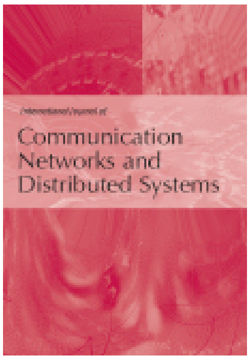


International Journal of Communication Networks and Distributed Systems

ISSN online: 1754-3924 - ISSN print: 1754-3916
https://www.inderscience.com/ijcnds

Subcarrier power control for URLLC communication system via multi-agent deep reinforcement learning in IoT network

Haiyan Wang, Xinmin Li, Feiying Luo, Jiahui Li, Xiaoqiang Zhang

DOI: 10.1504/IJCNDS.2024.10062215

Article History:

Received:	04 September 2023
Last revised:	17 November 2023
Accepted:	17 November 2023
Published online:	30 April 2024

Subcarrier power control for URLLC communication system via multi-agent deep reinforcement learning in IoT network

Haiyan Wang

School of Internet of Things and Intelligent Engineering,
Jiangsu Vocational Institute of Commerce,
Nanjing, China
Email: 050009@jvic.edu.cn

Xinmin Li*

Key Laboratory of Medicinal and Edible Plant Resources
Development of Sichuan Education Department,
Chengdu University,
Chengdu, China
and
Guangdong Provincial Key Laboratory of
Future Networks of Intelligence,
The Chinese University of Hong Kong,
Shenzhen, China
Email: lixm@swust.edu.cn
*Corresponding author

Feiying Luo

CEC Jinjiang Information Industry Co., Ltd.,
Chengdu, China
and
School of Information Engineering,
Southwest University of Science and Technology,
Mianyang, China
Email: 18708307075@163.com

Jiahui Li and Xiaoqiang Zhang

School of Information Engineering,
Southwest University of Science and Technology,
Mianyang, China
Email: leila_ljh@163.com
Email: xqzhang@swust.edu.cn

Abstract: Designing an intelligent resource allocation scheme to achieve the performance requirements of internet of things (IoT) devices for the future ultra-reliable low-latency communication (URLLC) network is a challenging task. In this paper, we formulate a joint blocklength allocation and power control optimisation problem to maximise the sum-rate performance with the short data packet in an uplink URLLC communication system. To alleviate this non-convex optimisation problem under the subcarrier power, blocklength and rate constraints, we firstly transfer it into a multi-agent reinforcement learning (RL) problem, in which each subcarrier works as the agent to decide its own power intelligently. Then a distributed blocklength allocation and power control scheme is proposed based on deep Q-network (DQN). To improve the rate performance in the dynamic communication environment, we design the segmented reward function depending on the communication rate and blocklength under different conditions, and adopt the experience replay strategy to avoid the dependency of training data. Finally, the simulation results show that the proposed scheme achieve the effectiveness and convergence under different settings compared to benchmark schemes.

Keywords: ultra-reliable low-latency communication; URLLC; blocklength allocation; power control; deep reinforcement learning.

Reference to this paper should be made as follows: Wang, H., Li, X., Luo, F., Li, J. and Zhang, X. (2024) ‘Subcarrier power control for URLLC communication system via multi-agent deep reinforcement learning in IoT network’, *Int. J. Communication Networks and Distributed Systems*, Vol. 30, No. 3, pp.374–392.

Biographical notes: Haiyan Wang received her BE and MS from the PLA University of Science and Technology, Nanjing, China, in 1997 and 2003, respectively. She is currently a Lecturer with the School of Internet of Things and Intelligent Engineering, Jiangsu Vocational Institute of Commerce. Her current research interests include network technique, computer software, communication engineering, and internet of things.

Xinmin Li received his PhD from the University of Science and Technology of China in 2017. He is with Chengdu University. He was with the School of Information Engineering, Southwest University of Science and Technology. His research interests include IoT network, learning algorithm, and data analytics.

Feiying Luo received her MS from the Southwest University of Science and Technology in 2023. She focuses on IoT network, learning algorithm, and FPGA platform.

Jiahui Li received her BS from the Southwest University of Science and Technology in 2021. She is pursuing her Master’s degree in Southwest University of Science and Technology. Her research interests include IoT network, and learning algorithm.

Xiaoqiang Zhang received his BE, MS and PhD from the Northwestern Polytechnical University, Xi’an, China, in 2010, 2013, and 2018, respectively. He is currently a Lecturer with the School of Information Engineering, Southwest University of Science and Technology. His current research interests include IoT system, synthetic aperture imaging, computational photography, and computer vision.

1 Introduction

With the rapid increasing of the global data traffic per year, the sixth generation (6G) mobile communication system need to support diverse applications and devices, e.g., internet of things (IoT), intelligent transportation, virtual reality (VR) and smart city (6G FLAGSHIP, 2019; Sulyman et al., 2017; Park and Bennis, 2018). It is forecasted that the number of the mobile devices reach more than ten billions in 2022 (Cisco, 2019), in which the number of mobile devices is five times of number of fixed devices. To meet the demands of massive connections and differentiated devices, the forthcoming fifth generation (5G) communication system takes the scenarios, key technologies into consideration to meet the extreme key performance indicators for each application scenario (3GPP TR38.913, 2017). There exist three major scenarios in the fifth generation system, i.e., enhanced mobile broadband (eMBB) service aims to guarantee the higher data rate, massive machine-type communication (mMTC) service aspires to increase the number of connections, and ultra-reliable low-latency communication (URLLC) service ensures the requirements of high reliability and low latency. Considering the time-sensitive applications in the 6G/5G network, e.g., vehicle-to-vehicle, industrial automation, augmented reality and VR, it is significant to design the solutions to guarantee the latency and reliability performance by applying various techniques (Yin et al., 2021). Recently, the third generation partnership project (3GPP) pointed out that URLLC system will provide 99.999% reliability and 1 ms latency for the future applications (Li et al., 2019). These metrics represent distinct key performance indicators compared to previous wireless communication systems. Table 1 presents the practical requirements of several typical URLLC applications in terms of latency and reliability.

Table 1 Practical requirements in URLLC applications

Application	*Latency (ms)*	*Reliability (%)*
Smart grid	3∼20	99.999
Augmented reality	0.4∼2	99.999
V2V	5	99.999
Professional audio	2	99.99999
Industrial automation	0.25∼10	99.9999999

Source: Sutton et al. (2019)

3GPP developed a new radio (NR) air interface to achieve the requirements of future 5G/6G networks (Anutusha et al., 2020) and designed the minislot to reduce the latency (3GPP TS38.211, 2018), in which the new slot structures are suitable for the short-packet communications. Unlike that, the current wireless communication systems adopt the long-packet transmissions, and the achievable rate is normally characterised by Shannon's capacity (Eggers et al., 2019). However, due to the demanding latency requirement in critical applications, the size of transmitted URLLC packets is small. Thus, the communication is no longer reliable and the decoding error probability is no longer negligible. As a result, Shannon's capacity is not applicable to characterise the maximum achievable rate of short URLLC packets. Otherwise, the performance of the latency and reliability will be underestimated (Giampaolo et al., 2023; Pase et al., 2022). This necessitates the achievable rate characterisation and relaying protocol design under the finite blocklength (FBL) regime (Polyanskiy et al., 2010; Liu et al., 2021). Thus, the finite blocklength theorem has been proposed in Polyanskiy et al. (2010) to obtain tight bounds on coding rate under short-packet transmission. Based on this theorem, the existing literature has carried out massive advanced works in URLLC systems (Hu et al., 2020; He et al., 2021). In Ramin et al. (2021), the authors explored the average achievable rate and error probability of systems assisted by reconfigurable intelligent surfaces in finite blocklength regime. For short packet communication between sources and robots used in automated factories, Jiang et al. (2021) considered different interference mitigation methods for full-duplex systems to optimise the latency and ehance the coverage and reliability. Ren et al. (2019) optimised the blocklength and power allocation jointly to minimise the decoding error probability. Therefore, to meet the performance requirements of URLLC systems with the finite blocklength, it is a challenging problem to allocate the system resources efficiently due to the limited blocklength resource.

To solve the above problems, the existing literature has exploited some promising solutions for URLLC systems to achieve different targets, e.g., spectrum efficiency (Chang et al., 2019; Mohamed et al., 2021), energy efficiency (Ayidh et al., 2020; Haque et al., 2023), power control (Wang and Zhang, 2019; Yang et al., 2021). Considering the ultra-reliable uplink transmission design between multiple robots and a central controller with stringent delay requirements, some scholars meet the requirements of system throughput and reliability by jointly optimising error probability, block allocation and transmit power allocation (Celebi et al., 2022; Cheng and Shen, 2022). In Chang et al. (2019), the authors studied the resource allocation scheme in real-time uplink URLLC control system, which aims at maximising the spectrum efficiency by adjusting the optimal spectrum. In Ayidh et al. (2020), the authors analysed the uplink energy efficiency of uplink URLLC communication system. To guarantee the ultra-reliable low-latency services for the resources-limited devices, Sui et al. (2023) proposed an energy-efficient frame allocation method to determine the size of the resource block. By analysing the characteristics of interference matrices, Wang and Zhang (2019) constructed a feasible power control scheme by scheduling the close-by links silent. In the multiple subcarrier systems, each subcarrier can be shared to multiple users, and each user is allowed to reuse the independent subcarriers, which will inevitably cause user interference. A method based on channel selection and power control scheme is used to solve the weighted rate maximisation problem in device-to-device (D2D) network (Tan et al., 2019). In the existing literature, methods such as stochastic geometry (Dai et al, 2017), and game theory (Zhou et al., 2020), have been applied to optimise the transmission power. However, due to the large-scale connections of massive users in future applications, various quality of service requirements need to be considered. In particular, each user making decision individually can cause a direct influence on interference level of other users, while the problem of power control and blocklength are coupled. In conclusion, it is challenging to obtain the optimal resource allocation scheme in the dynamic wireless communication system.

Different from the conventional optimisation methods, deep reinforcement learning (DRL) can store the history experience and select a beneficial action from the predefined action space based on history experience to obtain the maximal reward. Thus, it is more efficient to solve the complex decision-making tasks that optimise the transmit power and blocklength. Kasgari et al. (2020) proposed a resource allocation scheme based on the experienced DRL to guarantee the high reliability and low latency performance for each wireless user under data rate constraints. Gu et al. (2020) applied DRL method to solve the subcarrier power allocation problem for the dynamic link environment in D2D communication system. To ensure low access latency and high connectivity density, Tran et al. (2023) proposed a DRL method to suppose energy-efficient resource allocation strategies. In Wang et al. (2023), a multi-agent soft-actor-critic-discrete based URLLC-constrained scheme was proposed to maximise the throughput. To reduce the energy consumption and delay of mobile devices, Naveen et al. (2023) proposed the energy-saving resource allocation scheme based on DRL method. A method based on deep learning was also proposed for joint optimisation of reconfigurable intelligent surface and power allocation of access point over each subcarrier (Zhong et al., 2022).

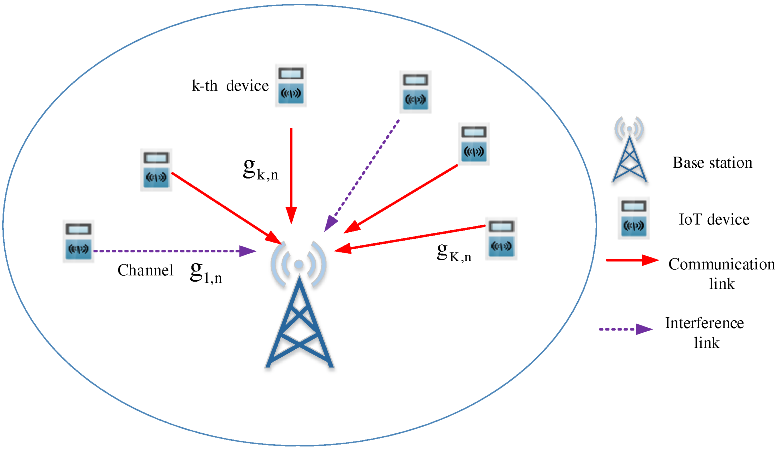
In this work, we consider an uplink URLLC communication network, where each IoT device uses to the shared subcarriers to send their information to the base station (BS) under blocklength and power constraints. In this URLLC communication system, we aim to optimise the channel blocklength and the subcarrier transmit power jointly to maximise the sum rate, subject to the maximum blocklength and transmit power. The main contributions of the work are summarised as follows:

- We adopt the finite blocklength theory to model the rate metric in uplink URLLC communication system and formulate the rate maximisation optimisation problem under the blocklength and transmit power of IoT devices constraints. The non-convex rate problem is transformed into the multi-agent decision process. Thus, each subcarrier has a capable of making intelligent decision in its own communication environment.
- A distributed deep Q-network (DQN)-based algorithm is proposed to optimise the transmit power and blocklength, in which each subcarrier works as the independent agent to select the best action. The independent state space, action space, and the piecewised reward function are constructed to satisfy the rate performance and the blocklength constraints. Moreover, the experience replay and random sampling strategies are adopted to decrease the dependency of agents' training data during the learning process.
- Extensive simulation results show that the proposed scheme achieves better rate performance compared with Q-learning scheme and the greedy scheme. The convergence of the proposed scheme is validated under different learning rates and training episodes.

2 System model

We consider an uplink URLLC IoT network consisting of a single-antenna BS and K single-antenna IoT devices, in which all devices share the subcarrier set $\mathcal{N} = \{1, 2, ..., N\}$ and transmit the information to the BS concurrently. All the IoT devices are serving by the BS depicted in Figure 1. For simplicity of exposition, let $\mathcal{N}_k$ and $\mathcal{K}_n$ denote the occupied subcarrier index of the k^{th} device and device indication using the n^{th} subcarrier, respectively. W is the total transmission bandwidth and each subcarrier spacing is $W_{\text{sc}} = \frac{W}{N}$. For simplicity, we assume that each device has N_{a} subcarriers, and the channel information between the BS and the IoT devices keep constant due to low mobility and short-range coverage.

Figure 1 System model of uplink URLLC IoT communication network (see online version for colours)



The channel gain between the BS and the k^{th} IoT device over the n^{th} subcarrier is expressed as follows:

$$g_{k,n} = \sqrt{\beta_k} h_{k,n} \tag{1}$$

where β_k represents the large-scale fading and $h_{k,n}$ denotes the small-scale fading between the k^{th} IoT device and the BS over the n^{th} subcarrier. The transmit symbol of the k^{th} IoT device on the n^{th} subcarrier, denoted as $s_{k,n}$, is assumed to be an independent and identical complex Gaussian variable, i.e., $s_{k,n} \sim \mathbb{CN}(0,1)$. Thus, the received signal at the BS over the n^{th} subcarrier is given by

$$y_n = \sum_{k \in \mathcal{K}_n} \sqrt{p_{k,n}} g_{k,n} s_{k,n} + z_n \tag{2}$$

where $p_{k,n}$ denotes the transmit power of k^{th} IoT device over the n^{th} subcarrier, and z_n is the received noise with zero-mean and variance σ^2 at the BS. The signal-to-interference-plus-noise ratio (SINR) $\gamma_{k,n}$ between the k^{th} device and the BS over n^{th} subcarrier can be given by

$$\gamma_{k,n} = \frac{p_{k,n} |g_{k,n}|^2}{\sum\limits_{i \in \mathcal{K}_n \setminus k} p_{i,n} |g_{i,n}|^2 + \sigma^2} \tag{3}$$

In the infinite blocklength communication, the reliable transmission can be achieved without decoding errors. However, packet size of IoT device is always small in URLLC system due to latency requirements or application characteristics, which means that the accurate encoding rate cannot be obtained from the perspective of Shannon's channel capacity. Thus, for given error probability and finite blocklength, the achievable rate $R_{k,n}$ between the k^{th} IoT device and the BS over n^{th} subcarrier can be approximated (Polyanskiy et al., 2010).

$$R_{k,n} = \log_2(1 + \gamma_{k,n}) - \sqrt{\frac{V(\gamma_{k,n})}{l_{k,n}}} \frac{\tilde{Q}^{-1}(\eta_k)}{\ln 2} \tag{4}$$

where η_k and $l_{k,n}$ denote the required error probability of the k^{th} device and the blocklength of the k^{th} device over the n^{th} subcarrier, respectively. According to Polyanskiy et al. (2010) and Fang et al. (2021), the approximation is very accurate when the blocklength is greater than 100. $V(\gamma_{k,n}) = 1 - 1/(1+\gamma_{k,n})^2$ is the channel dispersion and $\tilde{Q}^{-1}(x)$ is the inverse Gaussian Q-function with $\tilde{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp^{-t^2/2} dt$. Therefore, the communication rate of the k-device in URLLC communication system is written as

$$R_k = \sum_{n \in \mathcal{N}_k} R_{k,n} \tag{5}$$

Our target is to maximise the sum rate of URLLC communication system by optimising subcarrier's power and blocklength jointly, thus the optimisation problem is formulated as follows:

$$\text{(P1):} \max_{\{p_{k,n}, l_{k,n}\}} \sum_{k=1}^{K} R_k \tag{6}$$

s.t.

$$0 \le p_{k,n} \le P_{\max}, n \in \mathcal{N}, k = 1, ..., K \tag{7}$$

$$R_k \ge R_{\min}, \tag{8}$$

$$l_{k,n} \ge L_{\min}, \tag{9}$$

$$\sum_{k=1}^{K} \sum_{n=1}^{N} l_{k,n} \le L_{\max}, \tag{10}$$

where $P_{\max}$, $R_{\min}$, $L_{\min}$ and $L_{\max}$ denote the maximum transmit power, the minimum rate, the minimum blocklength and maximum blocklength, respectively. $L_{\max}$ is always related to the transmission duration T and system bandwidth W, i.e., $L_{\max} = WT$ (Li et al., 2019; Giuseppe et al., 2016; Feng et al, 2022), which means that the data transmission under the latency constraint, and the transmission has to be complete within $L_{\max}$ blocklength. On the one hand, the sum-rate depends on its dynamic environment, thus the objective function of optimisation problem is non-convex. On the other hand, the transmit power, the blocklength are coupling to have a direct effect on SINR and rate respectively, and there exists huge computation complexity to search multi-device's power and blocklength in unknown system. Thus, it is difficult to obtain the optimal solution via the standard convex optimisation method. To solve this non-convex problem, we transform problem (P1) into a decision process and propose an intelligent DRL-based scheme to select the power and allocate the blocklength.

3 Proposed power control scheme based on DRL

Recently, DRL is one of promising machine learning methods to solve the resource allocation problem to enable the intelligence of wireless communication systems since it has a capable of making a decision by selecting the potential action based on the stored experiences and using the deep neural network to learn instead of the massive number of values. Markov decision process (MDP) is applied to model the RL process. MDP can be modelled by a tuple $\langle \mathcal{S}, \mathcal{A}, R, \gamma \rangle$ with the state space $\mathcal{S}$, action $\mathcal{A}$, reward R and discount factor $\gamma \in [0,1]$ (Zhao et al., 2023). At the step t, the agent selects the action a_t by interacting with the system environment to maximise the long-term cumulative reward $R_t = r_t + \sum_{t'=1}^{t-1} \gamma^{(t-t')} r_{t'}$.

Different the conventional optimisation methods, we propose a multi-agent RL scheme for blocklength allocation and power control, which is suitable for the high-dimensional action space. To construct an efficient RL algorithm, we need to form the sate space of the environment and action space of agent, and model the specific reward function of the environment to satisfy the constraints and maximise the objective function of problem (P1).

3.1 State, action and reward function

- Agent: in our work, each subcarrier works as the agent in the RL process. Specifically speaking, when the n^{th} subcarrier of the k^{th} device is the current agent at the step t, the agent can independently decide the transmit power value $a_{k,n}^{p,t}$ and the blocklength $a_{k,n}^{l,t}$ derived from the current state $\mathbf{s}_{k,n}^{t}$ and reward r^t to satisfy the power and blocklength constraints, and maximise the sum rate performance.
- Action space: the action space includes the discrete transmit power values and the blocklength values, therefore each agent can select action $\mathbf{a}_{k,n}^{t} = \{a_{k,n}^{p,t}, a_{k,n}^{l,t}\} \in \mathcal{A} = \{\mathcal{A}^p, \mathcal{A}^l\}$ in any state to transition the next state during time slot t. For simplicity, we assume that each agent has the same the action space $\mathcal{A}^p = \{0, \frac{P_{\max}}{L_p-1}, \frac{2P_{\max}}{L_p-1}, ..., P_{\max}\}, \mathcal{A}^l = \{0, \frac{L_{\max}}{L_l-1}, \frac{2L_{\max}}{L_l-1}, ..., L_{\max}\}$, where $P_{\max}$ and $L_{\max}$ denote the maximum of transmit power and the system total blocklength, and L_p and L_l denote the length of the action space $\mathcal{A}^p, \mathcal{A}^l$, respectively. The agent can independently select the action $a_{k,n}^{p,t} \in \mathcal{A}^p, a_{k,n}^{l,t} \in \mathcal{A}^l$ to maximise the reward value.
- State space: the state space consists of desired power and interference power, respectively, i.e., the state of the k^{th} device over the n^{th} subcarrier at the step t is expressed as

  $$\mathbf{s}_{k,n}^{t} = [p_{k,n}^{t}|g_{k,n}|^2, ..., p_{i,n}^{t}|g_{i,n}|^2, ...], i \in \mathcal{K}_n \tag{11}$$

  In the initial state, i.e., $t = 0$, each agent can randomly select the subcarrier power and blocklength according to the constraints (7), (9) and (10). Because of the current state $\mathbf{s}_{k,n}^{t}$ and the action $\mathbf{a}_{k,n}^{t}$, the agent can obtain the next state $\mathbf{s}_{k,n}^{t+1}$.
- Reward function: to maximise the communication rate of the k-device under the blocklength constraint, we use the difference between the system blocklength and the used blocklength as

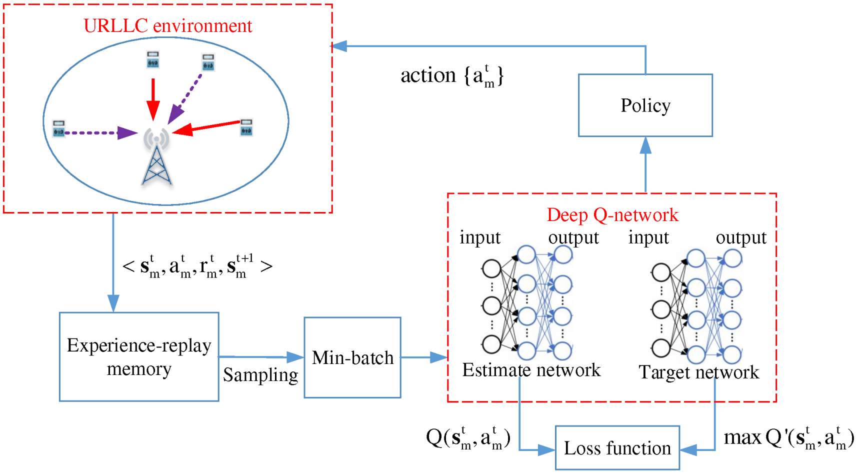
  $$\phi = \sum_{k=1}^{K} \sum_{n=1}^{N} l_{k,n} - L_{\max}, n \in \mathcal{N}, k = 1, ..., K \tag{12}$$

and design the segmented reward function of each agent as

$$r_{k,n}^{t}(\mathbf{s}_{k,n}^{t}, \mathbf{a}_{k,n}^{t}) = \begin{cases} -\lambda\phi - R_{k,n}^{t}, & \text{if } R_{k,n}^{t} < R_{\text{min}} \text{ and } \phi > L_{\text{max}}, \\ -\lambda\phi + R_{k,n}^{t}, & \text{if } R_{k,n}^{t} \geq R_{\text{min}} \text{ and } \phi > L_{\text{max}}, \\ -R_{k,n}^{t}, & \text{if } R_{k,n}^{t} < R_{\text{min}} \text{ and } \phi \leq L_{\text{max}}, \\ R_{k,n}^{t}, & \text{otherwise.} \end{cases} \tag{13}$$

where $\lambda \in (0, 1)$ is the weighted parameter. It is observed that the agent has the penalty value when the blocklength constraint (9) and (10) is not satisfied, and can obtain more reward value when blocklength becomes smaller. Thus, the agent can select the potential action to maximise the rate performance.

Figure 2 The framework of proposed DQN-based scheme (see online version for colours)



3.2 *Proposed deep Q-network algorithm*

Considering the continuous power and blocklength variables in the URLLC communication system are quantised into discrete values, DRL creates the state-action function that characterises the impact of chosen actions on performance in specific states. On the other hand, the complexity of the selected algorithm becomes an important index in this paper. The whole procedure of DRL algorithm mainly contains two parts: reward calculation and action selection.

The computational complexity of all agents to calculate the reward is $O(N.|s_{k,n}|)$ and it is determined by the number of agents. The complexity of action selection is usually determined by the network structure, and the appropriate algorithm is particularly important such as DQN and DDPG. The neural network structure of DQN agents includes a single neural network with three hidden layers and $3K$ hidden nodes in each layer. For the DQN network, the number of neurons in the m^{th} layer is U_m, and the number of layers in the DQN network is M. Thus, the computational complexity of the DQN networks for all agents is $O(K(|s_{k,n}|.U_2 + \sum_{m=3}^{M}(U_{m-1}U_m + U_mU_{m+1} + U_{M-1}\cdot|a_{k,n}|)))$ (Xi et al., 2021). However, there are two complex neural networks at each DDPG agent. Each actor and critic network consists of three hidden layers, with $3K$ hidden nodes in each layer. Therefore, the complexity of the actor network is $O_A(K(|s_{k,n}|.U_2 + \sum_{m=3}^{M_A}(U_{m-1}U_m + U_mU_{m+1} + U_{M-1}.|a_{k,n}|)))$, and the critic network can be expressed as $O_C(K(|s_{k,n}|.U_2 + \sum_{m=3}^{M_C}(U_{m-1}U_m + U_mU_{m+1} + U_{M-1}.|a_{k,n}|)))$. The computational complexity of the DDPG network is $O_A + O_C$ (Ciftler et al., 2022). The implementation of DDPG is more complex compared to DQN method. Thus, DQN method is more suitable for solving the optimisation problem (P1) in this work.

DQN method employs the neural networks to learn the policy instead of storing state-action values, effectively reducing the dimensionality of the action space (Wu et al., 2021). The proposed DQN-based design framework for the URLLC communication system is illustrated in Figure 2. With the control policy ξ for the n^{th} subcarrier, the Q-function is written as

$$Q^{\xi}(\mathbf{s}_n^t, \mathbf{a}_n^t) = E\left[r_n(\mathbf{s}_n^t, \mathbf{a}_n^t) + \sum_{j=0}^{t-1} \gamma^j r_n(\mathbf{s}_n^j, \mathbf{a}_n^j)\right] \tag{14}$$

where $\gamma \in [0, 1]$ is the discount factor. The Q-function is connected to the current reward when the discount factor $\gamma = 0$, implying that the agent's action select depends on the current reward $r_n(\mathbf{s}_n^t, \mathbf{a}_n^t)$. The optimal action to maximise the rate performance in (P1) is $\mathbf{a}_n^{t,*} = \arg\max_{\mathbf{a}_n^j \in \mathcal{A}} Q^{\xi}(\mathbf{s}_n^t, \mathbf{a}_n^j)$ by searching Q-value under different potential actions (Wu et al., 2021).

We can obtain the optimal control policy ξ^* by updating the Q-function as follows:

$$Q^{t+1}(\mathbf{s}_n^t, \mathbf{a}_n^t) = Q(\mathbf{s}_n^t, \mathbf{a}_n^t) + \nu\left(r(\mathbf{s}_n^t, \mathbf{a}_n^t) + \gamma \max_{\mathbf{a}_n^j \in \mathcal{A}} Q(\mathbf{s}_n^{t+1}, \mathbf{a}_n^j) - Q(\mathbf{s}_n^t, \mathbf{a}_n^t)\right) \tag{15}$$

where ν represents the learning rate. According to (15), each subcarrier has the capacity to update the Q-function and learn the control policy by selecting actions that maximise the stored Q-values, subsequently maximising rewards. To address action selection within the constraints of limited state-action information, an ϵ-greedy strategy is used for environment exploration. The exploration probability ϵ is defined as

$$\mathbf{a}_n^t = \begin{cases} \text{random}(\mathcal{A}), & \text{with} \quad \text{probalility } \epsilon, \\ \arg\max\limits_{\mathbf{a}_n^j \in \mathcal{A}} Q^{\xi}(\mathbf{s}_n^t, \mathbf{a}_n^j), & \text{with} \quad \text{probability } 1-\epsilon. \end{cases} \tag{16}$$

According to this strategy, the subcarrier exhibits a stochastic behaviour by taking random actions with a probability denoted as ϵ, thereby facilitating exploration within the URLLC communication environment. Considering the subcarrier's unknown state space demands extensive memory and leads to slow convergence, utilising a deep neural network to intelligently extract features from available datasets can alleviate computational complexity by predicting outputs. As shown in Figure 2, the tuple comprising state, action, reward, and next state serves as input for a deep neural network. This network produces Q-values, denoted as $Q(\mathbf{s}_n^{t+1}, \mathbf{a}_n^t|\theta_t)$ and $Q(\mathbf{s}_n^{t+1}, \mathbf{a}_n^t|\theta_t^-)$, within estimate and target neural networks. Here, θ_t and θ_t^- represent parameters of the estimate and target neural networks during the i^{th} training iteration, respectively (Li et al., 2021). In the deep neural network, the target neural network replicates the estimated neural network every N_{rep} steps. This strategy ensures their proximity for stability purposes. Thus, optimising neural network parameters θ_i through an appropriate loss function becomes crucial for obtaining the optimal Q-function. The defined loss function is as follows:

$$\mathcal{L}(\theta_t) = \left| r(\mathbf{s}_n^t, \mathbf{a}_n^t) + \gamma \max_{\mathbf{a}_n^j \in \mathcal{A}} Q'(\mathbf{s}_n^{t+1}, \mathbf{a}_n^j | \theta_t^-) - Q(\mathbf{s}_n^t, \mathbf{a}_n^t | \theta_t) \right|^2 \tag{17}$$

On the basis of the loss function and the training dataset, various optimiser, including the gradient descent algorithm, can be employed to acquire the optimal parameters for the neural network.

Algorithm 1 Multi-agent power control and blocklength allocation scheme based on deep Q-network for URLLC communication system

```
Initialise URLLC communication system: number of IoT devices $K$, subcarrier set $\mathcal{N}$,
subcarrier spacing $W_{\text{sc}}$ and error probability $\eta$.
Initialise DQN parameters: learning parameters $\{\nu, \gamma, \epsilon\}$, number of agents $M$, the
maximal episode $N_{\text{ep}}$, batch size $N_{\text{bat}}$ and neural network.
for $i = 1 : N_{\max}$ do
    Initialise the environment and state $\mathbf{s}_0$.
    for $j = 1 : M$ do
        Initialise the state $\mathbf{s}_n^0$, and obtain the corresponding action $\mathbf{a}_n^t$ from the action set
        $\mathcal{A}$ by $\epsilon$-greedy method.
        Execute the action $\mathbf{a}_n^t$, compute the reward $r_n^t$ according to (13) and obtain the
        next state $\mathbf{s}_n^{t+1}$. Store $< \mathbf{s}_n^t, \mathbf{a}_n^t, r_n^t, \mathbf{s}_n^{t+1} >$ into the experience-reply memory and select
        the batch samples from the memory randomly.
        If $t\%N_{\text{rep}}$ ==0, duplicate the estimate neural network to target neural network.
        Train the neural network by the loss function in (17) to optimise the parameter $\theta$,
        $t \leftarrow t + 1$.
    end for
end for
Generate the power value $p_n$ and blocklength value $l_n$ over each agent and compute the
rate performance.
```

The training data is crucial for effectively training the deep neural network. To address the dependency on training data, the experience replay and random sampling methods are employed. We use an experience replay memory, denoted as N_{mem}, to store tuples from the RL learning process. This data is updated every N_{tr} steps, ensuring the training data remains fresh. The random sampling scheme is employed to complete batches by randomly selecting experience data from the replay memory. This approach helps smooth the transition between historical data and current observations (Li et al., 2021). The proposed DQN-based scheme for URLLC communication systems is shown in Algorithm 1.

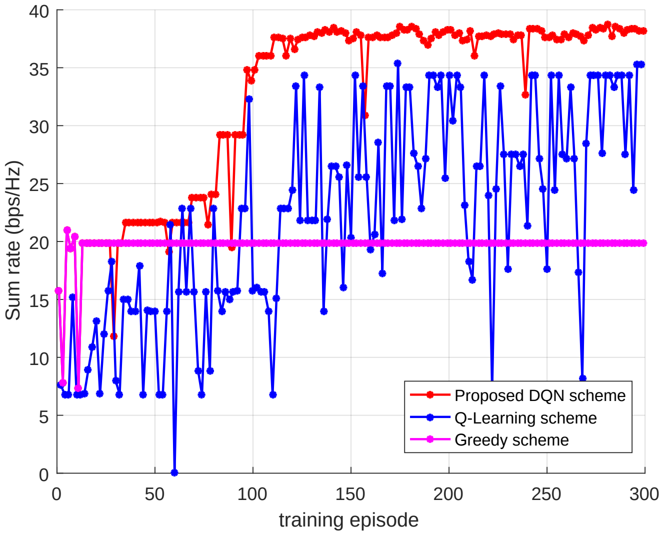
4 Simulation results

In this section, we demonstrate the simulation results that highlight the effectiveness and convergence of the proposed DQN-based algorithm for the power control and blocklength allocation scheme. Additionally, we analyse the impact of various parameters on the sum-rate performance. We consider that all IoT devices are randomly distributed in a circular area with a radius of 800 m centred on the base station. The large-scaling fading between the k^{th} device and the BS is modelled as $\beta_k = 128.1 + 37.6 \log 10(d_k)$ (Chang et al., 2019) with d_k denoting the distance and the small-scaling between the k^{th} device and the BS over n^{th} subcarrier is written as $h_{k,n} \sim \mathbb{CN}(0, 1)$. The neural network consists of one input layer, three hidden layers, and one output layer. To optimise the parameters of the neural network, this work employed the gradient descent method. And fresh training data consistently updates the oldest historical data in the experience memory. To ensure an efficient batch data size, training begins after N_{bat} steps.

Table 2 Simulation parameters for URLLC communication system

Symbol	*Description*	*Value*
W_{sc}	Bandwidth	30 KHz, 60 KHz
$P_{\max}$	Maximum transmit power of each subcarrier	20 dBm
σ^2	Noise power	–90 dBm
N_{a}	Subcarrier number of each IoT device	2
R_{s}	Radius of BS	800 m
L_p	Length of power control action space	8
L_l	Length of blocklength allocation action space	10
η	Error probability	$\{10^{-5}, 10^{-7}\}$
γ	Discount factor	0.9
N_{bat}	Batch size	64

Figure 3 The sum rate versus under different training episodes (see online version for colours)



The simulation environment is Intel i7 CPU, Python 3.5 and Tensorflow 1.10 to train the deep neural network. All results are averaged over 200 episodes. The simulation parameters are shown in Table 2. We adopt Q-learning scheme and greedy scheme as the benchmark schemes for comparison. The Q-learning algorithm is a class of value-based reinforcement learning methods. It involves organising states and actions within a Q-table to store corresponding Q values. This table is then utilised to select actions that obtain the maximum benefits. Similarly, the greedy algorithm, a commonly employed approach, focuses on choosing the optimal option within the current state to get the best solution.

Figure 4 The sum rate versus the learning rate ν = [0.1, 0.01, 0.001, 0001] (see online version for colours)

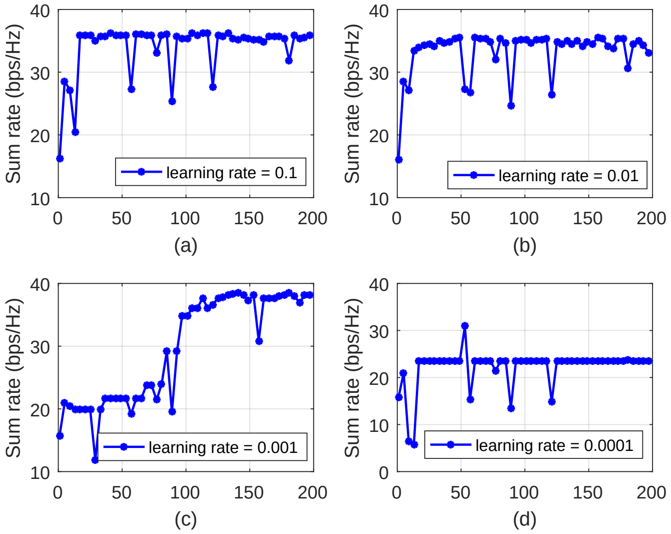


Figure 3 depicts the rate performance comparison between the proposed DQN-based scheme and benchmark schemes under distinct learning episodes. The performance of the proposed scheme based on DQN consistently outperforms that of the benchmarks. It is also seen that the rate performance of the proposed schemes and the Q-learning-based scheme converge to the optimal value as the training episode increases and the convergence rate of the proposed scheme is significantly faster and more stable. It is shown that the proposed scheme outperforms the Q-learning-based scheme by 16.08%. This is attributed to the utilisation of experience replay and random sampling from the batch during the learning process in the proposed scheme, enabling the agent to adeptly and efficiently adapt to dynamic environmental changes.

Figure 4 shows the influence of learning rate on rate of the proposed DQN-based scheme. It can be seen that when the learning rate is 0.0001, the proposed algorithm has poor convergence and can not obtain a good power allocation policy. We also can find that when the learning rate is 0.1 or 0.01, the rate converges quickly and it has a comparatively higher value. Thus, the proposed scheme can converge to a stable performance by selecting the optimal learning rate depending on the different communication environments

Figure 5 The effect of number of IoT devices on the average rate of the proposed scheme with fixed N (see online version for colours)

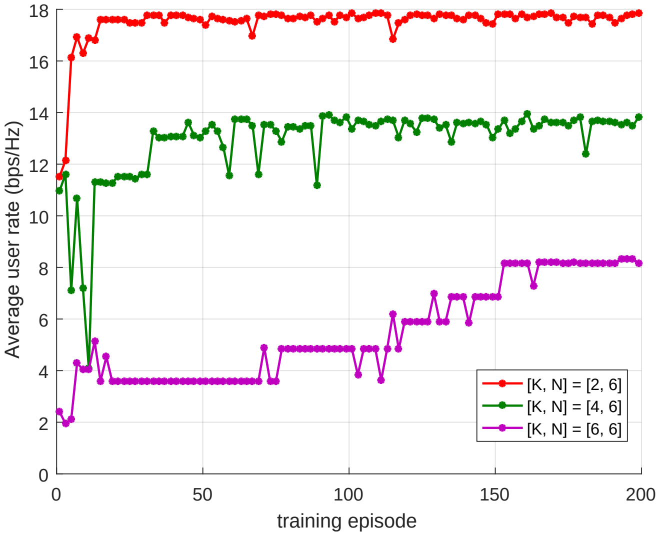


Figure 5 the effect of number of IoT devices on rate performance of the proposed scheme. It is observed that as the IoT device number and the increase of the ratio of the subcarrier number, the rate of each device decreases, e.g., the rate of IoT device with the ratio $\frac{K}{N} = 1/3$ achieves the gain of the rate performance up to 114.68% compared to the ratio $\frac{K}{N} = 1$. It is originated from the increasing of interference power as the ratio increases.

Figure 6 depicts the effect of the subcarrier spacing and the error probability on the rate performance. The 3GPP standard, release-15, introduces the concept of min-slot to support the URLLC applications by reducing the transmission time interval (3GPP TR38.912, 2016), and NR release-15 has scalable numerology with subcarrier spacing of 15 KHz, 30 KHz, and 60 KHz below 6 GHz, and 120 KHz or 240 KHz above 6 GHz (Joachim et al., 2018). It is observed that the rate improves as the subcarrier spacing increases. On the other hand, URLLC applications, such as industrial automation and autonomous vehicles, require extremely low error rates to ensure reliability and safety (Haque et al., 2023). To ensure the safety of such applications, the communication system must operate with extremely low error rates (Khan et al., 2022), e.g., the error probabilities using 10^{-5} and 10^{-7} reflect the safety-critical requirements of these applications. It is observed that as the required error probability η decreases, the rate performance decreased as well, which is consistent with the rate expression (4).

Different parameter weights affect the algorithm's convergence and stability. The effect of the weight parameter of the reward function on the sum rate performance is shown in Figure 7. It is observed that the convergence value of the proposed scheme decreases with the increase of the weight parameter, which originates from the designed reward function is related to the blocklength and sum rate. According to (13), the difference among the obtained reward results when the weight parameter $\lambda = 0.1$ is greater than that of reward results when the weight parameter $\lambda = 0.5$. Thus, the proposed scheme can choose the optimal actions for small weight parameters. However, small weight parameter may cause unstable sum rate.

Figure 6 The sum rate versus the subcarrier spacing with $W_{\mathrm{sc}} = [30, 60]$ KHz and the error probability with $\eta = [10^{-5}, 10^{-7}]$ (see online version for colours)

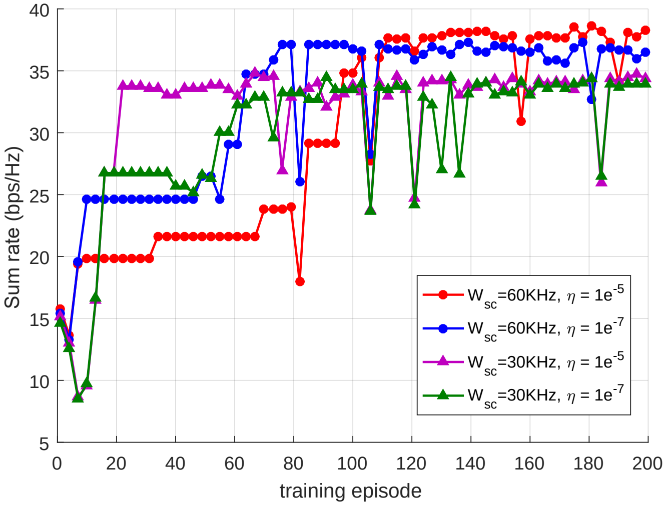
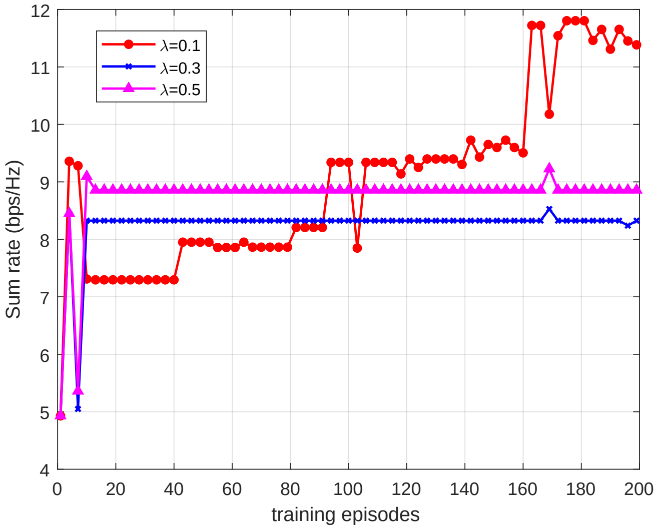


Figure 7 The sum rate versus the weight parameter of the reward function with $\lambda = [0.1, 0.3, 0.5]$ (see online version for colours)



5 Conclusions

In this paper, we studied the joint blocklength allocation and power control to maximise the rate performance in the uplink URLLC IoT network under the subcarrier power, blocklength and rate constraints. We decompose the non-convex optimisation problem into a multi-agent RL process, in which each subcarrier acts as an agent to intelligently determine its power. Depending on DQN, we introduced the joint scheme for blocklength allocation and power control, incorporating experience replay to circumvent training data dependencies. The simulation results demonstrate that the proposed scheme outperforms the benchmark scheme in terms of sum-rate performance and convergence. In the future, investigating the intelligence schemes of URLLC systems with multiple antennas are interesting.

Acknowledgements

The work was supported in part by Key Program of Jiangsu Vocational Institute of Commerce under Grant JSJM20008, in part by Key Laboratory of Medicinal and Edible Plant Resources Development of Sichuan Education Department, Chengdu University under Grant 10Y202201, in part by Guangdong Provincial Key Laboratory of Future Networks of Intelligence, The Chinese University of Hong Kong, Shenzhen under Grant 2022B1212010001-OF04, in part by National Natural Science Foundation of China under Grant 62201479, in part by the Natural Science Foundation of Sichuan Province under Grant 2023NSFSC1388 and the Sichuan Science and Technology Program under Grant 2022YFG0148.

References

3GPP TR38.912 (2016) *Study on New Radio Access Technology: Radio Access Architecture and Interfaces* [online] https://www.etsi.org/deliver/etsi_tr/138900_138999/138912/14.01.00_60/tr_138912v140100p.pdf.

3GPP TR38.913 (2017) *Study on Scenarios and Requirements for Next Generation Access Technologies* [online] https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.02.00_60/tr_138913v140200p.pdf.

3GPP TS38.211 (2018) *NR: Physical Channels and Modulation* [online] https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/15.02.00_60/ts_138211v150200p.pdf.

6G FLAGSHIP (2019) *6G White Paper: Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence* [online] http://jultika.oulu.fi/files/isbn9789526223544.pdf.

Anutusha, D., Rakesh, K.J. and Shubha, J. (2020) 'A survey on beyond 5G network with the advent of 6G: architecture and emerging technologies', *IEEE Access*, Vol. 9, No. 67, pp.512–547.

Ayidh, A. A, Chang, B., Zhao, G., Ghannam, R. and Imran, M. (2020) 'Energy-efficient power allocation in URLLC enabled wireless control for factory automation applications', *2020 IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp.1–6.

Celebi, H., Pitarokoilis, A. and Skoglund, M. (2022) 'A multi-objective optimization framework for URLLC with decoding complexity constraints', *IEEE Transactions on Wireless Communications*, Vol. 21, No. 4, pp.2786–2798.

Chang, B., Zhang, L., Li, L., Zhao, G. and Chen, Z. (2019) 'Optimizing resource allocation in URLLC for real-time wireless control systems', *IEEE Transactions on Vehicular Technology*, Vol. 68. No. 9, pp.8916–8927.

Cheng, J. and Shen, C. (2022) 'Relay-assisted uplink transmission design of URLLC packets', *IEEE Internet of Things Journal*, Vol. 9, No. 19, pp.18839–18853.

Ciftler, B., Alwarafy, A. and Abdallah, M. (2022) 'Distributed DRL-based downlink power allocation for hybrid RF/VLC networks', *IEEE Photonics Journal*, Vol. 14, No. 3, pp.1–10.

Cisco (2019) *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update*, pp.2017–2022 [online] https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf.

Dai, J., Liu, J., Shi, Y., Zhang, S. and Ma, J. (2017) 'Analytical modeling of resource allocation in D2D overlaying multihop multichannel uplink cellular networks', *IEEE Transactions on Vehicular Technology*, Vol. 66, No. 8, pp.6633–6644.

Eggers, P., Angjelichinoski, M. and Popovski, P. (2019) 'Wireless channel modeling perspectives for ultra-reliable communications', *IEEE Transactions on Wireless Communications*, Vol. 18, No. 4, pp.2229–2243.

Fang, M., Li, D., Zhang, H., Fan, L. and Trigui, I. (2021) 'Performance analysis of short-packet communications with incremental relaying', *Computer Communications*, Vol. 177, No. 1, pp.51–56.

Feng, R., Li, Z., Wang, Q. and Huang, J. (2022) 'An ADMM-based optimization method for URLLC-enabled UAV relay system', *IEEE Wireless Communications Letters*, Vol. 14, No. 8, pp.1–5.

Giampaolo, C., Pettersson J. and Condoluci, M. (2023) 'Analysis of a contention-based approach over 5G NR for federated learning in an industrial Internet of Things scenario', *IEEE Access*, Vol. 11, No. 5, pp.74473–74485.

Giuseppe, D., Tobias, K. and Petar, P. (2016) 'Toward massive, ultrareliable, and low-latency wireless communication with short packets', *Proceedings of the IEEE*, Vol. 104, No. 9, pp.1711–1726.

Gu, B., Zhang, X., Lin, Z. and Alazab, M. (2020) 'Deep multiagent reinforcement-learning-based resource allocation for internet of controllable things', *IEEE Internet of Things Journal*, Vol. 8, No. 5, pp.3066–3074.

Haque, M., Tariq, F., Khandaker, M., Wong, K-K. and Zhang, Y. (2023) 'A survey of scheduling in 5G URLLC and outlook for emerging 6G systems', *IEEE Access*, Vol. 31, No. 43, pp.372–396.

He, Q., Zhu, Y., Zheng, P., Hu, Y. and Schmeink, A. (2021) 'Multi-device low-latency IoT networks with blind retransmissions in the finite blocklength regime', *IEEE Transactions on Vehicular Technology*, Vol. 70, No. 12, pp.782–795.

Hu, Y., Li, Y., Gursoy, M.C., Velipasalar, S. and Schmeink, A. (2020) 'Throughput analysis of low-latency IoT systems with QoS constraints and finite blocklength codes', *IEEE Transactions on Vehicular Technology*, Vol. 69, No. 3, pp.3093–3104.

Jiang, Y., Duan, H., Zhu, X., Wei, Z., Wang, T. and Zheng, F. (2021) 'Toward URLLC: a full duplex relay system with self-interference utilization or cancellation', *IEEE Wireless Communications*, Vol. 28, No. 1, pp.74–81.

Joachim, S., Gustav, W., Torsten, D., Robert, B. and Kittipong, K. (2018) '5G radio network design for ultra-reliable low-latency communication', *IEEE Network*, Vol. 32, No. 2, pp.24–31.

Kasgari, A., Saad, W., Mozaffari, M. and Poor, H.V. (2020) 'Experienced deep reinforcement learning with generative adversarial networks (GANs) for model-free ultra reliable low latency communication', *IEEE Transactions on Communications*, Vol. 69, No. 2, pp.884–899.

Khan, B., Jangsher, S., Ahmed, A. and Dweik, A. (2022) 'URLLC and eMBB in 5G industrial IoT: a survey', *IEEE Open Journal of the Communications Society*, Vol. 3, No. 11, pp.34–63.

Li, X., Li, J. and Liu, D. (2021) 'Energy-efficient UAV trajectory design with information freshness constraint via deep reinforcement learning', *Mobile Information Systems*, Vol. 2021, pp.1–9.

Li, Y., Hu, C., Wang, J. and Xu, M. (2019) 'Optimization of URLLC and eMBB multiplexing via deep reinforcement learning', *IEEE/CIC International Conference on Communications Workshops in China (ICCC Workshops)*.

Liu, Y., Deng, Y., Elkashlan, M., Nallanathan, A. and Karagiannidis, G.K. (2021) 'Analyzing grant-free access for URLLC service', *IEEE Journal on Selected Areas in Communications*, Vol. 39, No. 3, pp.741–755.

Mohamed, K. B. and Salim, B. (2021) 'Spectrum allocation and power control for D2D communication underlay 5G cellular networks', *International Journal of Communication Networks and Distributed Systems*, Vol. 27, No. 3, pp.299–322.

Naveen, K. and Ahmad, A. (2023) 'Deep reinforcement learning empowered energy efficient task-offloading in cloud-radio access networks', *International Journal of Communication Networks and Distributed Systems*, Vol. 29, No. 3, pp.341–358.

Park, J. and Bennis, M. (2018) 'URLLC-eMBB slicing to support VR multimodal perceptions over wireless cellular systems', *2018 IEEE Global Communications Conference (GLOBECOM)*, pp.1–7.

Pase, F., Giordani, M., Cuozzo, G., Cavallero, S., Eichinger, J., Verdone, R. and Zorzi, M. (2022) 'Distributed resource allocation for URLLC in IIoT scenarios: a multi-armed bandit approach', *2022 IEEE Globecom Workshops (GCWkshps)*, pp.383–388.

Polyanskiy, Y., Poor, H.V. and Verdu, S. (2010) 'Channel coding rate in the finite blocklength regime', *IEEE Transactions on Information Theory*, Vol. 56, No. 5, pp.2307–2359.

Ramin, H., Samad, A., Nurul, M. and Matti, L. (2021) 'Average rate and error probability analysis in short packet communications over RIS-aided URLLC systems', *IEEE Transactions on Vehicular Technology*, Vol. 70, No. 10, pp.320–334.

Ren, H., Pan, C., Deng, Y., Elkashlan, M. and Nallanathan, A. (2019) 'Joint power and blocklength optimization for URLLC in a factory automation scenario', *IEEE Transactions on Wireless Communications*, Vol. 19, No. 3, pp.1786–1801.

Sui, W., Chen, X., Zhang, S., Jiang, Z. and Xu, S. (2021) 'Energy-efficient resource allocation with flexible frame structure for hybrid eMBB and URLLC services', *IEEE Transactions on Green Communications and Networking*, Vol. 5, No. 1, pp.72–83.

Sulyman, A., Oteafy, S. and Hassanein, H. (2017) 'Expanding the cellular-IoT umbrella: an architectural approach', *IEEE Wireless Communications*, Vol. 24, No. 3, pp.66–71.

Sutton, J., Jie, Z., Liu, R. and et al. (2019) 'Enabling technologies for ultra-reliable and low latency communications: from PHY and MAC layer perspectives', *IEEE Communications Surveys and Tutorials*, Vol. 21, No. 3, pp.2488–2524.

Tan, J., Zhang, L. and Liang, Y. (2019) 'Deep reinforcement learning for channel selection and power control in D2D networks', *2019 IEEE Global Communications Conference (GLOBECOM)*, pp.1–6.

Tran, D., Sharma, S., Ha, V., Chatzinotas, S. and Woungang, I. (2023) 'Multi-agent DRL approach for energy-efficient resource allocation in URLLC-enabled grant-free NOMA systems', *IEEE Open Journal of the Communications Society*, Vol. 4, No. 14, pp.70–86.

Wang, L. and Zhang, H. (2019) 'Analysis of joint scheduling and power control for predictable URLLC in industrial wireless networks', *2019 IEEE International Conference on Industrial Internet (ICII)*, pp.160–169.

Wang, Y., Wu, H., Jhaveri, R. and Djenouri, Y. (2023) 'DRL-based URLLC-constraint and energy-efficient task offloading for internet of health things', *IEEE Journal of Biomedical and Health Informatics*, pp.1–12.

Wu, Y., Dinh, T., Fu, Y., Lin, C. and Quek, T. (2021) 'A hybrid DQN and optimization approach for strategy and resource allocation in MEC networks', *IEEE Transactions on Wireless Communications*, Vol. 20, No. 7, pp.4282–4295.

Xi, X., Cao, X., Yang, P., Chen, J., Quek, T. and Wu, P. (2021) 'Network resource allocation for eMBB payload and URLLC control information communication multiplexing in a multi-UAV relay network', *IEEE Transactions on Communications*, Vol. 69, No. 3, pp.1802–1817.

Yang, P., Xi, X., Quek, T., Cao, X. and Chen, J. (2021) 'URLLC-enabled UAV system incorporated with DNN-based channel estimation', *IEEE Wireless Communications Letters*, Vol. 10, No. 5, pp.1018–1022.

Yin, H., Cao, L. and Deng, X. (2021) 'Scheduling and resource allocation for multi-hop URLLC network in 5G sidelink', *2021 IEEE Vehicular Technology Conference (VTC-Fall)*, pp.1–7.

Zhao, T., Li, F. and He, L. (2023) 'DRL-based joint resource allocation and device orchestration for hierarchical federated learning in NOMA-enabled industrial IoT', *IEEE Transactions on Industrial Informatics*, Vol. 19, No. 6, pp.7468–7479.

Zhong, R., Liu, Y., Mu, X., Chen, Y. and Song, L. (2022) 'AI empowered RIS-assisted NOMA networks: deep learning or reinforcement learning?', *IEEE Journal on Selected Areas in Communications*, Vol. 40, No. 1, pp.182–196.

Zhou, Z., Wang, B., Gu, B., Ai, B., Mumtaz, S., Rodriguez, J. and Guizani, M. (2020) 'Time-dependent pricing for bandwidth slicing under information asymmetry and price discrimination', *IEEE Transactions on Communications*, Vol. 68, No. 1, pp.6975–6989.