategy, which includes Markov process and dynamic programming [22]. The state space is defined as the set of all possible states of the system, including the energy level in the battery, channel conditions and current transmission strategy. The action space is defined as the set of all possible transmission strategies that can be adopted in each state. Based on the current state and the selected action, the transition probability between States is simulated, and the randomness of energy arrival and channel change is considered. At the same time, a reward function is defined based on system throughput or energy efficiency to quantify the performance of each pair of state-actions. Combining the above process, model transformation is implemented. The energy level represents the energy currently stored in the battery. Channel conditions include channel gain and channel state information. Action space is defined as the set of all possible transmission strategies that can be adopted in each state. The reward function quantifies the performance of each pair of state-actions based on system throughput or energy efficiency. If the state s_{t+1} of the system in the next time slot is only related to the current state s_t of the system, and there is a transition probability Formula (6), it indicates that the state has Markov properties.

$$P\left[s_{t+1}\middle|s_t\right] = P\left[s_{t+1}\middle|s_1, s_2, \ldots, s_t\right] \quad (6)$$

According to the causality of adjacent states, in the s_t state, past states $s_1 \sim s_{t-1}$ can be discarded. If the states of all time slots in the system have Markov properties, this is a Markov stochastic process. In EH-MIMO mathematical model solving, RL can obtain transmission strategies based on Markov decision-making processes. RL is a continuous interaction between agent and its environment. Through the interaction, the updating decision strategy of RL can be updated in real time to carry out the next step [23]. The essence of RL is to solve intelligent agents, thereby changing the update decision strategy, and ultimately maximizing rewards. The above process can be represented by Figure 4.

The transmission strategy refers to the action a method selected when the state is s, as shown in Formula (7).

$$\pi(a_t|s_t) = p\left[a_t \in A|s_t \in S\right] \quad (7)$$

In Formula (7), S is the set of system states. π is a strategy. a_t is an optional action. A is an optional action set. Formula (7) represents the action selection probability p that the agent can obtain through π when the system state is s_t. The agent can select a a_t in A through p. π is the method for selecting actions and remains constant. The intelligent agent continuously calculates the cumulative return function and obtains a suboptimal transmission strategy through this method. From the Bellman equation, it can be inferred that any strategy π corresponds to a certain action value function (AVF).

Therefore, it is required to calculate and solve the optimal AVF $q_\pi^*(s_t, a_t)$ to obtain the optimal strategy π^* and obtain the relatively optimal transmission strategy. When the optimal action is selected for any state in the system, this optimal set of actions is the optimal transmission strategy. In the study, Markov can be represented as a five tuple $\langle S, A, P_r, R, T\rangle$ based on this process. S is the set of states. A is a set of actions. P_r is the probability of state s_t transitioning to s_{t+1} at time slot $t+1$ after the agent selects action a_t during time slot t. R is the reward received by the state s_t after taking action a_t. T is the total number of time slots. In practical situations, the P_r of the model is an unknown number, so the model can be constructed as a model free Markov model, which adopts a model free rein RL method for the EH-MIMO model. There are generally two types of RL methods without models, namely Monte Carlo method and time difference method. Figure 5 displays the Monte Carlo method schematic diagram. This method obtains the value function by exploring multiple times to obtain the mean.

The time difference is also a commonly used method in RL. Its biggest difference from the Monte Carlo method lies in obtaining the value function, as shown in Figure 6.

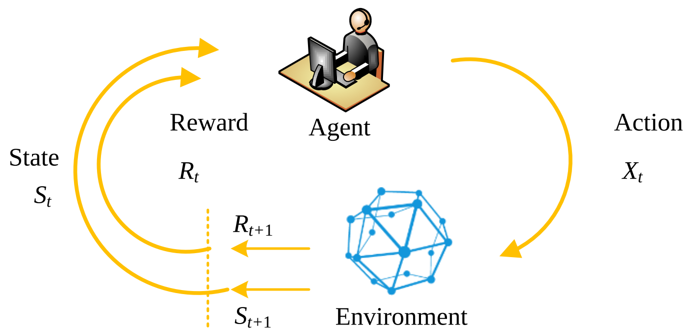

Figure 4: The training of RL

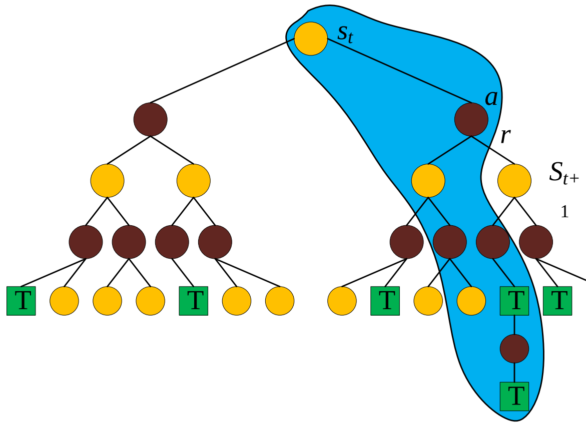

Figure 5: Schematic diagram of Monte Carlo method

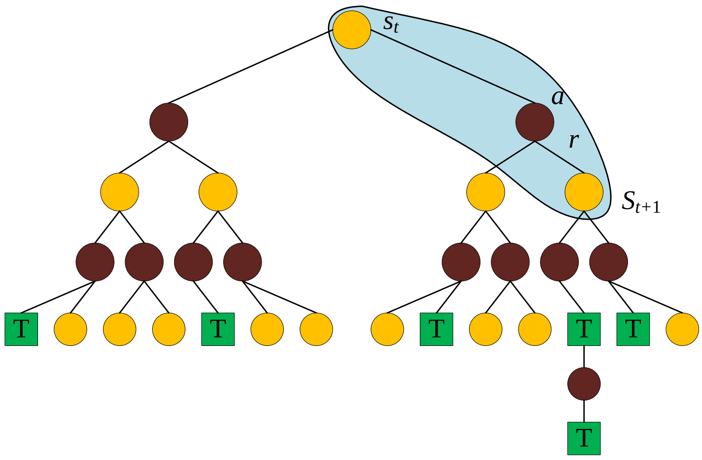

Figure 6: Schematic diagram of time difference method

In Figure 6, the time difference method does not need to go through all time slots and the resulting value function has a small variance. Therefore, the study applies time difference method to solve the model and obtain suboptimal transmission strategies. Q-learning (QL) is a common time difference separation line algorithm, which can obtain the MIMO system's power allocation in each time slot, and then obtain the optimal transmission power. However, the QL algorithm selects the maximum Q value action in a certain state s_{t+1}, which may lead to the algorithm ignoring other actions with the same value, resulting in insufficient exploration and affecting the final optimization strategy. Therefore, another algorithm in the time difference method, namely the SARSA, is applied to solve the model [24]. SARSA is an online algorithm. Different from QL algorithm, SARSA algorithm randomly selects the action with the maximum Q value based on a set probability when selecting actions, thus avoiding the defect of insufficient exploration in QL algorithm. In the SARSA algorithm, the update rules for the Q -table are shown in Formula (8).

$$\begin{aligned} Q(s_t,a_t) &\leftarrow Q(s_t,a_t) \\ &+\alpha\left(R_t+\gamma Q(s_{t+1},a_t)-Q(s_t,a_t)\right) \end{aligned} \tag{8}$$

In Formula (8), $Q(s_t,a_t)$ is the AVF corresponding to the state action pair.
α is the learning rate of the algorithm, which can control the speed of the algorithm environmental exploration. γ is a discount factor, mainly used to determine the importance of the current AVF and the action function for the next time slot. In RL, action selection strategies can affect the environmental exploration performance of the algorithm, thereby affecting its performance. Therefore, an appropriate action selection strategy is crucial for the SARSA algorithm. After comprehensive consideration, the study adopts a Greedy Softmax strategy that combines Softmax and Greedy. This strategy can effectively balance the degree of environmental exploration and algorithm convergence, and considering the structure of Markov decision-making processes makes it suitable for the

research content. The Greedy Softmax strategy is shown in Formula (9).

$$\pi(\varepsilon_1,\xi,s_t,a_t)=\begin{cases} Softmax\ policy & ,if\ \Delta\le\varepsilon_1 \\ \arg\max\limits_{a_t\in A} Q(s_t,a_t) & ,if\ \Delta>\varepsilon_1 \end{cases} \tag{9}$$

In Formula (9), Δ is a uniform random number generated for each time slot, with a value range of (0,1). ε_1 is a fixed value, with a value range of (0,1). ξ is a temperature parameter. In addition to RL for data resource allocation and value function, this study also introduces multiple quadrature amplitude modulation (MQAM) wireless communication system to enhance the transmission process through adaptive coding. The flexible rate-power adjustment of the adaptive model can improve the overall performance of the network. Two hypotheses are proposed. One is that the system satisfies linear modulation, and the adjustment time is an integer multiple of the code gap T_s. Second, the system pulse is selected in off-line Nyquist form, and the signal bandwidth is expressed as $B=\frac{1}{T_s}$. Taking the transmission speed method of the sender as an example, it can be expressed as $R_s=\frac{1}{T_s}$. The MQAM model may modulate different conditions simultaneously to achieve improved spectrum utilization. Under the background of additive white Gaussian noise channel, the theoretical bit error rate range of the model is calculated, as shown in Formula (10).

$$P_b\le 2e^{-1.5\eta(M-1)} \tag{10}$$

In Formula (10), P_b denotes the transmitting power. η is the signal-to-noise ratio. M represents constellation points.

3.3 EH-MIMO mathematical model solution based on ASARSA

In the previous content, the study utilizes the SARSA algorithm to solve the EH-MIMO mathematical model to obtain the suboptimal power transmission strategy. The learning of the SARSA algorithm is illustrated in Figure 7.

In MIMO systems, because there are multiple antennas at both the transmitting and receiving ends, the number of management state pairs in Q-table is very large, resulting in insufficient dimension, and the inability to construct the table, which greatly affects the performance of the algorithm. To solve this problem, an ASARSA algorithm based on linear value function is proposed. ASARSA algorithm plays a key role in solving the "dimensional disaster" problem of traditional QL method in high-dimensional state space.

ASARSA algorithm adopts linear value function approximation, which is a major difference from the traditional tabular method that needs to store separate values for each pair of states-actions. The ASARSA algorithm does not store the Q-table in the transmitter of the MIMO system, but replaces the Q-table with a constructed basis function. The basis function is shown in Formula (11).

$$f_m(s_t,a_t), m=1,2,\ldots,M \tag{11}$$

In Formula (11), M means the total amount of constructed basis functions. Next, the corresponding initial weights w_m are assigned to all basis functions. By utilizing the weights corresponding to the basis function and the basis function, an approximate AVF $\hat{Q}(s_t,a_t,w)$ can be obtained. This value to replace the $Q(s_t,a_t)$ value in the traditional SARSA algorithm. The approximate AVF $\hat{Q}(s_t,a_t,w)$ can be solved using Formula (12).

$$\hat{Q}(s_t,a_t,w)\approx Q(s_t,a_t)=f^T w \tag{12}$$

In Formula (12), $f\in R^{M\times1}$ is a matrix composed of basic functions. $w\in R^{M\times1}$ is a matrix constructed by the corresponding weights of the basis function. When using ASARSA, the closer the value $\hat{Q}(s_t,a_t,w)$ is to the value $Q(s_t,a_t)$, the better the performance of the algorithm. It uses the least squares difference to evaluate the approximation accuracy between the two, as shown in Formula (13).

$$J(w)=E_{\pi}\left[\left(Q(s_t,a_t)-\hat{Q}(s_t,a_t,w)\right)^2\right] \quad (13)$$

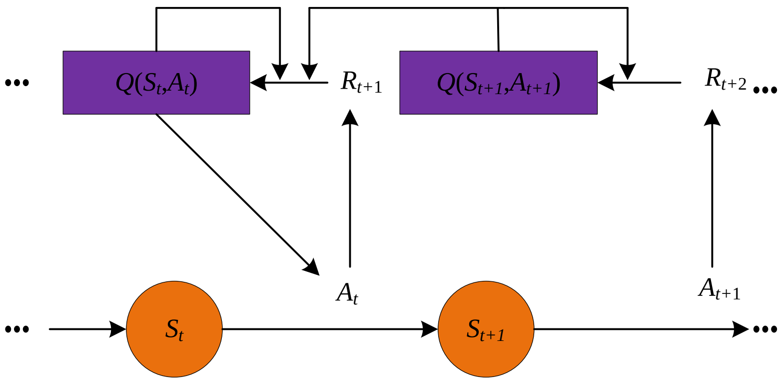

Figure 7: The learning process of SARSA algorithm

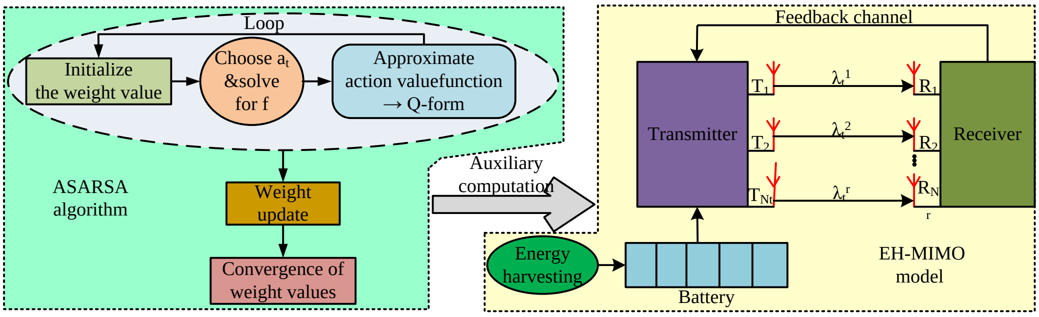

Figure 8: Overall EH-MIMO model based on ASARSA algorithm

It minimizes $J(w)$ to obtain the optimal approximation accuracy. Therefore, the gradient descent method is applied to calculate w. The gradient of $\hat{Q}(s_t,a_t,w)$ is shown in Formula (14)

$$\nabla\hat{Q}(s_t,a_t,w)=f \quad (14)$$

The value of w is adjusted according to the direction of gradient descent, so as to minimize the error between $\hat{Q}(s_t,a_t,w)$ and $Q(s_t,a_t)$. The weights of the ASARSA are updated according to Formula (15).

$$w\leftarrow w+\alpha_t\begin{bmatrix}R_t+\gamma Q(s_{t+1},a_{t+1},w)\\ -\hat{Q}(s_t,a_t,w)\end{bmatrix}f \quad (15)$$

According to the above design and formula, the model of the whole system can be obtained, as shown in Figure 8. In the ASARSA algorithm, the first step is to initialize the weight values corresponding to all basis functions, that is, to assign initial weight values to all basis functions. When in time slot t, it selects action a_t based on π in state s_t. Subsequently, it solves f and $\hat{Q}(s_t,a_t,w)$ according to Formula (12). Next, the state s_t shifts to s_{t+1}. Repeating the above operation can obtain $\hat{Q}(s_{t+1},a_{t+1},w)$. Then, the weight values corresponding to the basis function are updated using Formula (15). After the algorithm fully explores the environment, the weight values converge and the correlation between the state and action is obtained. When the transmitter of the MIMO system is in the utilization stage, based on this correlation, the corresponding a_t can be obtained at s_t. Because there is no Q-table in ASARSA algorithm, it can effectively avoid the dimensional disaster. Base on the above content, the ASARSA algorithm is constructed, and the EH-MIMO mathematical model is solved using this algorithm to optimize resource allocation in MIMO communication systems. At the same time, the study sets the exploration probability and learning rate as 1/k, where k is the number of iterations. This setting makes the algorithm tend to explore at the beginning, and gradually shift towards using known strategies as the iteration progresses to promote rapid convergence. Decreasing learning rate helps to learn quickly in the early stage, reduce the updating range in the later stage and avoid shock. The temperature parameter is initially set to 100 and gradually decreases as learning progresses to increase the tendency to select the best action. These parameters are selected to balance exploration and utilization and ensure the effectiveness and stability of the algorithm.

4 Performance analysis of system resource allocation optimization strategy based on ASARSA

To prove the optimization effect of ASARSA algorithm on resource allocation of MIMO communication system, simulation experiments are conducted in this study. To highlight the superior performance of ASARSA algorithm, this study chooses to compare and verify it with SARSA algorithm and QL algorithm. The experiment tests the performance of the model from the perspectives of convergence, F1, Recall, MAE, and RMSE values, throughput of EH-MIMO wireless communication system under different algorithms, and ROC curves of different algorithms.

The parameters of the simulation experiment here mainly consider offline strategy, greedy strategy, conservative strategy and random strategy. Simulation parameters are shown in Table 2.

In Table 2, k represents the number of learning rounds for the current agent, which compares the effectiveness of ASARSA, SARSA, and QL in solving MIMO resource allocation optimization mathematical models. Firstly, it is required to compare the convergence of ASARSA, SARSA, and QL when solving the EH-MIMO resource allocation optimization mathematical model. The loss value is used as the judgment index in the experiment, as shown in Figure 9. In Figure 9(a), the ASARSA achieved near target accuracy after 76 iterations, which was 25 and 77 fewer than SARSA and QL, respectively. In Figure 9(b), the ASARSA algorithm approached the target accuracy after 10.0 seconds, which was 4.8 seconds less than the SARSA and 9.7 seconds less than the QL, respectively.

This study uses F1 and recall rates to evaluate and compare the performance of ASARSA, SARSA, and QL. F1 is a binary model accuracy measure related to model precision and recall rate. The recall rate is a measure of the model recall rate, as shown in Figure 10. In Figure 10(a), the F1 values of ASARSA were 96.52%, 1.05%, and 1.34% higher than SARSA and QL, respectively. In Figure 10(b), the recall value of ASARSA was 96.33%, which was 0.27% and 0.36% higher than SARSA and QL, respectively.

Table 2: Simulation parameter settings

Parameter	Unit	Value
Number of antennas at the transmitting end	-	2
Number of antennas at the receiving end	-	2
Noise	W/Hz	0.2
Bandwidth	Hz	105
Time interval	s	1
Total time slots	-	100
The total number of rounds of intelligent agent learning	-	1000000
Temperature parameters	-	100
Exploration probability	-	1/k
Learning rate	-	1/k
Initial battery energy	J	0.15
Battery capacity	J	0.25
Maximum collected energy	J	0.20
Energy quantification step size	-	0.05

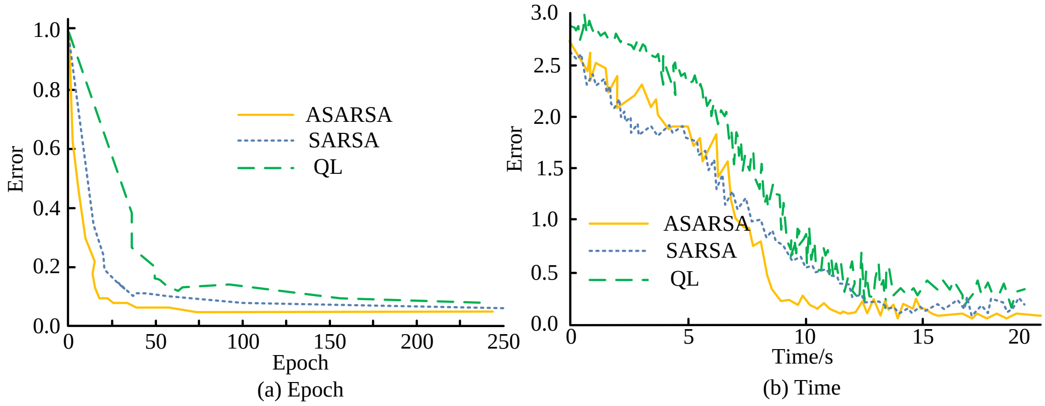

Figure 9: Convergence analysis of algorithms

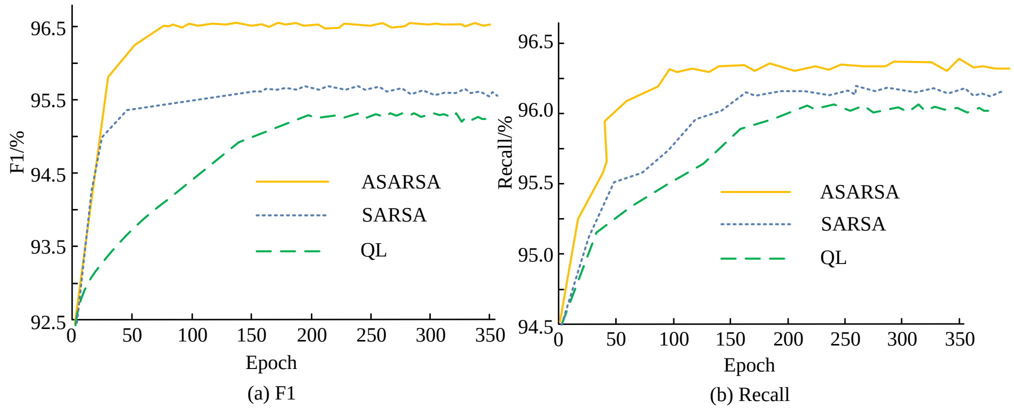

Figure 10: F1 value and recall value of the algorithm

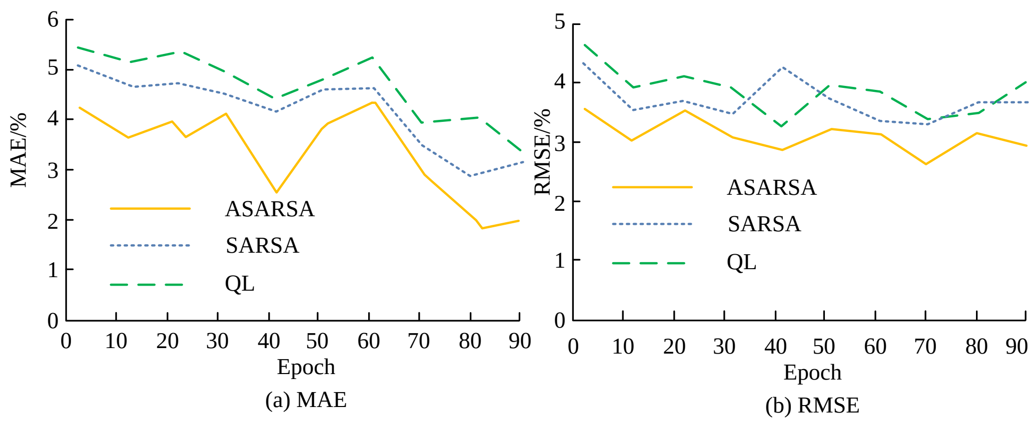

Figure 11: MAE value and RMSE of the algorithm

This study evaluates and compares the performance of the ASARSA, SARSA, and QL using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values, which are shown in Figure 11. In Figure 11(a), the average MAE value of the ASARSA was 3.54%, which was 1.27% and 2.01% lower than the SARSA and the QL, respectively. In Figure 11(b), the average RMSE value of the ASARSA was 3.10%, which was 0.58% and 1.12% lower than the SARSA and the QL, respectively.

The throughput of the EH-MIMO wireless communication system varies with the number of slots under different algorithm strategies, as shown in Figure 12. The throughput of the system under the two strategies of ASARSA and SARSA was relatively close. Under the QL strategy, the throughput of the system was significantly lower than that of ASARSA and SARSA. When the number of slots was 100, the throughput of the system under the ASARSA strategy was 15.0×105 bit, which was 0.2×105 bit and 3.6×105 bit higher than SARSA and QL, respectively.

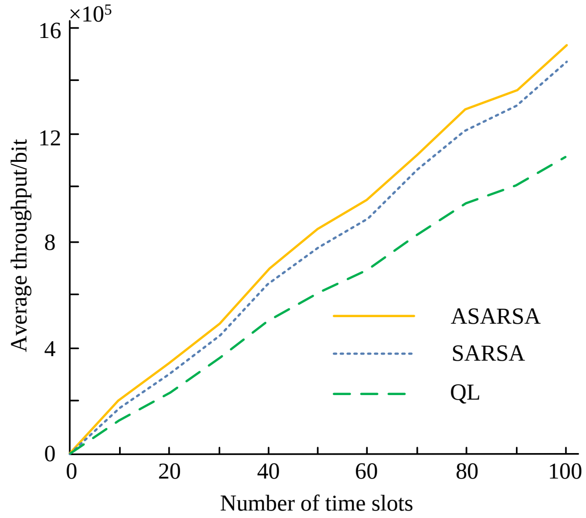

Figure 12: The variation of wireless communication system throughput under different time slot numbers

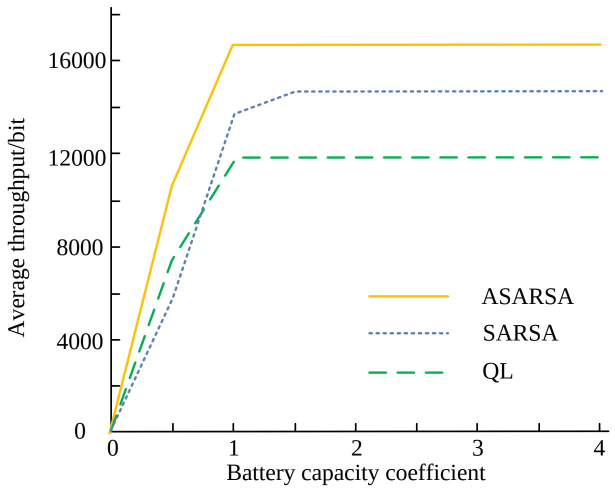

Figure 13: The variation of wireless communication system throughput with the change of battery capacity coefficient

Under the RA strategy obtained after the EH-MIMO resource allocation optimization mathematical model, different algorithms are used to solve the throughput of the EH-MIMO wireless communication system when comparing different battery capacity coefficients (ρ). The battery capacity coefficient varies according to different algorithm strategies, as shown in Figure 13. The throughput of the system under the two strategies of ASARSA and SARSA was relatively close, while under the QL strategy, the throughput of the system was significantly lower than that of ASARSA and SARSA. When the battery capacity coefficient was 4, the throughput of the system under the ASARSA strategy was 16900 bit, which was 1600 bit and 5000 bit higher than the SARSA and QL, respectively.

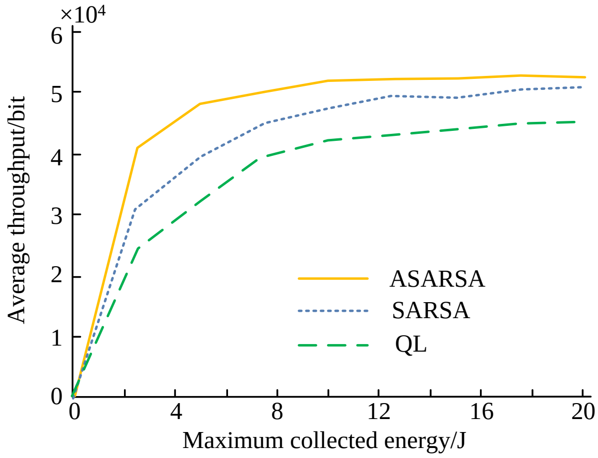

Figure 14: The change in throughput of wireless communication systems with the maximum collected energy

Figure 14 shows the throughput variation of EH-MIMO wireless communication system under different algorithm strategies at maximum collected energy. The throughput of the system under the two strategies of ASARSA and SARSA was relatively close, while under the QL strategy, the throughput of the system was significantly lower than that of ASARSA and SARSA. When the maximum value of collected energy was 20, the throughput of the system under the ASARSA strategy was 5.1×104 bit, 0.1×104 bit and 0.7×104 bit higher than SARSA and QL, respectively.

This paper uses ROC curves to analyze the comprehensive performance of ASARSA, SARSA, and QL, as shown in Figure 15. This paper uses ROC curves to analyze the comprehensive performance of ASARSA, SARSA, and QL, as shown in Figure 15. The AUC value of the ASARSA was 0.963, which was 0.016 and 0.027 higher than the SARSA and the QL, respectively. On this basis, in order to further verify the robustness of ASARSA algorithm, the sensitivity analysis of key parameters is conducted. The exploration probability and learning rate vary from 0.1 to 0.01, while the temperature parameter varies from 50 to 200. Table 3 shows detailed information.

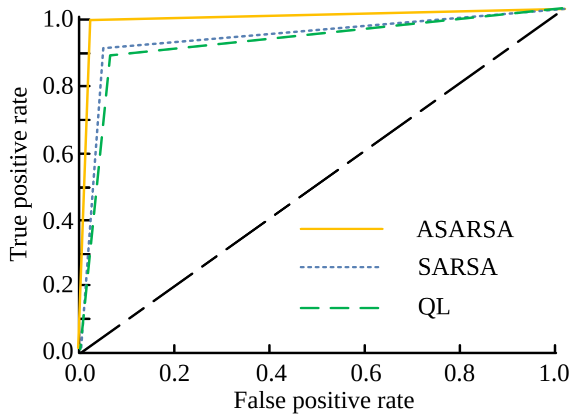

Figure 15: ROC curve analysis of the algorithm

Table 3: Parameter sensitivity analysis

Index	Exploration probability/learning rate			Temperature parameter (℃)		
	0.1	0.05	0.01	50	100	200
System throughput (bits)	14.5×105	14.8×105	15.0×105	14.2×105	15.0×105	14.3×105

Convergence speed (times)	90	80	100	120	100	110

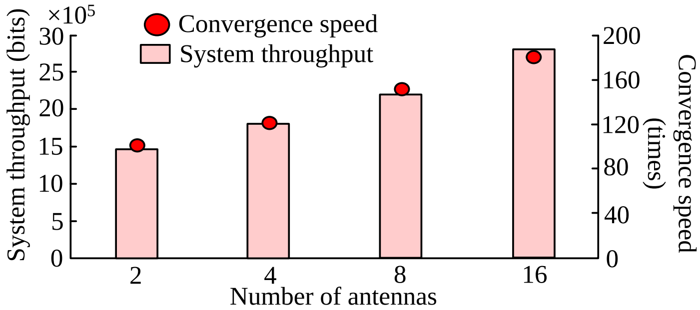

Figure 16: Simulation results under different antenna numbers

From Table 3, the algorithm is robust to parameter changes within a certain range, but its performance obviously declines beyond a specific range. This indicates that although parameter selection can affect algorithm performance, adjusting parameters within a reasonable range will not affect the effectiveness of the ASARSA algorithm. At the same time, the performance of the algorithm under different antenna numbers is further simulated. Scenes with 2, 4, 8 and 16 antennas are simulated, and the system throughput and algorithm convergence speed in each case are recorded, as shown in Figure 16.

From Figure 16, the results show that with the increase of the number of antennas, ASARSA algorithm can still maintain a high system throughput, and the convergence speed is only slightly reduced. This discovery proves that ASARSA algorithm has good scalability in EH-MIMO systems of different scales. In summary, the ASARSA proposed in the study performs better and has higher efficiency in solving the EH-MIMO resource allocation optimization mathematical model. Under the ASARSA algorithm strategy, the throughput of the EH-MIMO system is higher and the optimization effect is better. Therefore, ASARSA can effectively optimize the allocation of EH-MIMO resources, thereby solving the spectrum resource shortage to a certain extent, and promoting the development of EH-MIMO technology.

5 Discussion

The combination of EH technology and MIMO system can achieve the energy-saving of wireless communication system and solve the spectrum resource shortage, which is one of the development trends of green communication in the future. The current EH-MIMO wireless communication system resource allocation optimization algorithm has the defects of insufficient prior information and high algorithm complexity, which can not effectively realize the communication system resource allocation optimization. In view of this, to solve the above problems, this study transformed the resource allocation optimization problem of EH-MIMO communication system into a Markov decision-making problem. An ASARSA algorithm based on RL was proposed to obtain the suboptimal transmission strategy, so as to maximize the system throughput and finally complete the resource allocation optimization of EH-MIMO communication system.

Under the condition of 100 time slots, the ASARSA algorithm achieved a system throughput of 15.0×105 bits, which was significantly higher compared with the SARSA and QL algorithms. In terms of convergence speed, the ASARSA algorithm approached the target accuracy after 76 iterations, 25 and 77 fewer than the SARSA and QL algorithms, respectively. These results indicate that the ASARSA algorithm has higher efficiency and better optimization effects in resource allocation optimization. The innovation of the ASARSA algorithm lies in its ability to overcome the dimensional disaster. In high-dimensional state spaces, the traditional SARSA algorithm needs to store a large number of state-action pairs, which not only occupies a lot of memory, but also increases computational complexity. The ASARSA algorithm effectively solves the dimensional disaster problem through the linear AVF, avoiding the need to store the Q-table. Moreover, the ASARSA algorithm achieves a good balance between computational efficiency and accuracy. Although it uses an approximation method, its performance is still superior to or close to other models.

In high-dimensional spaces, one of the main challenges faced by the SARSA algorithm is the dimensional disaster, that is, the performance of the algorithm drops sharply as the dimension of the state space increases. The ASARSA algorithm effectively avoids this problem through a basis function method to approximate the AVF. Compared with the traditional QL method, the ASARSA algorithm does not need to store separate values for each state-action pair, but constructs the AVF through basis functions and corresponding weights, which greatly reduces memory requirements and computational complexity.

Although the ASARSA algorithm improves computational efficiency through approximation methods, this may also introduce approximation errors. To quantify this trade-off, the research evaluates the approximation accuracy of the ASARSA algorithm. The results show that while maintaining high computational efficiency, the ASARSA algorithm can still maintain low approximation errors, proving its effectiveness and reliability in practical applications. In summary, the ASARSA algorithm has obvious advantages in solving the resource allocation optimization problems in EH-MIMO communication systems, especially when dealing with high-dimensional state spaces. At the same time, the innovation and computational trade-offs of the ASARSA algorithm in dealing with the dimensional disaster also provide important reference value for its practical application. Future work will further improve the generalization ability of the algorithm and verify it in a wider range of communication systems.

6 Conclusion

In the EH-MIMO wireless communication system, reasonable resource allocation can effectively improve system efficiency and save system resources. To this end, a mathematical model is constructed for optimizing RA in the EH-MIMO system, and an ASARSA is proposed to solve it. The experimental results showed that the ASARSA achieved close to target accuracy after 76 iterations, which was 25 and 77 fewer than SARSA and QL, respectively. After 10.0 seconds of iteration, the target accuracy was approached, which was 4.8 seconds less than SARSA and 9.7 seconds less than QL, respectively. The F1 value was 96.52%, which was 1.05% and 1.34% higher than SARSA and QL, respectively. The Recall value was 96.33%, which was 0.27% and 0.36% higher than SARSA and QL, respectively. The MAE value was 3.54%, which was 1.27% and 2.01% lower than SARSA and QL, respectively. The RMSE value was 3.10%, which was 0.58% and 1.12% lower than SARSA and QL, respectively. When the number of time slots was 100, the system throughput was 15.0×105 bit, 0.2×105 bit and 3.6×105 bit higher than SARSA and QL, respectively. When the battery capacity coefficient was 4, the system throughput was 16900 bit, which was 1600 bit and 5000 bit higher than SARSA and QL, respectively. When the maximum value of collected energy was 20, the system throughput was 5.1×104 bit, 0.1×104 bit and 0.7×104 bit higher than SARSA and QL, respectively. The AUC value was 0.963, which was 0.016 and 0.027 higher than SARSA and QL, respectively. This research innovatively extracts the characteristics of resource allocation optimization problems in EH-MIMO communication systems, transforms them into a Markov decision-making process, and uses SARSA algorithm to obtain the suboptimal transmission strategy. In addition, an ASARSA algorithm was proposed to solve the dimensional disaster problem, so as to improve the resource allocation optimization effect of EH-MIMO communication systems. The experimental results showed that ASARSA algorithm effectively achieved EH-MIMO resource allocation optimization, and then solved the spectrum resource shortage to a certain extent, promoting the development of EH-MIMO technology. The limitation of this study is that the number of antennas set in the simulation experiment parameters is small, which is different from the actual situation. Therefore, there may be some deviation between the obtained experimental results and the actual situation. Subsequently, the number of antennas should be increased to improve the reliability of the experiment.

7 Funding statement

The research is supported by Research and Analysis on the Big Data Laboratory Platform Construction System of the Industry-University Cooperative Education Project of Ministry of Education in 2022. (No. 220506090201041).

References

[1] Y. Zhang, Y. Wu, A. Liu, X. Xia, T. Pan, and X. Liu, "Deep learning-based channel prediction for LEO satellite massive MIMO communication system," IEEE Wireless Communications Letters, vol. 10, no. 5, pp. 1835-1839, 2021. https://doi.org/10.1109/LWC.2021.3083267

[2] Y. S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "A compressive sensing approach for federated learning over massive MIMO communication systems," IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1990-2004, 2020. https://doi.org/10.1109/TWC.2020.3038407

[3] X. Ma, Z. Chen, W. Chen, Z. Li, Y. Chi, C. Han, and S. Li, "Joint channel estimation and data rate maximization for intelligent reflecting surface assisted terahertz MIMO communication systems," IEEE Access, vol. 8, pp. 99565-99581, 2020. https://doi.org/10.1109/ACCESS.2020.2994100

[4] X. Liu, T. Huang, N. Shlezinger, Y. Liu, and Y. C. Eldar, "Joint transmit beamforming for multiuser MIMO communications and MIMO radar," IEEE Transactions on Signal Processing, vol. 68, pp. 3929-3944, 2020. https://doi.org/10.1109/TSP.2020.3004739

[5] Z. Ma, B. Ai, R. He, G. Wang, Y. Niu, M. Yang, J. Wang, Y. Li, and Z. Zhong, “Impact of UAV rotation on MIMO channel characterization for air-to-ground communication systems,” IEEE Transactions on Vehicular Technology, vol. 69, no. 11, pp. 12418-12431, 2020. https://doi.org/10.1109/TVT.2020.3028301

[6] H. N. Dang, H. T. Nguyen, and T. V. Nguyen, “Joint detection and decoding of mixed-adc large-scale mimo communication systems with protograph ldpc codes,” IEEE Access, vol. 9, pp. 101013-101029, 2021. https://doi.org/10.1109/ACCESS.2021.3097444

[7] J. Wang, C. X. Wang, J. Huang, H. Wang, and X. Q. Gao, “A general 3D space-time-frequency non-stationary THz channel model for 6G ultra-massive MIMO wireless communication systems,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 6, pp. 1576-1589, 2021. https://doi.org/10.1109/JSAC.2021.3071850

[8] D. Chang, H. Jiang, J. Zhou, H. Zhang, and M. Mukherjee, “Capacity optimization using augmented lagrange method in intelligent reflecting surface-based MIMO communication systems,” China Communications, vol. 17, no. 12, pp. 123-138, 2020. https://doi.org/10.23919/JCC.2020.12.009

[9] E. Grossi, M. Lops, and L. Venturino, “Joint design of surveillance radar and MIMO communication in cluttered environments,” IEEE Transactions on Signal Processing, vol. 68, pp. 1544-1557, 2020. https://doi.org/10.1109/TSP.2020.2974708

[10] M. Temiz, E. Alsusa, and M. W. Baidas, “Optimized precoders for massive MIMO OFDM dual radar-communication systems,” IEEE Transactions on Communications, vol. 69, no. 7, pp. 4781-4794, 2021. https://doi.org/10.1109/TCOMM.2021.3068485

[11] S. Zhang, and R. Zhang, “Capacity characterization for intelligent reflecting surface aided MIMO communication,” IEEE Journal on Selected Areas in Communications, vol. 38, no. 8, pp. 1823-1838, 2020. https://doi.org/10.1109/JSAC.2020.3000814

[12] A. E. Hassanien, and J. Mononteliza, “Autonomous driving path planning based on sarsa-dyna algorithm,” Asia-pacific Journal of Convergent Research Interchange, vol. 6, no.7, pp. 59-70, 2020. https://doi.org/10.47116/apjcri.2020.07.06

[13] T. Alfakih, M. M. Hassan, A. Gumaei, C. Savaglio, and G. Fortino, “Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on SARSA,” IEEE Access, vol. 8, pp. 54074-54084, 2020. https://doi.org/10.1109/ACCESS.2020.2981434

[14] Y. Chen, Z. Xu, and W. Yu, “Agent-based artificial financial market with evolutionary algorithm,” Economic Research-Ekonomska Istraživanja, vol. 35, no. 1, pp. 5037-5057, 2021. https://doi.org/10.1080/1331677X.2021.2021098

[15] M. S. Rais, R. Boudour, K. Zouaidia, and L. Bougueroua, “Decision making for autonomous vehicles in highway scenarios using Harmonic SK Deep SARSA,” Applied Intelligence, vol. 53, no.3 pp. 2488-2505, 2023. https://doi.org/10.1007/s10489-022-03357-y

[16] S. Mohamed, and R. Ejbali, “Deep SARSA-based reinforcement learning approach for anomaly network intrusion detection system,” International Journal of Information Security, vol. 22, no.1, pp. 235-247, 2022. https://doi.org/10.1007/s10207-022-00634-2

[17] J. Ren, C. Ye, and F. Yang, “Solving flow-shop scheduling problem with a reinforcement learning algorithm that generalizes the value function with neural network,” Alexandria Engineering Journal, vol. 60, no.3, pp. 2787-280, 2021. https://doi.org/10.1016/j.aej.2021.01.030

[18] Z. Shi, J. Zhu, and H. Wei, “SARSA-based delay-aware route selection for SDN-enabled wireless-PLC power distribution IoT,” Alexandria Engineering Journal, vol. 61, no. 8, pp. 5795-5803, 2022. https://doi.org/10.1016/j.aej.2021.11.029

[19] T. M. Aljohani, and O. Mohammed, “A real-time energy consumption minimization framework for electric vehicles routing optimization based on sarsa reinforcement learning,” Vehicles, vol. 4, no.4, pp. 1176-1194, 2022. https://doi.org/10.3390/vehicles4040062

[20] R. Brociek, M. Goik, J. Miarka, M. Pleszczyński, and C. Napoli, "Solution of inverse problem for diffusion equation with fractional derivatives using metaheuristic optimization algorithm," Informatica, vol. 35, no 3, pp. 453-481, 2024. https://doi.org/10.15388/24-INFOR563

[21] N. Sindhwani, and M. Singh, “A joint optimization based sub-band expediency scheduling technique for MIMO communication system,” Wireless Personal Communications, vol. 115, pp. 2437-2455, 2020. https://doi.org/10.1007/s11277-020-07689-1

[22] R. Chen, H. Zhou, W. X. Long, and M. Moretti, “Spectral and energy efficiency of line-of-sight OAM-MIMO communication systems,” China Communications, vol. 17, no. 9, pp. 119-127, 2020. https://doi.org/10.23919/JCC.2020.09.010

[23] J. Wang, D. Yang, K. Chen, and X. Sun, “Cruise dynamic pricing based on SARSA algorithm,” Maritime Policy & Management, vol. 48, no. 2, pp. 259-282, 2021. https://doi.org/10.1080/03088839.2021.1887529

[24] F. de Arriba-Pérez, S. García-Méndez, F. Leal, B. Malheiro, and J. C. Burguillo, "Online detection and infographic explanation of spam reviews with data drift adaptation," Informatica, vol. 35, no. 3, pp. 483-507, 2024. https://doi.org/10.15388/24-INFOR562