DRL-Based Energy Efficient Power Adaptation for Fast HARQ in the Finite Blocklength Regime

Xinyi Wu and Deli Qiao
School of Communication and Electronic Engineering, East China Normal University, Shanghai, China
Email: 51215904045@stu.ecnu.edu.cn, dlqiao@ce.ecnu.edu.cn

***Abstract*—In this paper, a point-to-point communication system with low latency and high reliability is studied. A fast hybrid automatic repeat request (HARQ) protocol is applied, where some HARQ feedback is omitted and the associated channel uses are incorporated for data transmission in fast HARQ. Based on relevant results on the decoding error probability over finite blocklength (FBL) codes, a long-term bit energy minimization problem is formulated in the presence of feedback delay and reliability constraints. Considering the non-convexity of the optimization problem and small decoding error probabilities, a finite-episode Markov Decision Process (MDP) with a double-layer penalty reward is formulated. An actor-critic based deep reinforcement learning (DRL) algorithm is subsequently designed. Through numerical evaluations, it is shown that compared with the conventional HARQ and the existing fast HARQ protocol, the proposed scheme is more energy efficient especially when the packet size is large.**

I. Introduction

Different from the typical human-centered communication traffic which prioritizes high throughput, the Internet of Things (IoT) networks in Fifth Generation (5G) wireless communications of numerous critical industries such as industrial automation, smart manufacturing, healthcare, and virtual reality have strict requirements for delay and reliability [1]. As one of the crucial application scenarios in 5G, ultra-reliable and low-latency communication (URLLC) is anticipated to provide reliable and low-latency communication services [2].

Combining forward error correction (FEC) and automatic repeat request (ARQ), hybrid ARQ (HARQ) is an essential protocol applied in wireless networks to enhance data transmission performance [3]. Nonetheless, it is also characterized by high latency with latency arising from multiple retransmissions and feedback [4]. The performance of HARQ protocols has been extensively studied in various works. The authors in [6] have designed strategy to improve the performance of a delay-sensitive communication system via finding adaptive transmission rates. In order to reduce transmission delay, short packet transmissions were considered in [7], where the Shannon capacity with infinite code length was no longer applicable and finite blocklength (FBL) analyses can provide more accurate characterizations.

Regarding the analyses in the FBL regime, some key works have considered HARQ for URLLC. In [8], different HARQ schemes, i.e., HARQ with incremental redundancy (HARQ-IR) and HARQ with chase-combined, were analyzed in URLLC systems, respectively. A trade-off between energy and latency in a HARQ-IR scheme for URLLC communication was studied in [9]. Besides, to further reduce the delay, the authors in [4] proposed an improved HARQ strategy in FBL regime without waiting for feedback on the basis of channel condition. Note that these models are commonly complicated due to non-convex optimization problems. High computing overheads are generally required to execute the algorithms.

Faced with challenges, deep reinforcement learning (DRL) has shown great potential for the analysis of URLLC systems. In [10], a DRL-based framework was developed for downlink URLLC systems to obtain maximum long-term throughput constrained by latency in NR-Unlicensed and WiFi coexistence systems. The authors in [11] studied a resource allocation problem in a joint eMBB and URLLC system, where a multi-agent DRL-based algorithm was proposed satisfying the reliability constraint and QoS requirement of URLLC and eMBB, respectively. The authors solved the spectrum measurement problem in an uncertain environment by combining a model-free DRL-based solution with a proactive dynamic spectrum sharing (PDSS) scheme in [12]. Nonetheless, the above works rarely involve transmission events with small transition probabilities.

In this paper, we consider a fast HARQ protocol in the finite blocklength regime with low transmission delay constraints, where the feedback delay is integrated. Compared with the conventional HARQ protocol, some feedback is omitted, where the associated channel uses among the feedback not utilized before can be involved for data transmission. Then, we formulate the problem as a long-term bit energy minimization problem in the presence of reliability, delay, and peak power constraints. We model the non-convex problem as a finite-episode MDP with a double-layer penalty reward function and propose an Advantage Actor-Critic (A2C) based algorithm to solve it. Particularly, considering the transmission events with small transition probabilities, a term related to transition probabilities is added to the reward function to facilitate better training results. Numerical results show that the proposed scheme can achieve better performance in terms of energy efficiency compared to the conventional HARQ and the existing fast HARQ protocol, especially when the packets carry considerable information bits.

II. Preliminaries

In this section, we briefly discuss the system model, the fast HARQ scheme and the decoding error probability. Additionally,

This work is supported in part by the National Natural Science Foundation of China (61671205, 61271204), and in part by the Shanghai Rising Star Program (21QA1402700).

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

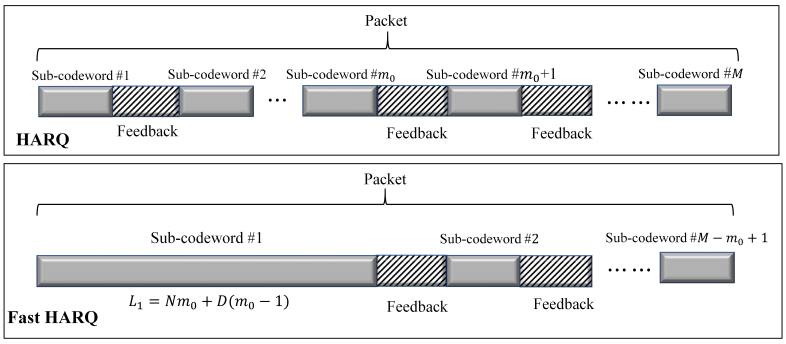

Fig. 1. Schematic of the packet transmission in the fast HARQ approach in comparison to the HARQ scheduling.

the problem formulation is given.

A. System Model

Throughout this paper, we assume a point-to-point block fading channel with single-antenna nodes with perfect channel state information (CSI) at the receiver only, where the fading coefficients stay constant for a coherence block of N symbols and change independently from one block to another. The relationship between the output and the input in the i^{th} block is given by

$$y_j = h_i x_j + z_j \quad j = 1, 2, \ldots, N, \tag{1}$$

where x_j and y_j are the channel input and output, respectively, in the j^{th} symbol duration of the i^{th} block, h_i denotes the fading coefficient, and z_j is the additive white Gaussian noise with zero mean and variance n_0, i.e., $z_j \sim \mathcal{CN}(0, n_0)$. In the following, we define $g_i = |h_i|^2$ as the channel gain and $G = \mathbb{E}[g(i)]$ as the expectation of the channel gain.

B. Fast HARQ

The comparison between the conventional HARQ scheme and the fast HARQ protocol is depicted in Fig. 1. Firstly, we consider a conventional fixed-rate HARQ scheme which requires a maximum number M of transmissions. Each packet of B information bits are encoded into a codeword of overall length MN channel uses, divided into M subcodewords. If the maximum number of transmissions is not reached, the receiver will respond with an ACK or NACK signal, depending on whether the decoding is successful or not. At this point, a NACK signal prompts another transmission. When the maximum transmission round is reached, the message will be dropped without requiring feedback. As a stark difference from most of prior works, we take the feedback delay in terms of D channel uses into account in this paper. Then, the traditional scheme would wait $D(M-1)$ channel uses for feedback, which are not utilized for data transmission.

Alternatively, for the fast HARQ scheme, it is assumed that the transmitter sends $m_0\,(m_0 \leq M)$ subcodewords that would take m_0 transmission rounds in the conventional HARQ scheme together in the first transmission round. Then, the channel uses of length $(m_0-1)D$ waiting for feedback can be incorporated for data transmission without affecting the transmission delay. Thus, for the fast HARQ, each packet of B information bits is encoded into a codeword of overall length $MN+(m_0-1)D$ channel uses. In the first transmission round, the subcodeword of length L_1 is sent to the receiver, where $L_1 = m_0 N + (m_0+1)D$. If decoding fails, the receiver sends a NACK signal, and another round of retransmissions is requested. In subsequent retransmission rounds, a subcodeword of length N is sent each time until the message is successfully decoded or the maximum number of transmission rounds is reached. Note that when $m_0 = 1$, the fast HARQ scheme reduces to the conventional HARQ scheme.

C. Decoding Error Probability

For the considered fast HARQ protocol, the subcodeword will pass through multiple fading blocks in the first transmission round. We assume that each time the feedback signal is transmitted through d fading blocks, where d is an integer and $D = dN$. Thus, during the first transmission round, the subcodeword is transmitted over $m_0 + (m_0-1)d$ blocks. Note that, if $d = 0$, the transmitter can receive the ACK/NACK signal from the receiver instantaneously without any feedback delay. It is worth noting that any other values of feedback can be quantized into integer multiples of N without affecting the subsequent analysis.

Define the transmit power of the k^{th} transmission round as p_k and the received signal-to-noise ratio of the specific block i duration of the k^{th} transmission round as γ_{k_i}. γ_{k_i} can be expressed as

$$\gamma_{k_i} = \frac{p_k g_i}{n_0}. \tag{2}$$

In the case of a single antenna and perfect CSI at the receiver only, the achievable coding rate is given by [13]

$$R \approx C - \sqrt{\frac{V(\gamma)}{L}} Q^{-1}(\epsilon), \tag{3}$$

where L denotes the sum number of channel uses which are used for coding, ϵ is the decoding error probability, and $Q^{-1}(.)$ is the inverse function of the complementary Gaussian cumulative distribution function $Q(.)$. We have

$$C(\gamma) = \mathbb{E}\{\log_2(1+\gamma)\}, \tag{4}$$

$$V(\gamma) = NVar(\log_2(1+\gamma)) + \frac{1}{\log_e^2 2}(V_{avg}(\gamma) + Var(\frac{\gamma}{1+\gamma})), \tag{5}$$

with $Var(X)$ representing the variance of random variable X, and

$$V_{avg} = \mathbb{E}\left\{\frac{\gamma(2+\gamma)}{(1+\gamma)^2}\right\}. \tag{6}$$

Considering the normal approximations [14], the decoding error probability of transmission round k can be approximately expressed as

$$\epsilon_k \approx Q\left(\frac{\sum_{l=1}^{k} n_l C_l - B}{\sqrt{\sum_{l=1}^{k} n_l V_l}}\right) \quad k \in [1, K], \tag{7}$$

where

$$C_l = \begin{cases} \frac{1}{m_0+(m_0-1)d}\sum_{i=1}^{m_0+(m_0-1)d}\log_2(1+\gamma_{k_i}), & l=1, \\ \log_2(1+\gamma_{k_i}) & \text{else.} \end{cases} \tag{8}$$

$$V_l = \begin{cases} \frac{1}{m_0+(m_0-1)d}\sum_{i=1}^{m_0+(m_0-1)d}\frac{\gamma_{k_i}(2+\gamma_{k_i})}{(1+\gamma_{k_i})^2}, & l=1, \\ \frac{\gamma_{k_i}(2+\gamma_{k_i})}{(1+\gamma_{k_i})^2}, & \text{else.} \end{cases} \tag{9}$$

Above, C_l and V_l represent the channel capacity and dispersion of the l^{th} transmission round, respectively, n_l is the length of the channel used for information coding in the l^{th} transmission round, and $K = M - m_0 + 1$ denotes the maximum number of transmission rounds in the proposed scheme.

D. Problem Formulation

Delay and reliability are two factors that should be taken into account considering the URLLC scenarios. In the following, we assume $N \times \Delta T = T_C$, where T_C represents the coherent time and ΔT is the symbol duration. D_{sum} represents the delay of the proposed transmission scheme, which indicates the sum number of channel uses in each packet transmission period including feedback. Denote ϵ_K as the decoding error probability of the maximum transmission round. Considering that in the proposed scheme, a transmission failure event only occurs in the maximum transmission round, and hence the reliability of the system can be represented by ϵ_K. Also, there exists a tradeoff between energy consumption and transmission reliability. Specifically, larger transmit energy can guarantee higher reliability, whereas low energy consumption is desired but may not guarantee the reliability constraints.

In view of the above considerations, we formulate the problem as follows

$$\min_{p(t)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} p(t)l(t)\Delta T \tag{10a}$$

$$\text{s.t.} \quad \epsilon_K \le \epsilon_{max}, \tag{10b}$$

$$D_{sum} \le D_{max}, \tag{10c}$$

$$0 < p(t) \le p_{max}. \tag{10d}$$

Problem (10) is a long-term energy minimization problem, where $l(t)$ denotes the length of channel uses for data transmission in the t^{th} step, the objective function in (10a) is the long-term average energy consumption of the continuous packet transmission process, (10b) and (10c) are the constraints of reliability and feedback delay, respectively, and (10d) is the peak power constraint for each transmission round. Since this optimization problem is non-convex and computationally intensive, we employ a DRL-based approach to solve it.

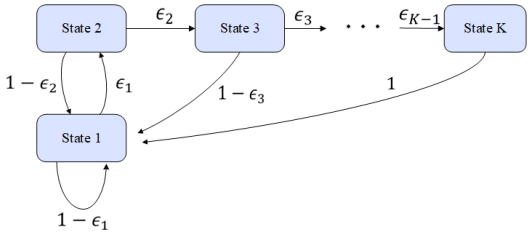

Fig. 2. The state transition model of Fast-HARQ.

III. Deep Reinforcement Learning Based Approach

In this section, we first formulate the problem as an Markov Decision Process (MDP) and then introduce an A2C-based algorithm to solve it.

A. MDP Formulation

To solve the optimal transmission power allocation problem of the fast HARQ scheme, we formulate a finite-episode MDP, where each episode consists of T steps that correspond to T subcodewords. The specific settings are as follows.

State space: The state space $\mathcal{S}$ is characterized by the number of transmission rounds, limited to the maximum number specified in the fast HARQ subsection. Thus, the state of each step can be given as

$$s(t) = \{k(t)|k(t) \in [1, K]\}. \tag{11}$$

Action space: The transmitter selects transmit power when each transmission attempt happens in slot t. Accordingly, the action of each step can be expressed as

$$a(t) = \{p(t)\}. \tag{12}$$

Transition dynamics: The state transition model of the proposed Fast-HARQ is depicted in Fig. 2. State 1 denotes the first transmission round in a fresh packet period, and State k denotes k^{th} transmission round, where the state transfers from State k to State $k+1$ with the decoding error probability ϵ_k when the maximum transmission round K is not reached. In accordance with the proposed transmission principle, State K enters State 1 with probability 1. Thus, the state transition matrix is given by

$$\mathbf{P} = \begin{bmatrix} 1-\epsilon_1 & \epsilon_1 & 0 & \dots & 0 \\ 1-\epsilon_2 & 0 & \epsilon_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1-\epsilon_{K-1} & 0 & 0 & \vdots & \epsilon_{K-1} \\ 1 & 0 & 0 & \vdots & 0 \end{bmatrix}, \tag{13}$$

where P_{ij} denotes the probability of state transition from State i to State j.

Reward function: The reward function plays a crucial role in learning performance and efficiency. It should be defined to be consistent with the design objective of minimizing long-term

average energy while satisfying the constraint of reliability. We define the normalized energy of each step

$$e(t) = \frac{p(t)l(t)}{N}. \tag{14}$$

Then, the reward function, including penalties for energy consumption and transmission failures, is expressed by

$$r(t) = \begin{cases} -e(t) * m(t), & \text{else}, \\ -e(t) * m(t) - v(t), & S^{'}(t) = \text{failure}, k(t) = K. \end{cases} \tag{15}$$

and

$$m(t) = \min\{-\log \epsilon_{k(t)}, \Delta_1\}, \tag{16}$$

$$v(t) = \min\{C(\frac{f(t)}{F})^{\alpha}, \Delta_2\}, \tag{17}$$

where $S^{'}$ denotes the decoding result of the current transmission round.

Specifically, we propose a function $m(t)$ related to the decoding error probability, which forms the penalty term of energy consumption combined with the normalized energy. For some transmission events with small transition probabilities, even though the relationships between transmit power and probabilities are known, it is difficult to exploit them due to the overall small values of the transition probabilities. In this section, the function $m(t)$ converts a small decoding error probability into a positive number that is easy to handle, and the value of it increases with transmit power until reaching the upper bound we set. To avoid excessive energy consumption while ensuring high reliability, $m(t)$ is used to increase the weight of energy consumption in reward function.

What's more, in terms of the penalty term for reliability, we construct a double-layer penalty term. The first layer penalty relies on the transmission result of the package, once a transmission failure event occurs, the penalty term emerges. Nevertheless, only focusing on this penalty will make the agent attempt to transmit successfully rather than ensuring the constraint of transmission reliability. Hence, the second layer penalty term in (15) is given, where C is a positive penalty coefficient and F is the maximum number of transmission failure events within the limitation of reliability, which is equal to $T\epsilon_{max}$. α is a positive integer controlling the increasing speed of the reliability penalty term with respect to the cumulative number of transmission failures, which is given by the function $f(t)$

$$f(t) = \sum_{i=1}^{t} \mathbb{I}[S^{'}(i) = \text{failure} \& k(i) = K], \tag{18}$$

where $\mathbb{I}\{.\}$ is the indicator function. Obviously, $\lim_{T\to\infty} \frac{f(t)}{t} = \epsilon_K$. $v(t)$ increases sharply as $f(t)$ increases to avoid transmission failure when $f(t)$ approaches or exceeds target number F.

In addition, Δ_1 and Δ_2 are constants which denote different upper bounds for both $m(t)$ and $v(t)$ to avoid excessive values for any penalty term. Hence, this reward function encourages the agent to consume less energy while satisfying the transmission reliability constraint.

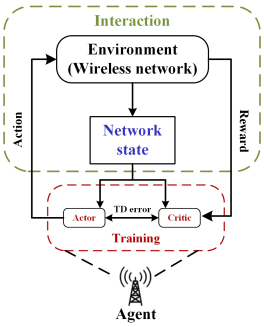

Fig. 3. Framework of the proposed A2C algorithm.

B. A2C-based Algorithm

In this part, we introduce the actor-critic algorithm A2C to solve the above MDP problem. In the following of this section, s_t, a_t and r_t are used to denote the state, action, and reward of the agent in each step t.

The framework of the proposed A2C algorithm is shown in Fig. 3. We can briefly summarize the algorithm into two parts: interaction and training. During the interaction, the agent receives the current state $s_t \in \mathcal{S}$ from the network environment and selects a $a_t \in \mathcal{A}$. Considering the transmission event with small transition probabilities in this page, we propose a more effective method for action exploration and evolution. Combining traditional methods for action exploration and the idea of the simulated annealing (SA) algorithm, we denote the dimension of power space n_a, exploration random integer of each episode $n_t \in [-n_a, n_a]$, random number $a \in [0, 1]$, and the total training episodes n_T. Then, T_e which indicates temperature in SA, is given by

$$T_e = b \exp\{\frac{-n_{episodes}}{c}\}, \tag{19}$$

where b, c are constants. If $a < Te$, we obtain the probability distribution of actions $p(s_t, a_n)$ by feeding s_t into the actor network $\pi(s_t|\theta_\pi)$. Obtain the index $i = argmax(p(s_t, a_n))$. Then, we add the exploration random integer of each episode to the index and limit the value of i within $[0, n_a - 1]$. If $a \geq Te$, we look for p_{avg} corresponding to the maximum average reward of T steps in the previous episodes. Round p_{avg} and obtain index value i in power space $\mathcal{P}$. Finally, the agent chooses a_t in $\mathcal{P}$ according to the index i.

Performing action a_t, the agent receives the reward r_t and reaches the next state s_{t+1}. The discount return reward at step t can be defined as $G_t = \sum_{T=0}^{\infty} \lambda^T r_{t+T}$, where λ is the discount factor. Define the action-value function $Q(s_t, a_t) = \mathbb{E}[G_t|s_t, a_t]$ and the state-value function $V(s_t) = \mathbb{E}[G_t|s_t]$, which estimate the expected return for selecting action a_t in state s_t and the average expected return from state s_t, respectively. The objective of the agent is to maximize the expected return from each state s_t, which can be estimated by $Q(s_t, a_t)$ and $V(s_t)$. The A2C has been proved to be an effective approach using only the state-value function $V(s_t)$, which reduces the number of parameters and simplifies the learning process. Particularly, the actor network

in A2C uses the advantage function to solve the gradient, which is defined as $A_t(s_t, a_t; \theta_\pi, \theta_v)$ and can be substituted by the TD error approximately, which is given by

$$\begin{aligned} A_t(s_t, a_t; \theta_\pi, \theta_v) &= Q(s_t, a_t) - V(s_t) = \mathbb{E}[G_t | s_t, a_t] - V(s_t) \\ &\approx r_t + \lambda V(s_{t+1}) - V(s_t) = \delta_t. \end{aligned} \tag{20}$$

In this way, the action is evaluated not only on how good the behavior is but also on how much it can be improved.

The training part is essentially the update and iteration of network parameters θ_π and θ_v, which parameterize the actor and critic networks, respectively. Adding average entropy to the loss function of the actor network, we have

$$\mathcal{L}_\pi = -\log \pi(a_t | s_t; \theta_\pi) \delta_t(\theta_v) - \rho E(\pi(a_t | s_t; \theta_\pi)), \tag{21}$$

where ρ is the weight of the average entropy. Then, the parameter θ_π is updated as

$$\theta_\pi \leftarrow \theta_\pi + \beta_\pi \nabla_{\theta_\pi} \mathcal{L}_\pi, \tag{22}$$

where β_π is the learning rate of actor network.

Considering the common MSE function as the loss function of the critic network, we have

$$\mathcal{L}_v = (r_t + \lambda V(s_{t+1}; \theta_v) - V(s_t; \theta_v))^2. \tag{23}$$

Then, the parameter θ_v is updated as

$$\theta_v \leftarrow \theta_v + \beta_v \delta_t(\theta_v) \nabla_{\theta_v} V(s_t; \theta_v), \tag{24}$$

where β_v is the learning rate of critic network.

The A2C-based power adaptation algorithm can be summarized as algorithm 1.

TABLE I
PARAMETERS OF NETWORKS

Parameters	Values
Actor network size	128×128
Critic network size	128×128
Activation function	ReLu
Training episodes	200
Batch size	128
Actor learning rate	10^{-6}
critic learning rate	10^{-5}
Reward discount factor λ	0.99

IV. Numerical Results

In the numerical results, we assume the fading distributions experience Rayleigh fading with $G = 1$, $n_0 = 1$ and $T_C = 1$ ms. Considering a target transmission probability $1 - \epsilon_{max} = 99.99\%$, we set $T = 1 \times 10^5$ to ensure the constraint of reliability in each training episode. We assume $M = 3, m_0 = 2$ and $N = D = 100$ channel uses, unless specified otherwise. The upper bounds of two penalty terms are $\Delta_1 = 10$ and $\Delta_2 = 2000$, respectively. The penalty coefficient $C = 50$ and $\alpha = 10$. The hyperparameters of

Algorithm 1 A2C-based power adaptation algorithm

1: **Initialization:** θ_π and θ_v;
2: **repeat**
3: Observe initial state s_t;
4: Choose action a_t according to the action exploration and evolution in section III-B;
5: Receive the reward r_t and the next state s_{t+1};
6: Calculate $V(s_{t+1})$ by feeding s_{t+1} into the value network;
7: **Update**
8: θ_π according to (19);
9: θ_v according to (21);
10: **until** The predefined maximum number of training episodes has been completed.

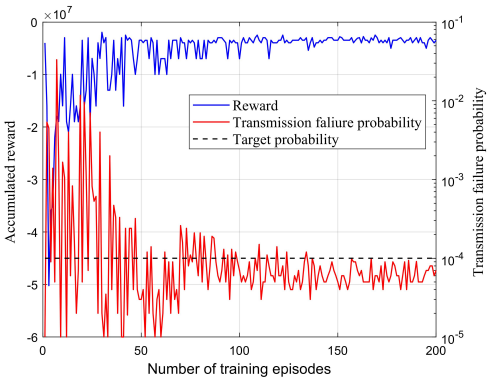

Fig. 4. Training curves of the network.

A2C networks are shown in table I. Hidden layers of the actor network and the critic network are fully connected structure, and the output layer of the actor network is set to softmax. The agent selects discrete actions in power space $\mathcal{P}$, where the elements in $\mathcal{P}$ are evenly distributed between 0 dBw and 30 dBw.

In Fig. 4, we plot the training curves during training. We assume $B = 2^7$ bits. From the figure, we can find that the training achieves convergence and the proposed algorithm can eventually satisfy the transmission reliability constraint.

In Fig. 5, we plot the comparison of bit energy as the packet size increases. We compare the proposed schemes with the conventional HARQ policy with power adaptation and the existing fast HARQ protocol with constant power. In this figure, we assume the same constraint of transmission reliability for different protocols. We can find that the proposed fast HARQ scheme is more energy efficient especially when the packet size is large.

In Fig. 6, we plot the energy efficiency versus m_0 under different feedback delay D. We assume $M = 5$ and $B = 2^{11}$ bits. It is interesting that in the presence of feedback delay, the fast HARQ scheme can achieve higher energy efficiency, especially for the case of $m_0 = 4$, i.e., only one retransmission round is allowed. Besides, we can see that the proposed scheme can achieve higher energy efficiency in the presence of larger feedback delay, since larger chunks can be expected for data transmissions whereas the transmission power can be significantly reduced.

In Fig. 7, we plot the values of m_0 that achieve optimal energy efficiency for different M values. Assuming $D = 100$, we find that the fast HARQ protocol can achieve optimal energy

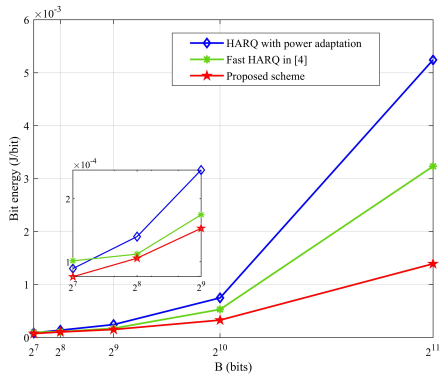

Fig. 5. Comparison of the bit energy.

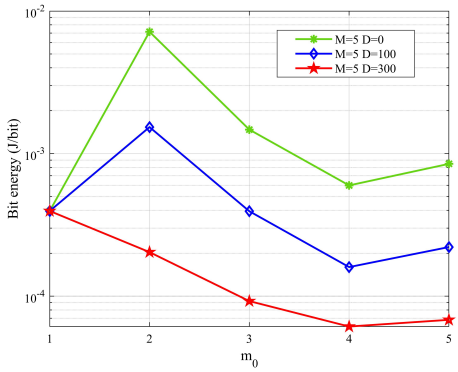

Fig. 6. Bit energy versus m_0.

efficiency when $m_0 = M - 1$ for both $B = 2^7$ and $B = 2^{11}$ bits. Meanwhile, We can find that without any feedback delay, the HARQ $(m_0 = 1)$ with power adaptation achieves best performance in energy efficiency when M is large.

V. Conclusion

In this paper, we have investigated a power adaptation policy under the constraints of feedback delay and reliability for a communication system in the FBL regime. Particularly, we have considered a fast HARQ scheme subject to the reliability constraints and transmission delay constraints. We have proposed to utilize the channel uses waiting for feedback delay in the conventional HARQ scheme for data transmission. Considering a long-term bit energy minimization problem, we have formulated a finite-episode MDP. Then, we trained a DRL agent to apply the A2C based algorithm to solve the problem considering the small decoding error probabilities. Numerical results have shown the effectiveness of the DRL algorithm in solving the non-convex problem. Additionally, compared with the HARQ and the existing fast HARQ protocol, the proposed protocol can achieve higher energy efficiency, especially when the packet size is large and the feedback delay is large.

Future efforts should aim to address the limitations identified in our study, including conducting further parameter simulations to examine the performance of this strategy. Additionally, considering a more realistic channel model may offer more robust or comprehensive results.

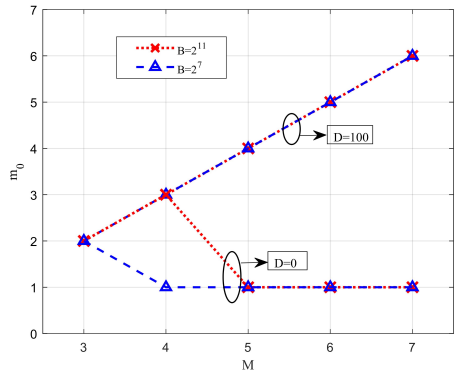

Fig. 7. Optimal m_0 versus M.

References

[1] "Study on scenarios and requirements for next generation access technologies, 3GPP, Sophia Antipolis, France, 3GPP Rep. *TR 38.913 V16.0.0*, Jul. 2020.

[2] C. She *et al.*, "A Tutorial on Ultrareliable and Low-Latency Communications in 6G: Integrating Domain Knowledge Into Deep Learning," *Proc. IEEE*, vol. 109, no. 3, pp. 204-246, Mar 2021.

[3] A. Ahmed, A. Al-Dweik, Y. Iraqi, H. Mukhtar, M. Naeem and E. Hossain, "Hybrid Automatic Repeat Request (HARQ) in Wireless Communications Systems and Standards: A Contemporary Survey," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 4, pp. 2711-2752, Fourthquarter 2021.

[4] B. Makki, T. Svensson, G. Caire and M. Zorzi, "Fast HARQ Over Finite Blocklength Codes: A Technique for Low-Latency Reliable Communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 194-209, Jan. 2019.

[5] H. Mukhtar, A. Al-Dweik, M. Al-Mualla, and A. Shami,"Low complexity power optimization algorithm for multimedia transmission over wireless networks, *IEEE Trans. Sig. Proc*, vol. 9, no. 1, pp. 113-124, Feb. 2015.

[6] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun*, vol. 58, no. 4, pp. 1129-1141, Apr 2010.

[7] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ, *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529C532, Oct. 2014.

[8] F. Nadeem, Y. Li, B. Vucetic and M. Shirvanimoghaddam, "Analysis and Optimization of HARQ for URLLC," 2021 *IEEE Globecom Workshops (GC Wkshps)*, Madrid, Spain, 2021, pp. 1-6.

[9] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions, *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475-2485, Nov. 2018.

[10] Y. Liu, H. Zhou, Y. Deng and A. N. Allanathan, "DRL-based Channel Access in NR Unlicensed Spectrum for Downlink URLLC," *IEEE Global Commun. Conf.(GLOBECOM)*, Rio de Janeiro, Brazil, pp. 591-596, 2022.

[11] M. Alsenwi, E. Lagunas and S. Chatzinotas, "Coexistence of eMBB and URLLC in Open Radio Access Networks: A Distributed Learning Framework," *IEEE Global Commun. Conf.(GLOBECOM)*, Rio de Janeiro, Brazil, pp. 4601-4606, 2022.

[12] X. Chen, L. Shan, X. Li, N. Deng and N. Zhao, "Proactive Dynamic Spectrum Sharing for URLLC Services Under Uncertain Environment via Deep Reinforcement Learning," *IEEE Wireless Commun. Netw. Conf.(WCNC)*, Austin, TX, USA, 2022.

[13] D. Qiao, M. C. Gursoy and S. Velipasalar, "Throughput-Delay Tradeoffs With Finite Blocklength Coding Over Multiple Coherence Blocks," *IEEE Trans. Commun*, vol. 67, no. 8, pp. 5892-5904, Aug. 2019.

[14] Y. Zhu, Y. Hu, X. Yuan, M. C. Gursoy, H. V. Poor and A. Schmeink, "Joint Convexity of Error Probability in Blocklength and Transmit Power in the Finite Blocklength Regime," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2409-2423, April 2023.