Adaptive Modulation Scheme for Satellite Communication Channel Based on RLNN

Yongxian Yu*, Luyong Zhang
School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, 100876, China
*Corresponding author’s e-mail: yuyongxian2018@bupt.edu.cn

Abstract. This paper discusses the potential applications of reinforcement learning in communication systems. In order to improve the spectral efficiency of wireless communication channels, we propose an adaptive modulation scheme based on reinforcement learning and deep neural network (RLNN) with a average network exploration (AE) strategy. Although the application of machine learning technology in link adaptation has attracted widespread attention, most of the solutions currently proposed are based on offline training algorithms, which are not suitable for real-time operations. Classical reinforcement learning is only effective in discrete actions and states. The proposed technical solution does not depend on the offline training mode. The proposed technical solution uses AE strategy to drive exploration, which can improve the exploration efficiency of the agent. The simulation result shows that we can increase the data transmission rate under limited bandwidth and independently improves link availability, throughput and the efficiency of the wireless communication system.

1. Introduction
Adaptive Modulation and Coding (AMC), as an alternative to improve the throughput of wireless communication systems, has aroused great interest. AMC uses the knowledge of channel state information (CSI) to adjust transmission parameters to maximize link throughput. In order to achieve this goal, machine learning algorithms can be applied to automatically calculate decision tasks[1]. Under different transmission strategies, the transmission performance of effective capacity of cognitive radio network is demonstrated[2]. The performances of cognitive radio system have been studied with primary user emulators existing[3]. The performance of data transmission in cognitive wireless networks is studied[4]. The RL model selects the appropriate modulation sequence and encoding rate in the underwater communication channel. At the same time, the RL scheme proposed in this paper can maximize the data throughput in [5] by determining the modulation sequence and encoding rate within the SNR range of the OFDM system.

2. Adaptive communication model
We need to understand transmitters and receivers and their feedback paths. Modulation mode is selected based on channel environment. The transmitter sends a pilot signal to the receiver to estimate the CSI. The estimator extracts envelope and phase information from pilot signals. When the channel information is estimated, the phase and envelope are estimated and compensated by the ideal phase discriminator and gain controller. The receiver then sends the CSI back to the transmitter via a feedback path.
We consider the Rayleigh-faded channel model over wireless communication channels, which is typically applied for mobile systems with dominant multipath component [6]. The SNR distribution of the Rayleigh-faded channel model can be obtained as [6]:

\[ p_r(\gamma) = \frac{1}{\bar{\gamma}} \exp\left(-\frac{\gamma}{\bar{\gamma}}\right), \quad \gamma > 0 \]

(1)

where \( \gamma \) is the instantaneous SNR per symbol by \( \gamma = (a^2 E_s)/N_0 \). \( \bar{\gamma} \) represent the average SNR per symbol. It is obtained as \( \bar{\gamma} = (\Omega E_s)/N_0 \), \( \alpha \) represents the fading amplitude. \( \Omega = \bar{a}^2 \), which is the square of the average fading amplitude. \( E_s \) represents the energy of per symbol. \( N_0 \) is the one-sided power spectral density of additive white Gaussian noise (AWGN), and the unit is W/Hz.

3. Reinforcement learning model

Cognitive radio [7] learn its own wireless channel according to the surrounding environment. In this case, intelligent algorithms are the key to successfully solving this problem. The cognitive engine uses this knowledge to adjust its decisions on transmitter parameters to achieve the desired performance goals.

By choosing the actions to be taken, and then perceiving the effects of these actions with the environment, these effects will be transformed into new states and reward functions.

Reinforcement learning theory can be formalized in terms of Markov theory, which can be defined as a quaternion: \( S = \{ s_1, s_2, \ldots, s_n \} \) represents a set of \( N \) possible states describing environment dynamics, \( A = \{ a_1, a_2, \ldots, a_m \} \) represents a finite set of \( M \) possible behaviors for selection, \( P: S \times A \times S \) is a Markov state transition matrix, \( P(s, a, s') \) represents the transition probability matrix of the state after taking the action, \( R: S \times A \times S \to R \) represents the reward return function, \( R(s, a, s') \) represents the reward of environmental feedback after taking action.

In the algorithm, \( s_{t+1} \) is the state at time \( t + 1 \), \( s_t \) is the state at time \( t \) and \( a_t \) represents the action to be taken according to the optimal policy. We know from MDP’s model that the reward function is independent of time.

A good policy defines a mapping from state to behaviour. The \( V^\pi(s) \) represents the status value. It is the value of the reward function calculated by the state according to the optimal strategy [8]. The value of the policy is defined as:

\[ V^\pi(s) = E \left\{ \sum_{t=0}^{\infty} \gamma^t r_t \right\} \]

(2)

\( r_t \) is the value of the reward at time \( t \), \( 0 \leq \gamma \leq 1 \) is a discount factor. The discount factor determines the importance of future rewards. The small value of \( \gamma \) means that the current reward is considered more. The big value of \( \gamma \) means that the long-term returns is considered more. In order to maximize the value of the defined reward function, we need to find an optimal strategy for each state.

A more convenient way to describe a policy is to use Q-function instead of V-function. The Q function indicates how good it is to follow the optimal strategy in the current state [9]. It gives the return obtained for an agent taking a given action from a given state, and then following strategy \( \pi \). It is defined as

\[ Q^\pi(s, a) = E \left\{ \sum_{t=0}^{\infty} \gamma^t r_t | s_t = s, a_t = a \right\} \]

(3)

It can be proven that the optimal Q-function, \( Q^*(s, a) \), is the one that satisfies \( Q^*(s, a) = \max_{\pi} Q^\pi(s, a) \). As a consequence, the value of the optimum policy is \( V^*(s) = \max_{\pi \in \Pi} Q^*(s, a) \).
\( Q^*(s, a) \) is known, the optimal policy can be determined by taking the action with the highest value among \( Q^*(s, a) \) for each state

\[
\pi^*(s) = \max Q^*(s, a), a \in A
\]  

(4)

When the transfer model of the environment is known, the optimal strategy can be obtained by using dynamic programming and other techniques to solve the nonlinear equations. Reinforcement learning is independent of the environment model. In this case, the environment needs to be explored to query the model, which is achieved through algorithms such as SARSA and Q-Learning [10].

In this study, we use Q-learning algorithm to solve reinforcement learning problem. It finds \( Q^*(s, a) \) recursively using the 4-tuple \((s, a, s', r)\), where \( s \) and \( s' \) are the states at time \( t \) and \( t+1 \), \( a \) is the action taken when in \( s \) and \( r \) is the immediate reward due to taking \( a \) at \( s \). The updating rule is:

\[
Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max Q(s', a) - Q(s, a)]
\]

(5)

where \( \alpha \) is the learning rate.

4. Adaptive modulation scheme based on RLNN

The modulation schemes we use are BPSK, QPSK, 8PSK and 16PSK. Each modulation sequence is assigned to a corresponding rate region \([y_k, y_{k+1}]\). When an instantaneous SNR \( \gamma \) falls into the region, its assigned modulation order is selected. The spectral efficiency (SE) of the proposed AM scheme can be obtained as

\[
\frac{D}{B} = \sum_{k=0}^{N-1} m_k \int_{y_k}^{y_{k+1}} p_c(\gamma) d\gamma
\]

(7)

where \( D/B \) is the average data rate per unit bandwidth.

We consider a BER approximation function in the AM scheme. The average BER of fading channel can be described as [11]

\[
P_e = \int_0^\infty P_e(\gamma)p_e(\gamma) d\gamma
\]

(8)

where \( P_e(\gamma) \) represents the BER of Gaussian channel. In Gaussian channel, the BER of MPSK can be obtained as

\[
P_e(\gamma) \approx \frac{2}{\max(\log_2 M, 2)} \sum_{k=1}^{K} Q_G \left( \sqrt{2\gamma \log_2 M \times \sin \frac{(2i-1)\pi}{M}} \right)
\]

(9)

where \( Q_G(\cdot) \) is the Gaussian Q-function and \( K = \max(4/M, 1) \). Substitute (9) into (8), and use (1) to calculate the average BER of each rate region.

RLNN is known to have the following advantages. Firstly, RLNN algorithm combines RL and deep neural network, using the concept of virtual exploration to reduce the exploration time of agents. By setting the network's output as an exploration option, the neural network allows all actions to be explored simultaneously. Secondly, RLNN algorithm eliminates the correlation between input sequences through the experience replay strategy, which makes the learning process of agent more stable. Further, the RLNN algorithm separates the target network from the main network and removes the correlation between the value of Q function and the target value, which are expressed by \( r(s, a) + d \max Q(s', a') \). It may prevent both the Q-function and the target value in (5) from being updated. Using these properties, the RLNN algorithm solves the instability problem of RL [12].

The RLNN algorithm use the following loss function to update the parameters at iteration i
\[ L_i(\theta_i) = \left[ Q(s, a; \theta_i) - \{ r(s, a) + \max_a Q^*(s', a'; \theta_i^-) \} \right]^2 \]  

(10)

where \( Q^*(\cdot) \) is the Q-function of the target network, and \( \theta_i \) and \( \theta_i^- \) are the network parameters of the main and the target networks, respectively.

Each iteration is the exploration of multiple actions. At the end of the iteration, the maximum spectral efficiency and the corresponding state in the iteration process will be saved in the state buffer. The agent selects a small batch of random samples from the experience pool, uses the small batch samples to determine the target value, and then optimizes through the loss function to update the network parameter \( \theta \).

Before each iteration, the agent averages the previously learned network parameters to obtain new network parameters, as follows

\[
W = \left[ \begin{array}{c} w^1_p = \frac{1}{K} \sum_{k=1}^{K} w^1_k \\ \vdots \\ w^n_p = \frac{1}{K} \sum_{k=1}^{K} w^n_k \end{array} \right] 
\]

(11)

\[
B = \left[ \begin{array}{c} b^1_p = \frac{1}{K} \sum_{k=1}^{K} b^1_k \\ \vdots \\ b^n_p = \frac{1}{K} \sum_{k=1}^{K} b^n_k \end{array} \right] 
\]

(12)

\( w^p_n \) is the nth layer neural network weight network parameter matrix of the kth network parameter, \( b^p_n \) is the nth layer neural network bias parameter matrix of the kth network parameter. \( W \) is the weight parameter matrix, \( B \) is the bias parameter matrix, \( K \) is the number of network parameters. The formula for selecting an action is as follows \( a_t = \arg \max_{a} Q_p(s_t, a; \theta^P_t) \) in AE strategy.

![Figure 1. The proposed RLNN structure with AE strategy](image)

In Figure 1, the online value network parameters obtained in different periods are mostly different. Average these online value network parameters to obtain a new neural network parameter. This neural network parameter is different from the latest network parameter, and the small changes in the network will be more in the future. This time step causes continuous, internal, and complex changes. This change may make it change a lot when selecting actions in the next stage, thereby improving exploration capabilities.

5. Simulation Results

In this section, the performance of the link adaptive scheme using reinforcement learning is evaluated by simulation. In the simulation settings, we make the following assumptions: We use 4-layered fully connected network structure. The number of network parameters, \( K \) is set to be 20. The value of
discounting factor, \( d \) is determined to be 0.9. The size of mini-batch is set to be 10. The value of \( \epsilon \) will decrease from 1 to 0.01.

In Figure 2, the simulation results of the maximum spectral efficiency under different strategies are compared: The AE strategy, the BER constraint strategy [12]. In the BER constraint strategy, \( \epsilon - \)greedy strategy selects action. The size and attenuation of \( \epsilon \) need to be manually adjusted, and improper adjustment will lead to poor performance. For AE/BER constraints, the maximum stride length count is set to 100. The switching step, which is the amount of the state change by an action, is set to be 0.05dB for the AE/BER constraint. As seen in Figure 2, the maximum spectral efficiency on episodes of different strategies are showed when SNR is 10dB, 20dB. Compared with the BER constraint strategy, the AE strategy can achieve the maximum spectral efficiency in the case of fewer episodes. It is observed that the AE strategy outperform the BER Constraint strategy. The spectral efficiency of the system is different in different SNR environments. With the increasing of SNR, the spectral efficiency of the system is also increasing. AE strategy has a significantly faster convergence speed than the BER Constraint.

![Figure 2. The maximum SE on episodes of different strategies in different SNR](image1)

The simulation results of the maximum spectral efficiency of Rayleigh fading channel are given in Figure 3. The performance comparisons between the AE strategy and the BER Constraint are shown. As seen in Figure 3, the maximum spectral efficiency can be improved compared with the BER Constraint. It can be seen from the figure that after the number of average SNR increases, the maximum spectral efficiency obtained by AE strategy is higher than BER Constraint strategy.

![Figure 3. The maximum SE over Rayleigh fading channel](image2)
6. Conclusion

In this paper, we propose an AM scheme based on RLNN with an average network exploration (AE) strategy. The goal of the AE strategy is to achieve faster convergence. With the AE strategy, this scheme can significantly improve spectral efficiency. Simulation results show that the spectral efficiency can be improved compared with the BER Constraint. We can find its application in systems that need to transfer large amounts of data with limited bandwidth.

References

[1] P. V. R. Ferreira et al, “Multi-objective reinforcement learning for cognitive satellite communications using deep neural network ensembles,” IEEE J. Sel. Areas Commun., vol. 36, no. 5, pp. 1030–1041, May 2018.
[2] L. Zhang, Y. Yang, X. Li, J. Chen and Y. Chi, ”Effective capacity in cognitive radio networks with relay and primary user emulator,” in China Communications, vol. 16, no. 11, pp. 130-145, Nov. 2019.
[3] Yang Y , Zhang L , Li X , et al. Detection Performances and Effective Capacity of Cognitive Radio with Primary User Emulators[J]. Iet Communications, 2018.
[4] L. Zhang, Y. Yang and J. Chen, ”Cognitive Communication in Link-Layer Evaluation Based Cellular-Vehicular Networks,” in IEEE Access, vol. 6, pp. 75293-75301, 2018.
[5] J. P. Leite, P. H. P. de Carvalho, and R. D. Vieira, “A flexible framework based on reinforcement learning for adaptive modulation and coding in OFDM wireless systems,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Shanghai, China, Apr. 2012, pp. 809–814.
[6] M. K. Simon and M.-S. Alouini, Digital Communication Over Fading Channels: A Unified Approach to Performance Analysis. Hoboken, NJ, USA: Wiley, 2000.
[7] Q. Zhao, L. Tong, A. Swami, and Y. Chen, “Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework,” IEEE J. Sel. Areas Commun., vol. 25, no. 3, pp. 589–600, Mar. 2007.
[8] E. Alpaydin, Introduction to Machine Learning, 2nd ed. MA: MIT Press, 2010.
[9] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Upper Saddle River, New Jersey: Prentice Hall, 2009.
[10] X. Song and J. Cheng, “Quantifying the accuracy of high SNR BER approximation of MPSK in fading channels,” in Proc. 27th Biennial Symp. Commun. (QBSC), Kingston, ON, Canada, Jun. 2014, pp. 105–108.
[11] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
[12] V. Mnih et al., ”Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, Feb. 2015.