# 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 long term average throughput (LTAT) is maximized by properly choosing the incremental information rate for each HARQ round on the basis of the outdated channel state information (CSI) available at the transmitter. The rate selection problem is first converted into a Markov decision process (MDP), which is then solved by capitalizing on the algorithm of deep deterministic policy gradient (DDPG) with prioritized experience replay. The simulation results finally corroborate the superiority of the proposed XP-HARQ scheme over the conventional HARQ with incremental redundancy (HARQ-IR) and the XP-HARQ with only statistical 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*Index Terms*—Cross-packet hybrid automatic repeat request (XP-HARQ), deep reinforcement learning (DRL), outdated channel state information, rate selection.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, [illegible] of the current message [illegible]

[illegible] Mohammadi *et al.* examined the long term average throughput (LTAT) of XP-HARQ [illegible] XP-HARQ [illegible] the LTAT of XP-HARQ in [?] [illegible] closed-form [illegible] XP-HARQ over independent Rayleigh fading channels [illegible] optimal diversity of XP-HARQ was proved. [illegible]

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 [illegible], in part by [illegible] Basic and Applied Basic Research Foundation [illegible] Grant 2023A1515010900, [illegible] Major Talent Program of Guangdong Provincial [illegible] 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 [illegible] Jinan University, Zhuhai 519070, China (e-mail: [illegible]; [illegible]@stu2021.jnu.edu.cn; [illegible]).

H. Lei is with Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: [illegible]).

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

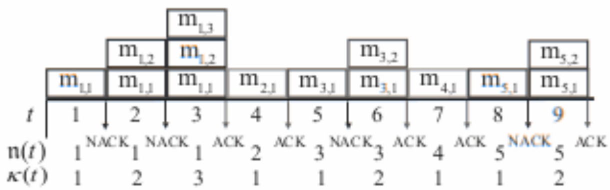

Fig. 1. An example of the XP-HARQ scheme with $K = 3$.

optimization problem is first formulated as a Markov decision process (MDP). By taking into account the continuous state and action spaces, the problem is 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 $\mathfrak{n}(t) \in \mathbb{Z}^+$ 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 $\mathfrak{n}(t)$-th HARQ cycle, the message $\mathrm{m}_{\mathfrak{n}(t),1}$ is encoded as a codeword $\mathbf{x}_{\mathfrak{n}(t),1}$ with a transmission rate $R_1$. The received signal reads as

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

where $h_{\mathfrak{n}(t),1}$ denotes the channel coefficient of the first round of the $\mathfrak{n}(t)$-th HARQ cycle with $\mathbb{E}(|h_{\mathfrak{n}(t),1}|^2) = 1$, $\mathbf{n}_{\mathfrak{n}(t),1}$ stands for the complex additive white Gaussian noise (AWGN) having zero mean and a variance of $\sigma^2$, and $P_1$ is the average transmit power in the initial HARQ round. If $\mathbf{x}_{\mathfrak{n}(t),1}$ is successfully decoded, a positive acknowledgement (ACK) will be sent back to confirm the successful reception of $\mathrm{m}_{\mathfrak{n}(t),1}$ 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 exploit wireless resources. Accordingly, prior to the $\kappa(t)$-th transmission of the $\mathfrak{n}(t)$-th HARQ cycle, the previously failed messages $\mathrm{m}_{\mathfrak{n}(t),1}, \cdots, \mathrm{m}_{\mathfrak{n}(t),\kappa(t)-1}$ are combined with the currently received message $\mathrm{m}_{\mathfrak{n}(t),\kappa(t)}$ to form a longer message $\mathrm{m}_{\mathfrak{n}(t),[\kappa(t)]}$. The concatenated message $\mathrm{m}_{\mathfrak{n}(t),[\kappa(t)]}$ is encoded as a codeword $\mathbf{x}_{\mathfrak{n}(t),\kappa(t)}$ with a nominal transmission rate $\sum_{\kappa=1}^{\kappa(t)} R_\kappa \triangleq R^{\Sigma}_{\kappa(t)}$, where the increment of the transmission rate, i.e., $R_\kappa$, originates from the new information bits involved in the $\kappa$-th transmission. Therefore, the signal $\mathbf{y}_{\mathfrak{n}(t),\kappa(t)}$ received in the $\kappa(t)$-th round of the current XP-HARQ cycle is written as

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

where $h_{\mathfrak{n}(t),\kappa(t)}$, $\mathbf{n}_{\mathfrak{n}(t),\kappa(t)}$, and $P_{\kappa(t)}$ follow the similar definitions as $h_{\mathfrak{n}(t),1}$, $\mathbf{n}_{\mathfrak{n}(t),1}$, and $P_1$, respectively, which are omitted here to save space. The messages $\mathrm{m}_{\mathfrak{n}(t),1}, \cdots, \mathrm{m}_{\mathfrak{n}(t),\kappa(t)}$ are jointly decoded by using the observations $y_1, \cdots, y_{\kappa(t)}$. 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 $h_{\mathfrak{n}(t),\kappa(t)}$. 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 $\rho$ 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 information in each HARQ round is below the transmission rate. More specifically, the outage probability of XP-HARQ after $K$ HARQ rounds is given by [?]

$$p_{out} = \Pr\left(I_1 < R_1, I_2 < R^{\Sigma}_2, \cdots, I_K < R^{\Sigma}_K\right), \quad (4)$$

where $I_\ell = \sum_{k=1}^{\ell} \log_2(1 + P_k |h_k|^2/\sigma^2)$ is the accumulated mutual

2) *Long Term Average Throughput:* The long term average throughput (LTAT) is a frequently used performance metric to evaluate the expected throughput of HARQ systems [?]. The LTAT of XP-HARQ system is defined as [?]

$$f_K = \Pr\left(I_1 < R_1, I_2 < R_2^{\Sigma}, \cdots, I_K < R_K^{\Sigma}\right), \quad (4)$$

where $I_k = \sum_{l=1}^{k} \log_2(1 + |h_l|^2 P_l/\sigma^2)$ stands for the accumulated mutual information until the $l$-th transmission.

$$\eta_K = \lim_{T\to\infty} \frac{\mathcal{R}(T)}{T} = \frac{\sum_{k=1}^{K} R_k (f_{k-1} - f_K)}{1 + \sum_{k=1}^{K-1} f_k} \quad (5)$$

where $\mathcal{R}(t)$ refers to the total number of successfully received information bits till time $t$, and the second equality in (??) is derived in [?], [?] by capitalizing on the renewal theory if only the statistical CSI is available at the transmitter.

*D. Maximization of LTAT*

This paper aims to maximize the LTAT through optimal rate selection if only the aged channel state information (CSI) is available at the transmitter. The optimization problem of the transmission rates can be formulated as

$$\max_{R_1,\cdots,R_K} \ \eta_K \quad \text{s.t.} \quad 0 \le R_k \le \bar{R},\ k \in [1, K], \quad (6)$$

where the transmission rate $\{R_k, k \in [1, K]\}$ is upper bounded by $\bar{R}$ to avoid frequent outages because of the limited resources. However, due to the time correlation among fading channels in (??) and the involved outage definition in (??), it is hardly possible to get the explicit outage expression. Hence, it is unlikely to solve the LTAT maximization problem in (??) with the conventional optimization tools. To overcome this difficulty, we recourse to the deep reinforcement learning (DRL) for the optimal solution of the transmission rate.

## III. DRL Empowered Rate Selection

Due to the rapid change of time-varying fading channels, it results in a prohibitively high system overhead to acquire the instantaneous CSI. Therefore, we assume that only the outdated and statistical CSIs are available at the transmitter, including the channel state of the previous slot $h_{t-1}$ and the correlation coefficient $\rho$. Moreover, the transmission rate of the current transmission round for XP-HARQ is determined by the transmission status (success or failure), rates, and channel states in the previous transmission rounds. Towards this end, the proposed optimization problem is transformed into a Markov decision process (MDP), which can be solved with DRL methods.

*A. Problem Reformulation and MDP*

By using the definition of the LTAT and replacing the limit operation with the expectation (the time average converges to the ensemble average for ergodic processes), the original problem (??) can be reformulated as

$$\max_{R(t)} \ \mathbb{E}\left(\frac{\mathcal{R}(T)}{T}\right) = \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{R}_{\mathfrak{n}(t),\kappa(t)}\right) \quad \text{s.t.} \quad 0 \le R(t) \le \bar{R}, \quad (7)$$

where the expectation is taken over the randomness of the channel states, $R(t)$ is the effective transmission rate for the new information bits in the time slot $t$, $\mathcal{R}_{\mathfrak{n}(t),\kappa(t)}$ denotes the effective transmission rate for the successfully received information bits after $\kappa(t)$ rounds during the $\mathfrak{n}(t)$-th HARQ cycle. According to the Shannon theory, the successful decoding occurs if and only if the transmission rate is less than the channel capacity. Therefore, $\mathcal{R}_{\mathfrak{n}(t),\kappa(t)}$ can be obtained as

$$\mathcal{R}_{\mathfrak{n}(t),\kappa(t)} = \begin{cases} R^{\Sigma}_{\kappa(t)}, & I_{\kappa(t)} \ge R^{\Sigma}_{\kappa(t)} \\ 0, & \text{else} \end{cases}. \quad (8)$$

With the problem reformulation of (??), the adaptive rate selection scheme can be modeled as an MDP, which can be solved by leveraging reinforcement learning (RL) method. The MDP essentially comprises four elements, including environment $\mathcal{E}$, state space $\mathcal{S}$, action space $\mathcal{A}$, and reward space $\mathcal{R}$. More specifically, at the time step $t$, the process is in state $s_t \in \mathcal{S}$. According to the current state, the agent makes a decision to choose an action $a_t \in \mathcal{A}$. After taking the action $a_t$, the next state $s_{t+1}$ is observed along with a reward $r_t \in \mathcal{R}$ received from the environment $\mathcal{E}$. By mapping the optimal rate selection of XP-HARQ as an MDP, the states, actions, and rewards are designed as follows.

1) *State $s_t$:* To capture the channel aging effect, the historical channel state $h_t$ is considered into the observation of environment. Moreover, the decoding status of XP-HARQ essentially depends on the accumulated mutual information and rate. Accordingly, the state $s_t$ is a vector consisting of the previously accumulated transmission rate and mutual information for the current HARQ cycle and the outdated channel state $h_{t-1}$, namely,

$$s_t \triangleq \begin{cases} \left(R^{\Sigma}_{\kappa(t-1)}, I_{\kappa(t-1)}, h_{t-1}\right), & \mathfrak{n}(t-1) = \mathfrak{n}(t) \\ (0, 0, h_{t-1}), & \text{else} \end{cases}, \quad (9)$$

wherein the accumulated transmission rate and mutual information for the current HARQ cycle are zero if a new HARQ cycle is initiated, i.e., $\mathfrak{n}(t-1) \neq \mathfrak{n}(t)$.

2) *Action $a_t$:* The action is defined as the effective transmission rate for the new information bits in the next HARQ round, i.e.,

$$a_t \triangleq R(t). \quad (10)$$

3) *Reward $r_t$:* The reward function can be defined as the effective transmission rate of the successfully received information bits for the current HARQ cycle $\mathfrak{n}(t)$, i.e.,

$$r_t = r(s_t, a_t, s_{t+1}) \triangleq \mathcal{R}_{\mathfrak{n}(t),\kappa(t)}. \quad (11)$$

By noticing the continuous space of the states and actions, the MDP problem can be solved with the DRL, which combines the reinforcement learning and deep neural networks, to learn the policy. The details are deferred to the next subsection.

*B. DRL Empowered Rate Selection*

A DRL-based rate selection scheme is proposed for the LTAT maximization of the XP-HARQ. By considering the continuous state and action spaces, a deep deterministic policy gradient (DDPG) with prioritized experience replay will be applied to develop the rate selection framework, as shown in Fig. ??. This framework consists of four neural networks, i.e., two policy networks (also termed as the actor network)

## B. DRL Empowered Rate Selection

A DRL-based rate selection scheme is proposed for the LTAT maximization of the XP-HARQ. By considering the continuous state and action spaces, a deep deterministic policy gradient (DDPG) with prioritized experience replay will be applied to develop the rate selection framework, as shown in Fig. ??. This framework consists of four neural networks, i.e., two policy networks (also termed as the actor network, i.e., $\mu(s_t;\boldsymbol{\theta})$ and $\mu(s_{t+1};\boldsymbol{\theta}^-)$) and two evaluation networks (also termed as the critic network, i.e., $Q(s_t,a_t;\boldsymbol{\omega})$ and $Q(s_{t+1},a_{t+1};\boldsymbol{\omega}^-)$), wherein the target-evaluation and target-policy networks are used to calculate the temporal-difference (TD) target to address the overestimation issue, and these neural networks are parameterized by $\boldsymbol{\theta}$, $\boldsymbol{\theta}^-$, $\boldsymbol{\omega}$, and $\boldsymbol{\omega}^-$. In addition, for the stability and fast convergence, a prioritized experience reply memory pool $\mathcal{M}$ is adopted to collect the agent's experience tuple $e_t = (s_i, a_i, r_i, s_{i+1})$ at each time $t$. At each time step, the four neural networks will be updated with a mini-batch of experience samples $\mathcal{B}_t$ that are drawn from $\mathcal{M}$ according to the priority of the playback experience, that is, $e_t \sim \mathcal{P}(\mathcal{M})$ for $\forall e_t \in \mathcal{B}_t$, where $\mathcal{P}$ is the probability function defined in (??). In what follows, priority experience playback mechanism and the training processes of the four neural networks are described in detail.

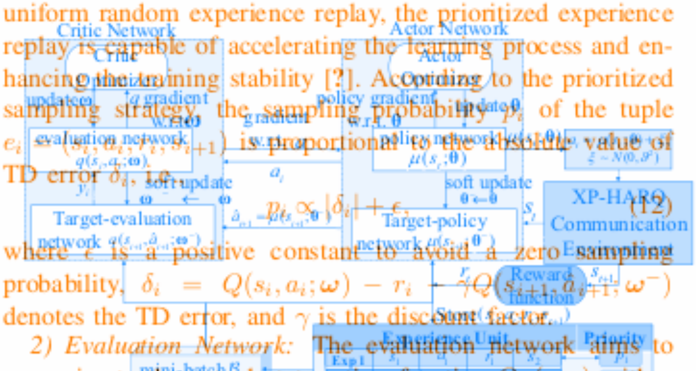


Fig. 2. The DDPG network for Rate Selection of XP-HARQ.

1) Prioritized Experience Replay: In contrast with the uniform random experience replay, the prioritized experience replay is capable of accelerating the learning process and enhancing the training stability [?]. According to the prioritized sampling strategy, the sampling probability $p_i$ of the tuple $e_i = (s_i, a_i, r_i, s_{i+1})$ is proportional to the absolute value of TD error $\delta_i$, i.e.,

$$p_i \propto |\delta_i| + \epsilon, \tag{12}$$

where $\epsilon$ is a positive constant to avoid a zero sampling probability, $\delta_i = Q(s_i, a_i;\boldsymbol{\omega}) - r_i - \gamma Q(s_{i+1}, a_{i+1};\boldsymbol{\omega}^-)$ denotes the TD error, and $\gamma$ is the discount factor.

2) Evaluation Network: The evaluation network aims to approximate the actual state-action function $Q_\pi(s,a)$ with a neural network parameterized by $\boldsymbol{\omega}$. The network parameters $\boldsymbol{\omega}$ can be updated with the TD algorithm. More specifically, the loss function is defined as the weighted squared TD error averaged over the sampled mini-batch $\mathcal{B}_t$, i.e.,

$$L(\boldsymbol{\omega}) = \frac{1}{2|\mathcal{B}_t|}\sum_{e_i \in \mathcal{B}_t} w_i \delta_i^2, \tag{13}$$

where $|\mathcal{B}_t|$ represents the batch size and the importance-sampling weight $w_i$ is used to eliminate the bias introduced by prioritized sampling and ensure the same learning rate of all samples. According to [?], $w_i$ is given by

$$w_i = (|\mathcal{B}_t| p_i)^{-\beta}, \tag{14}$$

which $\beta \in [0,1]$ is a hyperparameter that controls the extent of the correction. Then, the gradient descent algorithm is leveraged to update the network parameters $\boldsymbol{\omega}$ as

$$\boldsymbol{\omega}_{\text{new}} \leftarrow \boldsymbol{\omega}_{\text{now}} - \alpha \nabla_{\boldsymbol{\omega}} L(\boldsymbol{\omega}_{\text{now}}), \tag{15}$$

where $\nabla_{\boldsymbol{\omega}} L(\boldsymbol{\omega}) = \frac{1}{|\mathcal{B}_t|}\sum_{e_i \in \mathcal{B}_t} w_i \delta_i \nabla_{\boldsymbol{\omega}} Q(s_i, a_i;\boldsymbol{\omega})$ refers to the gradient of the loss function with respect to (w.r.t.) $\boldsymbol{\omega}$, and $\alpha$ is the learning rate.

3) Policy Network: The policy network $\mu(s_t;\boldsymbol{\theta})$ aims to learn action policy by mapping the states to the specific actions. Since the action-value function $Q_\pi(s,a)$ can evaluate the score of the current action policy, the performance objective for $\mu(s_t;\boldsymbol{\theta})$ can be defined as [?]

$$J(\boldsymbol{\theta}) = \frac{1}{|\mathcal{B}_t|}\sum_{e_i \in \mathcal{B}_t} Q(s_i, \mu(s_i;\boldsymbol{\theta});\boldsymbol{\omega}_{\text{now}}). \tag{16}$$

To learn the best policy, the parameters of the policy network can be optimized through the maximization of $J(\boldsymbol{\theta})$. Accordingly, the gradient ascend method is used to update $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{\theta}_{\text{new}} \leftarrow \boldsymbol{\theta}_{\text{now}} + \upsilon \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{\text{now}}), \tag{17}$$

where $\upsilon$ is the learning rate, and using chain rule yields $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{|\mathcal{B}_t|}\sum_{e_i \in \mathcal{B}_t} \nabla_{\boldsymbol{\theta}} \mu(s_i;\boldsymbol{\theta}) \nabla_a Q(s_i, a_i;\boldsymbol{\omega}_{\text{now}})$.

4) Target Evaluation/Policy Networks: To further improve the stability, the soft update strategy is applied to update the parameters of the target networks, i.e., $\boldsymbol{\omega}^-$ and $\boldsymbol{\theta}^-$. More specifically, with the new parameters $\boldsymbol{\omega}_{\text{new}}$ and $\boldsymbol{\theta}_{\text{new}}$ given by (??) and (??), respectively, the parameters of the two target networks will be updated as

$$\boldsymbol{\omega}^-_{\text{new}} \leftarrow \tau \boldsymbol{\omega}_{\text{new}} + (1-\tau)\boldsymbol{\omega}^-_{\text{now}}, \tag{18}$$

$$\boldsymbol{\theta}^-_{\text{new}} \leftarrow \tau \boldsymbol{\theta}_{\text{new}} + (1-\tau)\boldsymbol{\theta}^-_{\text{now}}, \tag{19}$$

where the hyperparameter $\tau \ll 1$.

## IV. Simulations and Discussions

In this section, simulated results are presented for verifications and discussions. For illustration, the system parameters are set as $\sigma^2 = 1$, $\rho = 0.1$, and $R = 16$ bps/Hz unless otherwise specified. Besides, we assume equal power allocation for XP-HARQ, i.e., $P_1 = \cdots = P_K$, and the average transmit signal-to-noise ratio (SNR) is defined as $P_1/\sigma^2 = \cdots = P_K/\sigma^2 \triangleq$ 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 $R$, 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 $\upsilon = \alpha = 0.001$. 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 $|\mathcal{M}| = 20000$,

## 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 $R = 10$ bps/Hz unless otherwise specified. Besides, we assume equal power allocation for XP-HARQ, i.e., $P_1 = \cdots = P_K$, and the average transmit signal-to-noise ratio (SNR) is defined as $P_1/\sigma^2 = \cdots = P_K/\sigma^2 \triangleq \mathrm{snr}$. 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 $R$, 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 $\mu = \alpha = 0.001$. 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 $|\mathcal{B}| = 512$. 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 $\vartheta^2 = 0.2$.

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 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.

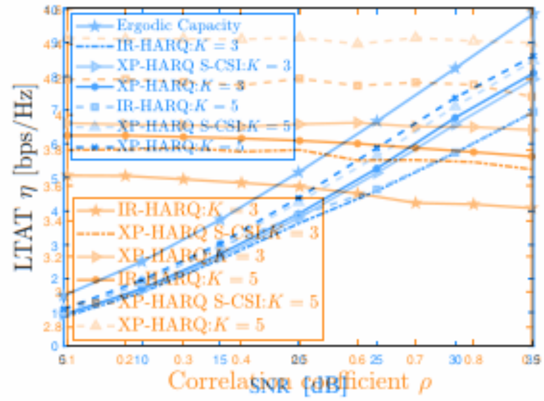


Fig. 3. The comparison of the LTAT for different HARQ schemes.

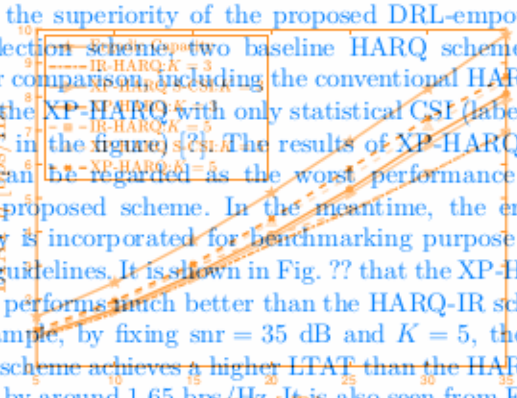

Fig. ?? investigates the impact of the time correlation coefficient on the LTAT given a fixed snr = 20 dB. Overall, it is not beyond our expectation that the time correlation has a detrimental effect on the LTAT. This is because more time diversity gain can be achieved from fading channels with a lower 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 $\rho$.

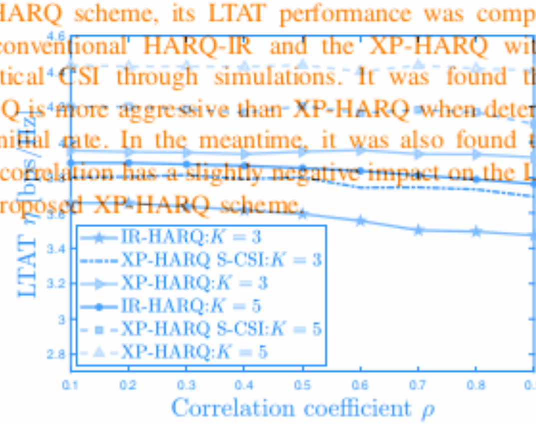


Fig. 4. Impact of correlation coefficient $\rho$.

## 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 a slightly negative impact on the LTAT of the proposed XP-HARQ scheme.