

Deep Reinforcement Learning Empowered Rate Selection of XP-HARQ

Da Wu, Jiahui Feng, Zheng Shi, Hongjiang Lei, Guanghua Yang, and Shaodan Ma

Abstract—The complex transmission mechanism of cross-packet hybrid automatic repeat request (XP-HARQ) hinders its optimal system design. To overcome this difficulty, this letter attempts to use the deep reinforcement learning (DRL) to solve the rate selection problem of XP-HARQ over correlated fading channels. In particular, the average throughput (LTAT) of XP-HARQ is maximized by properly choosing the information rate of each HARQ round of the based on the outdated channel state information (CSI) available at the transmitter. This rate selection problem is first converted into a Markov decision process (MDP), which is then solved by applying the deep deterministic policy gradient (DDPG) algorithm. The simulation results finally demonstrate the superiority of the proposed XP-HARQ scheme over the conventional HARQ with incremental redundancy (HARQ-IR) and XP-HARQ with only static CSI.

Index Terms—Cross-packet hybrid automatic repeat request (XP-HARQ), deep reinforcement learning (DRL), outdated channel state information, rate selection.

I. Introduction

Hybrid automatic repeat request (HARQ) is one of the key technologies that is capable of offering reliable transmissions. However, this benefit is essentially reaped at the price of large transmission delay, which is unfavorable for fulfilling the ultra-reliable and low-latency communications (URLLC). To resolve such a dilemma, there is a urgent need to develop a flexible HARQ transmission mechanism that could be reconfigurable to meet diverse URLLC requirements. In this letter, we focus on the cross-packet HARQ (XP-HARQ) that is an evolutionary version of HARQ with high spectral efficiency, albeit at the price

of high complexity [?], [?], [?]. Unlike the conventional HARQ schemes, new information bits are introduced in retransmissions such that surplus wireless resources are substantially exploited. Hence, it is unnecessary to wait for the end of the retransmissions of the current message before the delivery of the next message especially under benign channel conditions. As a consequence, the spectral efficiency of HARQ is boosted, meanwhile the average transmission delay is reduced.

Recently, the investigations on the XP-HARQ scheme are still in their fancy. Several efforts have been made to accurately evaluate and optimally design XP-HARQ schemes. In [?], Mohammed Jabi et al. examined the long term average throughput (LTAT) of XP-HARQ, with which the throughput improvement gained by XP-HARQ was verified. In [?], a two-layer coding scheme was developed to implement XP-HARQ to guarantee the inputs of the encoder with the same length, where puncturing and mixing operations were leveraged. The puncturing rates were then optimized with dynamic programming in [?]. The adaptive modulation and coding scheme was further introduced to boost the LTAT of XP-HARQ in [?]. In [?], the effective capacity of XP-HARQ was analyzed for buffer-limited XP-HARQ. However, the performance metrics of XP-HARQ in [?], [?], [?], [?] were obtained by conducting Monte-Carlo simulations and lacked insightful analysis. To fill this vacancy, the most fundamental performance metric, namely, outage probability, was derived in closed-form for XP-HARQ over independent Rayleigh fading channels in [?], with which full time diversity of XP-HARQ was proved. However, even under such a simple channel model, the outage analysis is too complex to further assist the optimal design of XP-HARQ, not to mention under more complicated fading channels.

To address the above issue, we resort to the data-driven deep reinforcement learning (DRL) for the optimal design of XP-HARQ over correlated fading channels. It should be noticed that only a few works attempted to devise the conventional HARQ schemes using the DRL methods. Particularly, in [?], a DRL enabled user scheduling policy was designed to minimize the age of information (AoI) for HARQ systems. In [?], a deep deterministic policy gradient (DDPG) algorithm was leveraged to maximize the throughput via optimizing the incremental redundancy bits. Unfortunately, the extension of the DRL methods to general HARQ schemes has never been reported. This letter maximizes the LTAT via adaptive rate selection by considering outdated channel state information (CSI). The

This work was supported in part by National Natural Science Foundation of China under Grants 62171200, 62171201, 61971080, and 62261160650, in part by Chongqing Key Laboratory of Mobile Communications Technology under Grant cqjpt-mct-202204, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010900, in part by Zhuhai Basic and Applied Basic Research Foundation under Grant ZH22017003210050PWC, in part by the Major Talent Program of Guangdong Provincial under Grant 2019QN01S103, and in part by the Science and Technology Development Fund, Macau SAR under Grants 0087/2022/AFJ and SKL-IOTSC(UM)-2021-2023. (Corresponding Author: Zheng Shi.)

Da Wu, Jiahui Feng, Zheng Shi, and Guanghua Yang are with the School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China (e-mails: 0x8a@stu2021.jnu.edu.cn; jiahui@stu2020.jnu.edu.cn; zhengshi@jnu.edu.cn; ghyang@jnu.edu.cn).

H. Lei is with Chongqing Key Lab of Mobile Communications Technology & Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: leihj@cqupt.edu.cn).

Shaodan Ma is the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China (e-mail: shaodanma@um.edu.mo).

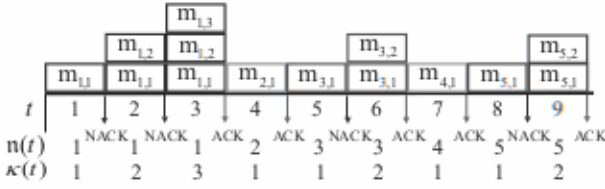


Fig. 1. An example of the XP-HARQ scheme with $K = 3$.

optimization problem is firstly formulated as a problem of Markov decision process (MDP). By taking into account the continuous state and action spaces, the problem is then solved by using DDPG with prioritized experience replay. By conducting Monte Carlo simulations, the proposed XP-HARQ scheme is proved to be superior to the conventional HARQ with incremental redundancy (HARQ-IR) and the XP-HARQ with only statistical CSI. Furthermore, it is found that the time correlation among fading channels does not lead to a significant impact upon the LTAT of the proposed XP-HARQ scheme.

The rest of this letter is outlined as follows. Section ?? introduces the system model. Section ?? develops a DRL empowered rate selection algorithm for XP-HARQ. The simulated results are presented in Section ?. Section ?? finally concludes this letter.

II. System Model

This letter considers a point-to-point communication system, in which XP-HARQ is adopted to enable the retransmissions of the message. To start, this section delineates the system model, including the XP-HARQ transmission mechanism, the channel model, performance metrics, and the rate selection problem.

A. XP-HARQ

As shown in Fig. ??, an example is used to illustrate the transmission mechanism of the XP-HARQ. To avoid network congestion in unfavorable propagation environment, the number of transmissions of XP-HARQ is limited up to K . For notational simplicity, let $n(t) \in \mathbb{Z}^+$ and $\kappa(t) \in [1, K]$ be the functions that map the time slot t to the current HARQ cycle and the current transmission round, respectively. In the initial transmission round of the $n(t)$ -th HARQ cycle, the message $m_{n(t),1}$ is encoded as a codeword $\mathbf{x}_{n(t),1}$ with a transmission rate R_1 . The received signal $\mathbf{y}_{n(t),1}$ reads as

$$\mathbf{y}_{n(t),1} = \sqrt{P_1} h_{n(t),1} \mathbf{x}_{n(t),1} + \mathbf{n}_{n(t),1}, \quad (1)$$

where $h_{n(t),1}$ denotes the channel coefficient of the first round of the $n(t)$ -th HARQ cycle with $\mathbb{E}(|h_{n(t),1}|^2) = 1$, $\mathbf{n}_{n(t),1}$ stands for the complex additive Gaussian noise (AWGN) having zero mean and a variance of σ^2 , and P_1 is the average transmit power in the initial HARQ round. If $\mathbf{x}_{n(t),1}$ is successfully decoded, a positive acknowledgement (ACK) will be sent back to confirm the successful reception of $m_{n(t),1}$ and the next HARQ cycle with index $t+1$ will be triggered immediately. Otherwise,

a negative acknowledgement (NACK) will be fed back to initiate the retransmissions. According to the coding strategy of XP-HARQ [?], as opposed to the conventional HARQ-IR that only redundant information bits are retransmitted, new information bits are introduced in the retransmissions by XP-HARQ to substantially exploited wireless resources. Accordingly, prior to the $\kappa(t)$ -th transmission of the $n(t)$ -th HARQ cycle, the previously failed messages $m_{n(t),1}, \dots, m_{n(t),\kappa(t)-1}$ are combined with the currently received message $m_{n(t),\kappa(t)}$ to form a longer message $m_{n(t),[\kappa(t)]}$. The concatenated message $m_{n(t),[\kappa(t)]}$ is encoded as a codeword $\mathbf{x}_{n(t),\kappa(t)}$ with a nominal transmission rate $\sum_{\kappa=1}^{\kappa(t)} R_{\kappa} \triangleq R_{\kappa(t)}^{\Sigma}$, where the increment of the transmission rate, i.e., R_{κ} , originates from the new information bits involved in the κ -th transmission. Therefore, the signal $\mathbf{y}_{n(t),\kappa(t)}$ received in the $\kappa(t)$ -th round of the current XP-HARQ cycle is written as

$$\mathbf{y}_{n(t),\kappa(t)} = \sqrt{P_{\kappa(t)}} h_{n(t),\kappa(t)} \mathbf{x}_{n(t),\kappa(t)} + \mathbf{n}_{n(t),\kappa(t)}, \quad (2)$$

where $h_{n(t),\kappa(t)}$, $\mathbf{n}_{n(t),\kappa(t)}$, and $P_{\kappa(t)}$ follow the similar definitions as $h_{n(t),1}$, $\mathbf{n}_{n(t),1}$, and P_1 , respectively, which are omitted here to save space. The messages $m_{n(t),1}, \dots, m_{n(t),\kappa(t)}$ are jointly decoded by using the observations $y_1, \dots, y_{\kappa(t)}$. The current XP-HARQ cycle stops and the next process begins once the receiver succeeds in reconstructing all the previously delivered messages or the maximum number of HARQ transmission attempts $\kappa(t)$ is used. Interested readers are referred to [?] for more details of the encoding/decoding implementation of XP-HARQ.

B. Channel Model

This letter considers time-correlated Rayleigh flat-fading channels, where the channel keeps constant during each codeword transmission slot and changes time-dependently across consecutive transmission slots. We define t as the index of the time slot in the sequel. For notational simplicity, we use the notation h_t to represent $h_{n(t),\kappa(t)}$. As a commonly used time-correlated channel model that takes place in the environment of low-to-medium mobility, h_t is modeled according to a first-order Gauss-Markov process as [?], i.e.,

$$h_t = \rho h_{t-1} + \sqrt{1 - \rho^2} w_t, \quad (3)$$

where ρ is the correlation coefficient between h_t and h_{t-1} , $w_t \sim \mathcal{CN}(0, \sigma^2)$ denotes the channel discrepancy and is independent of h_{t-1} . In order to account for the impact of channel aging, the outdated channel state h_{t-1} is sent back to the transmitter.

C. Performance Metrics

1) Outage Probability: The outage probability is an essential performance metric for evaluating the system reliability. The outage probability of XP-HARQ is the probability of the event that the accumulated mutual

IV. Simulations and Discussions

In this section, simulated results are presented for verifications and discussions. For illustration, the system parameters are set as $\sigma^2 = 1$, $\rho = 0.4$, and $\bar{R} = 10$ bps/Hz unless otherwise specified. Besides, we assume equal power allocation for XP-HARQ, i.e., $P_1 = \dots = P_K$, and the average transmit signal-to-noise ratio (SNR) is defined as $P_1/\sigma^2 = \dots = P_K/\sigma^2 \triangleq \text{snr}$. To deploy the DDPG, both the actor and critic networks consist of one input layer, three hidden layers, and one output layer. The number of the neurons in the three hidden layers are 100, 50, and 30 neurons, respectively. The three hidden layers of both networks use "ReLU" activation functions. The output layer of the actor network invokes "sigmoid" activation function to restrict the transmission rate within \bar{R} , while the critical network does not leverage any activation function in the output layer. Both the actor and critical networks capitalize on the adaptive moment estimation (Adam) optimizer to update the network parameters, and the learning rates are set to $v = \alpha = 0.001$. Furthermore, we assume that the number of epochs in the training state is 100, the number of time slots in each epoch is 6000, the size of the prioritized replay buffer is $|\mathcal{M}| = 20000$, the mini-batch size is $|\mathcal{B}_t| = 512$. In addition, we assume that the weight of the soft update $\tau = 0.01$, the discount factor $\gamma = 0.9$, the extent of the correction $\beta = 0.5$, and the noise variance of the behavior policy $\psi^2 = 0.2$.

Fig. ?? depicts the LTAT performance of XP-HARQ versus of the average transmit SNR under different K . To exhibit the superiority of the proposed DRL-empowered rate selection scheme, two baseline HARQ schemes are used for comparison, including the conventional HARQ-IR [?] and the XP-HARQ with only statistical CSI (labeled as "S-CSI" in the figure) [?]. The results of XP-HARQ with S-CSI can be regarded as the worst performance limit of our proposed scheme. In the meantime, the ergodic capacity is incorporated for benchmarking purpose or as design guidelines. It is shown in Fig. ?? that the XP-HARQ scheme performs much better than the HARQ-IR scheme. For example, by fixing $\text{snr} = 35$ dB and $K = 5$, the XP-HARQ scheme achieves a higher LTAT than the HARQ-IR scheme by around 1.65 bps/Hz. It is also seen from Fig. ?? that the proposed XP-HARQ scheme with outdated CSI surpasses the XP-HARQ scheme with statistical CSI by around 0.15 bps/Hz. Moreover, as the maximum number of transmissions K increases from 3 to 5, a remarkable performance gain can be attained by both XP-HARQ schemes with the outdated CSI and the statistical CSI, whereas the HARQ-IR scheme achieves a negligible LTAT enhancement particularly at high SNR. This advantage of XP-HARQ attributes to new information bits introduced in retransmissions. Moreover, this merit also brings about a reduced transmission delay.

Fig. ?? investigates the impact of the time correlation coefficient on the LTAT given a fixed $\text{snr} = 20$ dB. Overall, it is not beyond our expectation that the time correlation has a detrimental effect on the LTAT. This is because

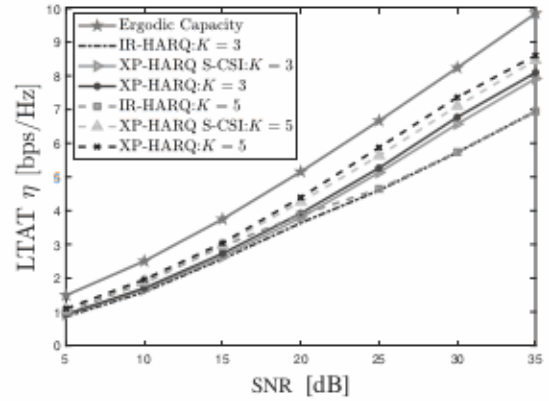


Fig. 3. The comparison of the LTAT for different HARQ schemes.

more time diversity gain can be achieved from fading channels with a lower time correlation [?]. Nevertheless, it is noteworthy that the superiority of the proposed XP-HARQ schemes essentially stems from utilizing the outdated CSI. Hence, a low channel correlation will result in less similarity of CSIs between two adjacent transmissions, which limits the time diversity gain from retransmissions. Accordingly, it can be seen from Fig. ?? that the LTAT curves slightly decrease with ρ .

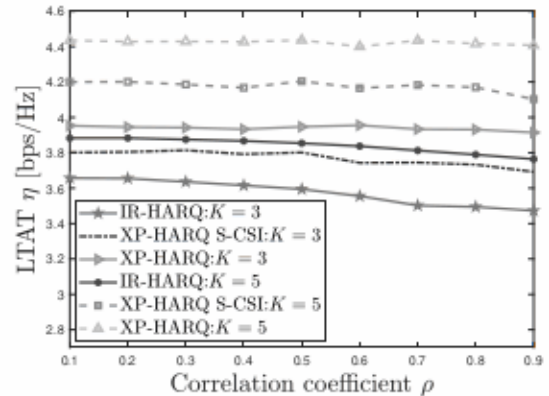


Fig. 4. Impact of correlation coefficient ρ .

V. Conclusion

Due to the lack of simple analytical results of the performance metrics of XP-HARQ, we applied the DRL to properly select the incremental information rate for XP-HARQ over correlated fading channels, without recourse to the traditional optimization tools. More specifically, the maximization of the LTAT was formulated as a problem of MDP, which can be solved by using the algorithm of DDPG with prioritized experience replay. To demonstrate the efficacy of the proposed XP-HARQ scheme, its LTAT performance was compared to the conventional HARQ-IR and the XP-HARQ with only statistical CSI through simulations. It was found that IR-HARQ is more aggressive than XP-HARQ when determining the initial rate. In the meantime, it was also found that the time correlation has