Online Radio Pattern Optimization Based on Dual Reinforcement-Learning Approach for 5G URLLC Networks

Esswie, Ali Abdelmawgood Ali Ali; Pedersen, Klaus Ingemann; E. Mogensen, Preben

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2020.3011026

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 23, 2020, accepted July 14, 2020, date of publication July 21, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3011026

Online Radio Pattern Optimization Based on Dual Reinforcement-Learning Approach for 5G URLLC Networks

ALI A. ESSWIE, (Member, IEEE), KLAUS I. PEDERSEN, (Senior Member, IEEE), AND PREBEN E. MOGENSEN, (Member, IEEE)

Nokia Bell Labs, 9220 Aalborg, Denmark
Department of Electronic Systems, Aalborg University, 9100 Aalborg, Denmark

Corresponding author: Ali A. Esswie (ali.esswie@nokia-bell-labs.com)

This work was supported by the Innovation Fund Denmark (IFD) under Grant 7038-00009B.

ABSTRACT The fifth generation (5G) radio access technology is designed to support highly delay-sensitive applications, i.e., ultra-reliable and low-latency communications (URLLC). For dynamic time division duplex (TDD) systems, the real-time optimization of the radio pattern selection becomes of a vital significance in achieving decent URLLC outage latency. In this study, a dual reinforcement machine learning (RML) approach is developed for online pattern optimization in 5G new radio TDD deployments. The proposed solution seeks to minimizing the maximum URLLC tail latency, i.e., *min-max* problem, by introducing nested RML instances. The directional and real-time traffic statistics are monitored and given to the primary RML layer to estimate the sufficient number of downlink (DL) and uplink (UL) symbols across the upcoming radio pattern. The secondary RML sub-networks determine the DL and UL symbol structure which best minimizes the URLLC outage latency. The proposed solution is evaluated by extensive and highly-detailed system level simulations, where our results demonstrate a considerable URLLC outage latency improvement with the proposed scheme, compared to the state-of-the-art dynamic-TDD proposals.

INDEX TERMS Dynamic-TDD, URLLC, 5G new radio, machine learning, reinforcement learning, Q-learning, cross link interference (CLI).

I. INTRODUCTION

One of the main drivers of the fifth generation (5G) radio standardization is the ultra-reliable and low-latency communications (URLLC) service class [1]. URLLC entail the transmission of sporadically-arriving small-payload packets with one-way radio latency of 1 ms and 99.999% success probability [2]. As the early 5G commercial enrollments are foreseen over the 3.5 GHz unpaired spectrum, due to its wide spectrum availability [3], time-division duplexing (TDD) technology is vital for the success of the 5G. With dynamic TDD, base-stations (BSs) independently utilize either a downlink (DL) or uplink (UL) transmission opportunity at a time in order to meet their capacity and latency demands, respectively [4].

Achieving the URLLC targets for dynamic TDD deployments is highly challenging [5] due to: (i) the non-concurrent availability of the DL and UL transmission opportunities, and (ii) the potentially strong cross-link interference (CLI) between neighboring BSs and user-equipment's (UEs), adopting opposite transmission directions. The fine selection of the DL and UL symbol structure during a TDD radio pattern has been demonstrated to immensely impact the achievable URLLC outage latency, even under hypothetically CLI-free conditions [5]. Moreover, the TDD radio pattern selection is an NP-hard problem for multi-cell multi-UE deployments, due to the simultaneous requests of conflicting link directions, and thus, this is the problem addressed in this work.

The associate editor coordinating the review of this manuscript and approving it for publication was Xijun Wang.

A. STATE OF THE ART DYNAMIC TDD STUDIES

The third generation partnership project (3GPP) has recently standardized a flexible frame structure for dynamic TDD 5G systems [6]. That is, BSs configure a 10-ms radio frame, consisting of multiple slot formats, each is composed of DL [D], UL [U], and flexible [F] symbols, respectively. The latter indicates the symbol set that can be dynamically configured, through a dedicated radio signaling from BS to UEs, either as DL or UL or act as a guard time among successive DL and UL symbols, respectively. Such design

offers a highly resilient framework for adapting the radio patterns to the time-variant offered traffic needs. One simple way to approach such frame flexibility is to semi-statically adapt the radio pattern configuration to the current average traffic conditions [7]. In particular, a common radio pattern is periodically updated and adopted by all neighboring BSs in order to meet the average network capacity demands, with minimal inter-BS signaling overhead.

Recent prior-art proposals seek to utilize the standardized pattern update flexibility. In [8], [9], a predefined set of radio frame configurations is adopted, with different possible DL and UL symbol ratios and pre-determined structures (aka - a frame-book). Thus, BSs dynamically select those patterns from the frame-book which best satisfy their individual link selection criteria, e.g., the currently buffered traffic.

However, as a consequence to the BS-specific pattern adaptation, neighboring BSs may simultaneously adopt opposite link directions, resulting in a severe CLI. For instance, the BS-BS CLI is demonstrated as a fundamental limitation of the achievable UL capacity [5], mainly due to the larger DL transmit power compared to the victim UL power. CLI mitigation and coordination schemes have therefore been widely investigated over recent prior art. In [10]–[12], coordinated cross-cell beam-forming, UL power control and cell muting are proposed to limit the residual network CLI, especially towards the more CLI-sensitive cell-edge UEs. Joint UL transceiver design [10], [13], [14], based on inter-cell signaling of the UEs' spatial signatures, is also introduced in order to isolate the BS-BS CLI spatial subspace from that is of the desired UL transmission. The drawback of those proposals is mainly the requirement of a large inter-cell signaling overhead space. Therefore, simpler and less-coordination-overhead demanding opportunistic CLI avoidance schemes [15], [16] have been suggested to offer attractive capacity and latency merits, where the BS-BS and UE-UE CLI is pre-averted on a best effort basis. This encompasses the design of a hybrid TDD pattern with a slot-aware dynamic UE scheduling. Although those proposals require simpler implementation complexity, they optimize the URLLC performance on a heuristic basis, which may jeopardize the achievable URLLC reliability and latency performance.

B. MACHINE LEARNING POTENTIAL IN DYNAMIC TDD SYSTEMS

Although the quoted TDD studies present clear advancements and valuable findings, the radio pattern selection procedure is yet deemed as a challenging problem towards the success of the 5G TDD deployments. This is particularly relevant for dynamic URLLC multi-cell multi-user TDD deployments, where the DL and UL traffic arrivals are highly sporadic in time, and with strict latency and reliability constraints. As stated, the problem of selecting the optimum TDD switching pattern is NP-hard and has so far been addressed by means of rather simple heuristic solutions. In this study, we go one step further where our hypothesis is that machine learning (ML) is a viable solution to be utilized at the BS nodes to dynamically select the best possible TDD switching pattern. That is, based on monitoring the past and current traffic and latency performance per BS, an ML capability shall learn and predict the best TDD switching pattern for the next radio frame.

ML techniques have been notably studied with the 5G wireless radio communications [17] for various radio design aspects such as interference management [18] and radio resource management [19]–[23]. Generally, ML can be divided into three categories [24] as: (1) supervised-ML (SML), where the input data is a priori known and well-labeled for model training. The SML model is continuously trained with the right *question-answer* pairs until it approaches the optimal model, (2) unsupervised-ML (UML), where the input data is neither a priori known nor labeled. Accordingly, data clustering and dimensionality reduction become necessary to extract the meaningful and independent feature vectors, and (3) reinforcement-ML (RML), where unlike SML and UML, it does not require offline model training. Thus, RML has been widely employed towards the real-time decision-making applications. RML algorithms are goal-oriented which consistently in time learn how to achieve a complex objective, through an iterative; however, simple, process of action exploration and environment observation, respectively. The model-free RML algorithms are mainly categorized to on-policy and off-policy techniques [25], respectively. The former directly learns the optimal policy while the latter approaches the near-optimal policy through more conservative exploration. On-policy ML algorithms, such as state–action–reward–state–action (SARSA) [26], have been demonstrated particularly attractive for the critical use cases where the learning agent is critically challenged with a tight training duration, and over which it cannot employ a sub-optimal policy, e.g., walking robots over a cliff.

For the latency-critical URLLC traffic, SML and UML are substantially challenging for practical deployments due to the required large size of dedicated training samples to reach a sufficient learning of the target URLLC 10^{-5} outage probability. Therefore, SML and UML methods are not adopted in this study as deemed too demanding for achieving the required level of model training. We prioritize RML as being more suitable for the type of system and objectives addressed in this paper, and hence, this is the focus of this study.

C. PAPER CONTRIBUTION

In this paper, a dual-RML based pattern optimization scheme is proposed for dynamic TDD 5G systems. The proposed solution targets minimizing the inflicted URLLC radio latency on a real-time basis, and accordingly, improving the achievable URLLC outage performance. The proposed scheme utilizes nested RML layers, where the primary layer estimates the number of the DL and UL symbols of the upcoming radio pattern to satisfy the foreseen offered traffic. Subsequently, the secondary RML sub-layers determine the DL and UL symbol structure that achieves the minimum

possible URLLC radio latency. The proposed algorithm neither requires inter-cell signaling exchange overhead nor offline dedicated training, i.e., online and distributed pattern optimization. Performance results show a significant improvement of the URLLC outage latency with the proposed solution, compared to state-of-the-art dynamic TDD proposals. The major contributions of this paper are listed as follows:

- We propose a novel dual reinforcement machine learning (RML) approach for online URLLC outage optimization for 5G-NR TDD networks.
- Unlike the state-of-the-art relevant TDD solutions [7]–[18], the proposed solution considers the joint capacity and latency statistics to optimize the URLLC outage latency performance. It is fully compliant with the current 3GPP 5G-NR standard specifications for dynamic-TDD deployments. The proposed framework neither requires inter-cell signaling exchange nor high processing complexity.
- Compared to the state-of-the-art TDD literature, the proposed scheme offers a considerable URLLC latency and reliability enhancement, under various DL and UL offered loads. It achieves 70% outage latency reduction compared to the standard dynamic TDD scheme.

Due to the complexity of the 5G new radio system design and the addressed problems herein, the proposed solution has been evaluated by extensive and highly-detailed system level simulations. Those simulations incorporate the major functionalities of the 5G new radio protocol stack, e.g., dynamic resource allocation and user scheduling, adaptive modulation and coding schemes (MCS), hybrid automatic repeat request (HARQ) re-transmissions, and the 3GPP 3D spatial channel modeling, respectively. Special care is given to ensure statistically-reliable results.

The paper is organized as follows. Section II presents the system modeling, while Section III formulates the problem addressed in this work. Section IV introduces a brief overview of the Q-reinforcement learning and Section V presents the detailed description of the proposed solution. Section VI introduces the state-of-the-art dynamic-TDD schemes, against which we evaluate the performance of the proposed solution. The performance evaluation results appear in Section VII, while conclusions are drawn in Section VIII.

II. SETTING THE SCENE

A. SYSTEM MODEL

We consider a macro 5G dynamic TDD deployment, where base-stations (BSs) are configured with 3-sector cell setting. Thus, there are a total of C cells, each is equipped with N antennas. Each cell serves an average of $K = K^{\mathrm{dl}}+K^{\mathrm{ul}}$ uniformly-distributed UEs, each equipped with M antennas, K^{dl} and K^{ul} denote the average numbers of the DL and UL UEs per cell. In this study, we assume that UEs are requesting DL and UL traffic with different DL and UL packet arrival rates, respectively. We adopt the URLLC-alike FTP3 traffic modeling with packet sizes of f^{dl} and f^{ul} bits, and a Poisson Arrival Process, with mean packet arrival rates of λ^{dl} and λ^{ul}, in the DL and UL directions, respectively [27]. The average offered load per cell in the DL direction is: $\Omega^{\mathrm{dl}} = K^{\mathrm{dl}} \times f^{dl} \times \lambda^{\mathrm{dl}}$, and UL direction: $\Omega^{\mathrm{ul}} = K^{\mathrm{ul}} \times f^{ul} \times \lambda^{\mathrm{ul}}$. The total offered load per cell is given as: $\Omega = \Omega^{\mathrm{dl}} + \Omega^{\mathrm{ul}}$.

We follow the 3GPP guidelines for the 5G TDD system modeling, as shown by Fig. 1. UEs are dynamically multiplexed using the orthogonal frequency division multiple access (OFDMA). In line with the 3GPP URLLC studies [27], the SCS is selected to equal 30 kHz as it offers sufficiently short symbol durations to fulfill the considered latency requirements, while still having enough cyclic prefix duration to cope with time-dispersion for the considered macro scenario, with the physical resource block (PRB) of twelve consecutive SCSs. Furthermore, we assume a short transmission time interval (TTI) duration of 4 OFDM symbols towards faster URLLC transmissions. Prior to the start of each radio frame [28], i.e., every 10-ms, the BS decides the next radio frame pattern based on the proposed RML solution. In this work, we assume a single guard OFDM symbol between every DL and UL symbols in the radio frame, in order to account for the DL channel delay spread before the UL transmissions are triggered.

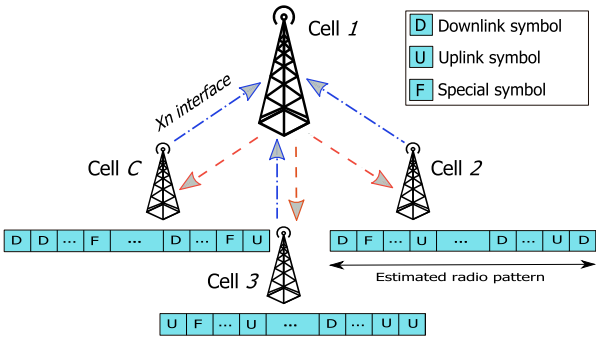


FIGURE 1. System model: dynamic-TDD macro deployment.

Accordingly, when a DL packet arrives at the cell, it is first processed by the serving cell, and thereafter, is buffered towards the first available DL transmission opportunity of the current TDD radio pattern, i.e., TDD pattern switching delay. The time to prepare a DL transmission block is taken explicitly into account in line with 3GPP 5G-NR specifications [29]. Then, the cell scheduler dynamically multiplexes all pending DL packets using the proportional fair criterion, where some DL packets can be further queued to the next DL transmission instant, i.e., scheduling queuing delay. The HARQ re-transmissions are always prioritized over new transmissions. Herein, dynamic link adaptation is also adopted, where the DL transmission MCS is adaptively selected such that it corresponds to a first-transmission block error rate (BLER) of 1%. The MCS selection is typically based on the most recently received channel quality indication (CQI) report from the UE. The scheduled users are notified with a scheduling grant (aka DL control information – DCI), and the overhead from the corresponding

physical-layer control signaling is taken explicitly into account in line with [30]. At the UE-side, DL reception is subject to processing time for decoding of the DL transmission. In case the transmitted DL packet is not successfully decoded by the intended UE, the UE triggers the transmission of a HARQ negative acknowledgment (NACK) during the next available UL transmission opportunity of the radio pattern, where the appropriate radio resources are allocated. Correspondingly, serving cell re-transmits the respective DL packet to be soft-combined at the UE.

For UL packet transmissions, we assume configured grant (CG) transmission (aka grant-free) with fixed MCS per UE [31], The use of CG means that as soon as a packet arrives at the UE, it is immediately prepared for UL transmission, and transmitted at the first coming UL TTI opportunity. Each CG transmission includes a robust preamble, so the receiving BS is able to detect from which UE the transmission is coming. The CG parameterization is such that UEs with high path-loss are transmitting on the full bandwidth with a conservative MCS corresponding to QPSK rate $\frac{1}{8}$, such that one URLLC payload of 32 bytes can be transmitted. In line with [31], UEs with better path-loss conditions are configured to transmit on one quarter of the carrier bandwidth with MCS QPSK rate $\frac{1}{2}$. Such UE classification, of high or low path-loss conditions, is based on a predefined coupling gain threshold $\hat{c}$. The UL transmit power $\Sigma\,[dBm]$ is configured to equal

$$\Sigma\,[dBm] = \min\left\{\Sigma_{\max},\ P0 + 10\log_{10}(\wp) + \alpha\eth + \nabla_{\text{MCS}}\right\}, \quad (1)$$

where $\Sigma_{\max}$ is the max UE transmit power, $P0$ is the target power spectral density, $\wp$ is the number of granted UL PRBs, α and $\eth$ denote the path-loss compensation factor and path-loss, respectively. ∇_{MCS} is an UL power boost factor where $\nabla_{\text{MCS}} = 10$ dB for QPSK1/2 and $\nabla_{\text{MCS}} = 0$ dB for QPSK1/8 in line with [31]. As CG transmissions from multiple users may occur at the same time on overlapping resources, uplink transmissions from UEs are subject to potential intra-cell interference, which only to a certain extent can be combated by the a linear BS multi-antenna receiver. If the BS fails to correctly decode a CG transmission from an UL UE, it immediately sends an uplink scheduling grant for the UE in the next coming DL TTI, issuing an UL HARQ re-transmission from the UE in the next UL TTI. The UL HARQ re-transmission is sent using the same configuration (bandwidth and MCS configurations) as the original transmission, but with a +3 dB transmission power boost to enhance the probability of decoding the HARQ re-transmission at the BS [31].

As an input to the proposed RML algorithm to dynamically select the radio frame configuration, cells should be aware of the directional traffic and latency statistics. Hence, in this work, we assume a realistic knowledge of those statistics at the cell side. Particularly, the DL traffic size, including buffered and new packets, is spontaneously known at the cell stack. However, in the UL direction, new UL packet transmissions are not a priori known at the cell. Those are only identified at the cell side when the first UL transmission attempt is either failed or correctly received. Thus, we only assume the UL HARQ-buffered traffic size is known at the cell side.

For capturing the latency statistics of the corresponding DL/UL buffers, we define the head of line delay (HoLD) per packet per UE as the time from the moment a DL/UL packet arrives at the transmitter packet data convergence protocol (PDCP) layer until it is successfully received at the receiver end, and forwarded to the PDCP layer. The exact DL HoLD is known at the cell.

For the UL direction, it is not known at the BS-side when a packet arrives at the UE-side. The BS only becomes aware of pending UL transmissions from the UE when it first tries to transmit those to the BS. The UL HoDL is therefore only monitored at the BS-side as the time from the first UL transmission attempt until successful decoding, i.e., essentially corresponding to the effective HARQ retransmission round trip time. Due to the adaptation of the TDD switching pattern and presence of both inter-cell and intra-cell interference from other UEs, as well as the potential BS-BS CLI, the UL HARQ round trip time is time-variant, and often dominant for the tail of the UL packet distribution.

Finally, the achievable one-way radio URLLC latency at the 10^{-5} outage probability is the main performance metric [5] of this work. It implies the delay from the moment when a URLLC packet arrives at the packet data convergence protocol layer of the transmitter until it is successfully received at the intended receiver, summing the BS and UE processing delays, buffering delay due to dynamic UE scheduling, delay to the first DL/UL transmission opportunity, and HARQ re-transmission delay.

B. SIGNAL MODEL

Assume $\mathfrak{B}_{\text{dl}}$, $\mathfrak{B}_{\text{ul}}$, $\mathcal{K}_{\text{dl}}$ and $\mathcal{K}_{\text{ul}}$ as the BS and UE sets with DL and UL transmissions, respectively. Thus, the DL signal at the k^{th} UE, where $k \in \mathcal{K}_{\text{dl}}$, $c_k \in \mathfrak{B}_{\text{dl}}$, is given as

$$\mathrm{y}^{\text{dl}}_{k,c_k} = \underbrace{\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{h}_k x_k}_{\text{Useful signal}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}\setminus k}\mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{h}_i x_i}_{\text{BS to UE interference}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}}\mathbf{G}_{k,j}\mathbf{o}_j x_j}_{\text{UE to UE interference}} + \mathbf{n}^{\text{dl}}_k, \quad (2)$$

where $\mathbf{H}^{\text{dl}}_{k,c_i} \in \mathcal{C}^{M\times N}$ is the DL 3D-UMA fading channel [32] from the cell serving the i^{th} UE, to the k^{th} UE, $\mathbf{h}_i \in \mathcal{C}^{N\times 1}$, $\mathbf{o}_k \in \mathcal{C}^{M\times 1}$ and x_k are the zero-forcing precoding vector at the c_i^{th} BS, precoding vector of the k^{th} UE, and the transmitted data symbol of the k^{th} UE, respectively, while $\mathbf{G}_{k,j} \in \mathcal{C}^{M\times M}$ implies the cross-link channel between the k^{th} and j^{th} UEs, and $\mathbf{n}^{\text{dl}}_k$ represents the additive white Gaussian noise. The UL signal at the c_k^{th} cell, $c_k \in \mathfrak{B}_{\text{ul}}$ from $k \in \mathcal{K}_{\text{ul}}$, is expressed by

$$\mathrm{y}^{\text{ul}}_{c_k,k} = \underbrace{\mathbf{H}^{\text{ul}}_{c_k,k}\mathbf{o}_k x_k}_{\text{Useful signal}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}\setminus k}\mathbf{H}^{\text{ul}}_{c_k,j}\mathbf{o}_j x_j}_{\text{UE to BS interference}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}}\mathbf{P}_{c_k,c_i}\mathbf{h}_i x_i}_{\text{BS to BS interference}} + \mathbf{n}^{\text{ul}}_{c_k}, \quad (3)$$

where $\mathbf{P}_{c_k,c_i} \in \mathcal{C}^{N\times N}$ is the BS-BS channel between the serving BSs of the k^{th} and i^{th} UEs, $k \in \mathcal{K}_{\text{ul}}$ and $i \in \mathcal{K}_{\text{dl}}$. Then, the post-receiver signal-to-interference ratio in the DL γ_k^{dl} and UL $\gamma_{c_k}^{\text{ul}}$ directions are expressed by,

$$\gamma_k^{\text{dl}} = \frac{\left\| \left(\mathbf{u}_k^{\text{dl}}\right)^{\text{H}} \mathbf{H}_{k,c_k}^{\text{dl}} \mathbf{h}_k \right\|^2}{\sum\limits_{i\in\mathcal{K}_{\text{dl}}\setminus k} \left\| \left(\mathbf{u}_k^{\text{dl}}\right)^{\text{H}} \mathbf{H}_{k,c_i}^{\text{dl}} \mathbf{h}_i \right\|^2 + \sum\limits_{j\in\mathcal{K}_{\text{ul}}} \left\| \left(\mathbf{u}_k^{\text{dl}}\right)^{\text{H}} \mathbf{G}_{k,j} \mathbf{o}_j \right\|^2}, \tag{4}$$

$$\gamma_{c_k}^{\text{ul}} = \frac{\left\| \left(\mathbf{u}_k^{\text{ul}}\right)^{\text{H}} \mathbf{H}_{c_k,k}^{\text{ul}} \mathbf{o}_k \right\|^2}{\sum\limits_{j\in\mathcal{K}_{\text{ul}}\setminus k} \left\| \left(\mathbf{u}_k^{\text{ul}}\right)^{\text{H}} \mathbf{H}_{c_k,j}^{\text{ul}} \mathbf{o}_j \right\|^2 + \sum\limits_{i\in\mathcal{K}_{\text{dl}}} \left\| \left(\mathbf{u}_k^{\text{ul}}\right)^{\text{H}} \mathbf{P}_{c_k,c_i} \mathbf{h}_i \right\|^2}, \tag{5}$$

where $\|\cdot\|^2$ is the second-norm, $\mathbf{u}_k^{\kappa} \in \mathcal{C}^{N/M\times 1}$, $\mathcal{X}^{\kappa}$, $\kappa\in\{\text{ul, dl}\}$, is the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver vector [33], with $(\cdot)^{\text{H}}$ as the Hermitian operation.

III. PROBLEM FORMULATION

The URLLC applications require a stringent radio latency bound and with a rare per-packet violation probability. In dynamic TDD systems, the URLLC outage performance is dominated by the number and the structure of the DL d_c and UL u_c symbols across the configured radio pattern. In this study, our objective is to optimize the radio pattern configuration, i.e., to determine the number and structure of d_c and u_c, for a faster and DL-UL balanced traffic transmission, and thus, an improved URLLC outage latency, as

$$\left(\frac{d_c}{u_c}\right)^* \triangleq \left\{ \frac{d^i}{u^i} \ : \ \frac{d^i}{u^i} \in \mathfrak{T} \right\}, \tag{6.a}$$

$$\left(\hat{w}_c\right)^* \triangleq \left\{ \hat{w}^j \ : \ \hat{w}^j \in \hat{\mathrm{W}} \right\}, \tag{6.b}$$

$$\textit{Subject to}: \begin{cases} \arg\min\limits_{c,t} \left(\Upsilon_{c,t}\right) \\ \arg\min\limits_{k} \left(\varphi_{c,k}\right), \ \forall k \in \mathcal{K}_{\text{ul/dl}} \end{cases} \tag{6}$$

where $\mathfrak{T}$ and $\hat{\mathrm{W}}$ are the inclusive sets of all possible $\frac{d_c}{u_c}$ ratios and structures, respectively. $\Upsilon_{c,t}$ denotes the buffered traffic difference of the c^{th} cell at time t between the amount of buffered DL and UL traffic volume, and $|\cdot|$ denotes the absolute value. $\varphi_{c,k}$ indicates the achievable one-way radio latency of the k^{th} UE.

The first constraint (6.a) implies that the selected TDD pattern at an arbitrary time should contribute to closing the gap among the buffered DL and UL traffic size over the pattern duration, regardless of the variant DL and UL PRB capacity and the offered traffic ratio $\frac{\Omega^{\text{dl}}}{\Omega^{\text{ul}}}$. The second constraint (6.b) ensures that the UE-specific latency performance is monotonically optimized.

IV. OVERVIEW OF THE Q REINFORCEMENT MACHINE LEARNING (Q-RML)

The RML [34], [35] is a vital branch of the machine learning. It has been widely applied in real-time decision-making problems such as autonomous driving and robot control. RML follows the mathematical framework of the Markov decision process [36], where the learning outcomes are partially random and tightly related to the environment. Accordingly, the goal of an RML agent is to obtain an optimal policy $\pi^* : S \rightarrow A$, which determines an action $a \in A$ under state $s \in S$, thus, to optimally maximize or minimize a pre-defined value function V^{π}. The value function is typically expressed in terms of the expected discounted cumulative reward or penalty at time epoch t, as

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma \ r_t(s_t, a_t) \,|s_0 = s \right], \tag{7}$$

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[r_t(s_t, a_t) + \gamma \ V^{\pi}(s_{t+1}) \,|s_0 = s \right], \tag{8}$$

where $\mathbb{E}(\cdot)$ implies the statistical mean, $r_t(s_t, a_t)$ is the immediate reward or penalty, observed from the environment after taking an action a under state s at time epoch t, and $\gamma \in [0, 1]$ is the discount factor on future rewards or penalties. Simple dynamic programming schemes can be utilized to solve eq. (7), when the state transition probabilities are a priori known. The RML aims to finding the optimal policy π^* when the system dynamics are not known through an iterative process of continuously adjusting its policy. In that sense, Q-RML is one of the most effective RML techniques. In this study, we adopt the baseline off-policy Q-learning approach to rapidly learn the optimal policy during the warm-up time. Thus, unlike the case with the on-policy RML techniques, we preserve a sufficiently enough pre-training time in order for the Q-RML approach to converge to the optimal greedy policy before impacting the actual inference performance.

A Q-RML agent applies the actions which closes the gap between the current policy π and the optimal target policy π^*, i.e., $\pi \xrightarrow{t} \pi^*$, such that the observed reward or penalty from the environment is monotonically optimized as

$$V(s_t) = F\left[r_t(s_t, a_t) + \gamma \ V^{\pi}(s_{t+1}) \right], \tag{9}$$

where the optimization function F is the optimization function, which defines the Q-RML learning goal, in terms of the corresponding value function, as given by

$$F \cong \begin{cases} \arg\max\limits_{a\in A} \left(\mathscr{F}\right), V(s) \rightarrow \text{reward} \\ \arg\min\limits_{a\in A} \left(\mathscr{F}\right), V(s) \rightarrow \text{penality} \end{cases}, \tag{10}$$

where $\mathscr{F}$ is the actual environment value function of the Q-RML instance, defined as a reward or penalty.

V. PROPOSED RML BASED PATTERN OPTIMIZATION

The proposed solution incorporates a dual RML approach for joint capacity and latency online optimization. We consider the model-free Q-reinforcement-machine-learning (Q-RML)

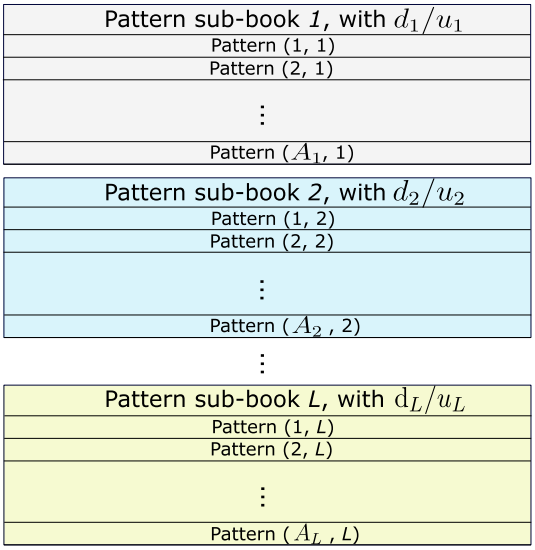


FIGURE 2. Proposed algorithm: pattern-book design.

algorithm [34] for its performance merits and low implementation complexity under a moderate state-action space size. As depicted by Fig. 2, a pattern-book is constructed, where there are L pattern sub-books, each is of a size $\mathbf{Card}\,(\mathcal{A}_l)$ radio patterns, where $\mathbf{Card}\,(\cdot)$ denotes the cardinality of a set, and $\mathcal{A}_l$ is the set of radio patterns in the l^{th} sub-book, $\forall l \in L$. All radio patterns within a single sub-book share the same d^l/u^l symbol ratio; although, with different symbol structures. The nested pattern book design allows for utilizing independent Q-RML instances to estimate the DL and UL symbol ratio as well as the respective symbol structure.

As depicted by Fig. 3, the primary Q-RML network, i.e., $Q{-}1$, estimates the number of the DL d_c and UL u_c symbols of the upcoming radio pattern. The $Q{-}1$ target is to select the symbol ratio which contributes into a faster; though, balanced DL and UL, traffic service over the pattern duration; however, adopting a default symbol structure. Then, the secondary Q-RML sub-networks, i.e., $Q{-}2{-}l$, determine the best possible DL and UL symbol structure, following the calculated $\frac{d^l}{u^l}$ ratio from $Q{-}1$, in order to minimize the filtered HoLD statistics, leading to a significantly improved and DL/UL fair URLLC outage performance. In the following, we represent the sole operation of $Q{-}1$ layer as Algorithm-1, and as Algorithm-2 when $Q{-}1$ and $Q{-}2{-}l$ layers are simultaneously incorporated.

A. PRIMARY Q-RML NETWORK FOR BALANCED DL/UL BUFFERING

In dynamic TDD macro systems, the achievable UL capacity is highly variant from the corresponding DL capacity, mainly due to the severe BS-BS CLI. For instance, a linear mapping from $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 1/1$ to $d^c/u^c = /1/1$ may not be sufficient. Accordingly, the TDD pattern adaptation process becomes fully dictated by the residual UL traffic, i.e., including buffered, and re-transmitted traffic, leading to a highly degraded DL capacity accordingly due to the subsequent starvation of the DL transmission opportunities. Thus, the Algorithm-1 RML instance seeks a rapid; but; balanced DL and UL traffic transmission, by estimating the sufficient d^c/u^c ratio for a given DL and UL traffic statistic every radio pattern duration.

In that regard, at the ς^{th} slot of the radio pattern, $\varsigma = 1, 2, \ldots, \xi$, with ξ as the number of slots per the configured radio pattern, the relative traffic ratio $\mu_{[t,c]}\,(\varrho)$ of the c^{th} BS at time epoch t is defined as

$$\mu_{[t,c]}\,(\varsigma) = \frac{Z^{\mathrm{dl}}_{[t,c]}\,(\varsigma)}{Z^{\mathrm{dl}}_{[t,c]}\,(\varsigma) + \left(\frac{1}{\iota}\right) Z^{\mathrm{ul}}_{[t,c]}\,(\varsigma)}, \tag{11}$$

where $Z^{\mathrm{dl}}_{[t,c]}\,(\varsigma)$ and $Z^{\mathrm{ul}}_{[t,c]}\,(\varsigma)$ are the aggregated DL and UL buffered traffic size of the ς^{th} slot during the current pattern, and ι is the first-transmission average UL BLER, experienced at the BS side. As discussed in Section II.A, $Z^{\mathrm{ul}}_{c}\,(\varsigma)$ implies only the UL HARQ-buffered packets. Accordingly, to ensure

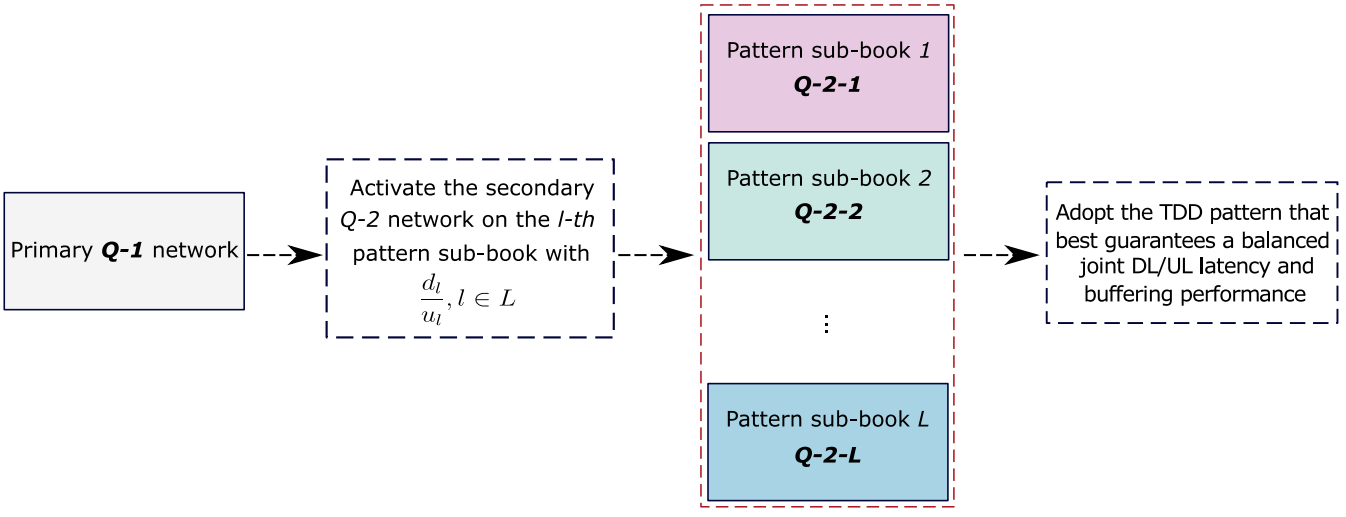


FIGURE 3. Proposed algorithm: nested Q-network design for TDD pattern selection.

fairness against $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$, the average ι is incorporated in eq. (11) such that the term $\left(\frac{1}{\iota}\right) Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ reflects the average total UL offered traffic size at the BS. The instantaneous traffic ratios $\mu_{[t,c]}(\varsigma)$ are linearly averaged over the duration of the TDD pattern as

$$\overline{\mu}_{[t,c]} = \frac{1}{\xi} \sum_{\varsigma=1}^{\xi} \mu_{[t,c]}(\varsigma), \tag{12}$$

with $\overline{\mu}_{[t,c]}$ as the relative traffic ratio at time epoch t. The traffic ratio $\overline{\mu}_{[t,c]} \rightarrow [0,1]$ reflects the combined buffering performance of the DL and UL traffic. For instance, $\overline{\mu}_{[t,c]} = 0.1$ denotes that the buffered UL traffic is 9x times the DL traffic. Accordingly, a state space $S^{(1)}$ is defined to represent the DL and UL traffic buffering conditions at an arbitrary time epoch t, as

$$S^{(1)}_t = \left\{ s^{(1)}_{1,t},\ s^{(1)}_{2,t},\ \ldots,\ s^{(1)}_{\mathfrak{J}_1,t} \right\}, \tag{13}$$

with $\mathfrak{J}_1$ as the size of the $Q-1$ state space. In principal, the state of the learning agent is determined as a function of the input performance metric, by an arbitrary mapping structure. In this work, we adopt a linear mapping of the quantized traffic volume to determine the BS state. Accordingly, the traffic-to-state mapping is designed as

$$s^{(1)}_t = \begin{cases} s^{(1)}_{1,t}, & \overline{\mu}_{[t,c]} < \mu_{\min} \\ s^{(1)}_{2,t}, & \mu_{\min} \leq \overline{\mu}_{[t,c]} < \mu_{\min}+\sigma \\ s^{(1)}_{3,t}, & \mu_{\min}+\sigma \leq \overline{\mu}_{[t,c]} < \mu_{\min}+2\sigma \\ \vdots & \vdots \\ s^{(1)}_{\mathfrak{J}_1,t}, & \overline{\mu}_{[t,c]} \geq \mu_{\max}, \end{cases} \tag{14}$$

where the traffic ratio quantization step σ is given as:

$$\sigma = \frac{\mu_{\max}-\mu_{\min}}{\mathfrak{J}_1-2}, \tag{15}$$

where $\mu_{\max}$ and $\mu_{\min}$ indicate the pre-defined minimum and maximum allowable levels of the traffic ratio $\overline{\mu}_{[t,c]}$. In that sense, $s^{(1)}_{1,t}$ indicates a traffic state where the buffered UL traffic is much larger than of the DL direction. Thus, an intermediate state is the system favorable target state to offer a balanced DL and UL buffering performance.

The action space $A^{(1)}$ is constructed to represent the set of all possible Algorithm-1 outcomes as

$$A^{(1)}_t = \left\{ a^{(1)}_{1,t},\ a^{(1)}_{2,t},\ \ldots,\ a^{(1)}_{L,t} \right\}, \tag{16}$$

where $a^{(1)}_{l,t} \equiv d^l/u^l, \forall l \in L$. Particularly, the Algorithm-1 instance determines the pattern sub-book, and hence, the corresponding $/d^l/u^l$ ratio, to be adopted over the upcoming radio pattern. Herein, we assume the immediate environment return $\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$ of Algorithm-1 represents a performance penalty, as

$$\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) = \left|\overline{\mu}_{[t,c]} - \mathfrak{m}^{(1)}\right|, \quad \forall i \in \mathfrak{J}_1,\ l \in L, \tag{17}$$

where $\mathfrak{m}^{(1)}$ denotes the mean value of the traffic ratio distribution $\overline{\mu}_{[t,c]}$. The mean value of the buffered traffic ratio $\mathfrak{m}^{(1)}$ is selected as the target of the primary Q-RML learning, since it allows for selecting the TDD pattern, with a certain DL-to-UL symbol ratio that is likely to preserve a balanced downlink and uplink buffered traffic performance. Specifically, $\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$ indicates the immediate cost, observed from the environment upon taking an action $a^{(1)}_{l,t}$ under state $s^{(1)}_{i,t}$, and is calculated in terms of how much deviant the traffic ratio $\overline{\mu}_{[t,c]}$ is from its balanced mean $\mathfrak{m}^{(1)}$. That is, a large Θ implies either unfavorable much buffered DL or UL traffic. At an arbitrary time epoch, the Algorithm-1 instance selects the action $a^{(1)}_{l,t} \equiv \frac{d^l}{u^l}$ which best minimizes the immediate cost as

$$\left(a^{(1)}_{l,t}\right)^* = \underset{a_{l,t} \in A^{(1)}}{\arg\min}\ \Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right). \tag{18}$$

Furthermore, the $\epsilon-$greedy policy is adopted to trade-off action exploration versus exploitation. Thus, at each step, a random number is drawn from a uniform distribution $\varrho^{(1)} \in \mathcal{U}(0,1)$, and is compared against the pre-defined exploration probability $\epsilon^{(1)}$. If $\varrho^{(1)} \leq \epsilon^{(1)}$ is satisfied, a random action is selected; otherwise, a greedy action according to eq. (18) is adopted. Finally, the value function entries $Q^{(1)}_{[t,c]}$ are iteratively updated to reflect the learning experiences as follows:

$$Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) \leftarrow \left(1-\alpha^{(1)}\right) Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) + \alpha^{(1)} \left[\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) + \gamma^{(1)} \underset{a_l \in A^{(1)}}{\arg\min}\ Q^{(1)}_{[t+1,c]}\left(s^{(1)}_{i,t+1}, a^{(1)}_{l}\right)\right], \tag{19}$$

where $\alpha^{(1)} \rightarrow [0,1]$ is the learning rate, which specifies how fast the learning occurs. For instance, if $\alpha^{(1)}$ is small, the learning rate of Algorithm-1 network shall exhibit a longer convergence time. $\gamma^{(1)} \rightarrow [0,1]$ implies the discounted factor, which determines how much significance is considered on future penalties. If $\gamma^{(1)}$ is large, the Algorithm-1 RML instance is biased towards adopting actions at time epoch t, which are highly probable to result in a further favorable state at $t+1$. The detailed primary RML network is summarized in Algorithm-1.

B. SECONDARY Q-RML SUB-NETWORKS FOR URLLC LATENCY MINIMIZATION

After the DL-to-UL symbol ratio $\frac{d^l}{u^l}$ is estimated from Algorithm-1 (layer 1), the corresponding Algorithm-2 sub-network is activated to estimate the best DL and UL symbol structure $\hat{w}_c$. For that, the DL and UL buffer latency samples per UE are monitored. Although, having monitored the latency for all the DL and UL packets for all active UEs in each cell represents a significant amount of statistics. Those samples are therefore further compressed into a more manageable metric that is meaningful for Algorithm-2 to learn and predict the best DL and UL symbol structure to minimize the overall cell latency outage performance. In this

Algorithm 1 Algorithm-1 for Balanced DL/UL Capacity

Initialize:
  for each $s^{(1)} \in S^{(1)}$ and $a^{(1)} \in A^{(1)}$ **do**
    Initialize the Q-value $Q^{(1)}_{[t_0,c]}\left(s^{(1)}_{i,t_0}, a^{(1)}_{l,t_0}\right)$
  end for
top:
  At the next pattern update time epoch t:
    Generate a random number $\varrho^{(1)} \in \mathcal{U}\,(0,\,1)$
    if $\left(\varrho^{(1)} \leq \epsilon^{(1)}\right)$, **then**
      Apply a random action $a^{(1)}_t \in A^{(1)}$
    else
      Apply the action $a^{(1)}_t \in A^{(1)}$, *accord. to eq. (18)*
    end if
    Observe DL and UL traffic statistics $\overline{\mu}_{[t,c]}$
    Get current cost $\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$, *accord. to eq. (17)*
    Determine system next state $s^{(1)}_{t+1}$, *accord. to eq. (14)*
    Update Q-value $Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$, *accord. to eq. (19)*
    Move time indexer: $t = t{+}1$, $s^{(1)}_t = s^{(1)}_{t+1}$, ***goto*** top.

paper, the adopted method is to separately filter the DL and UL latency samples using a Kaiser filter. The motivation for using such a filter is its flexibility to provide higher filter weights (priorities) to the most critical input latency samples. This fits in achieving the stringent URLLC performance, where the achievable overall URLLC outage latency target is typically dictated by the worst latency samples. Therefore, a single scalar latency indication for the cell is calculated to reflect the overall DL and UL latency performance of each cell, guiding the learning process of Algorithm-2.

The inter-UE DL/UL HoLD samples are filtered using a non-uniform spatial window. Precisely, the filtering is applied on the HoLD statistics in order to: (1) prioritize the delay samples of the UEs with the largest HoLD by assigning delay-proportional weights, and (2) safeguard Algorithm-2 learning convergence against the sudden changes of the per-packet HoLD samples. As shown by Fig. 4, we apply a mirrored Kaiser window $\Lambda\,[\vartheta\,]$ over the inter-UE HoLD statistics [7], where $\Lambda\,[\vartheta\,]$ is expressed in the digital domain

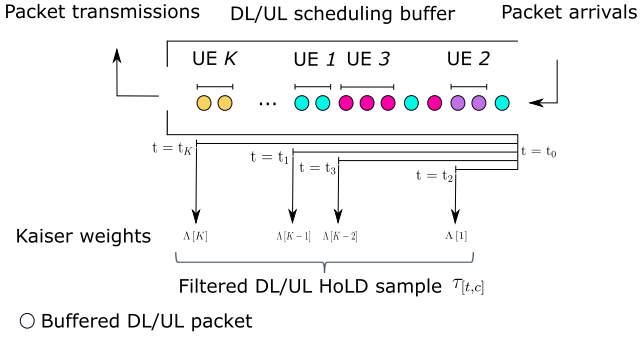


FIGURE 4. URLLC outage latency in DL/UL direction (ms).

by

$$\Lambda\,[\vartheta\,] = \frac{I_0\left[\beta\sqrt{1-\left(\frac{2\vartheta}{\theta}-1\right)^2}\right]}{I_0\,[\beta]}, \quad 1 \leq \vartheta \leq \theta{+}1, \tag{20}$$

where I_0 implies the zero-order modified Bessel function, β is a shaping factor, and $\theta{+}1$ denotes the window length, where $\Lambda\,[K] > \Lambda\,[K{-}1] > \ldots > \Lambda\,[1]$. Accordingly, the HoLD ratio $\tau_{[t,c]}\,(\varsigma)$ of the ς^{th} slot is defined as

$$\tau_{[t,c]}\,(\varsigma) = \frac{\tau^{\mathrm{dl}}_{[t,c]}\,(\varsigma)}{\tau^{\mathrm{dl}}_{[t,c]}\,(\varsigma){+}\tau^{\mathrm{ul}}_{[t,c]}\,(\varsigma)}, \tag{21}$$

where $\tau^{\mathrm{dl}}_{[t,c]}\,(\varsigma)$ and $\tau^{\mathrm{ul}}_{[t,c]}\,(\varsigma)$ are the Kaiser-filtered cell-specific HoLD samples in the DL and UL directions, respectively. Then, the average HoLD $\overline{\tau}_{[t,c]}$ across the radio pattern is then calculated by

$$\overline{\tau}_{[t,c]} = \frac{1}{\xi}\sum_{\varsigma=1}^{\xi}\tau_{[t,c]}\,(\varsigma)\,. \tag{22}$$

Equivalently to $\overline{\mu}_{[t,c]}$ of Algorithm-1, $\overline{\tau}_{[t,c]} \rightarrow [0,\,1]$ captures the directional HoLD performance. For instance, $\overline{\tau}_{[t,c]} = 0.8$ denotes that the DL HoLD is 4x times the corresponding UL HoLD. The state space of Algorithm-2 sub-networks is accordingly defined as

$$S^{(2,l)}_t = \left\{s^{(2,l)}_{1,t},\ s^{(2,l)}_{2,t},\ \ldots,\ s^{(2,l)}_{\mathfrak{I}_{2,l},t}\right\}, \tag{23}$$

where $\mathfrak{I}_{2,l}$ is the state space size of $Q{-}2{-}l$. Then, the corresponding HoLD-to-state mapping is defined as

$$s^{(2,l)}_t = \begin{cases} s^{(2,l)}_{1,t}, & \overline{\tau}_{[t,c]} < \tau_{\min} \\ s^{(2,l)}_{2,t}, & \tau_{\min} \leq \overline{\tau}_{[t,c]} < \tau_{\min}{+}\Upsilon_l \\ s^{(2,l)}_{3,t}, & \tau_{\min}{+}\Upsilon_l \leq \overline{\tau}_{[t,c]} < \tau_{\min}{+}2\Upsilon_l \\ \vdots & \vdots \\ s^{(2,l)}_{\mathfrak{I}_{2,l},t}, & \overline{\tau}_{[t,c]} \geq \tau_{\max}, \end{cases} \tag{24}$$

with the HoLD quantization step Υ given by

$$\Upsilon_l = \frac{\tau_{\max}{-}\tau_{\min}}{\mathfrak{I}_{2,l}{-}2}, \tag{25}$$

where $\tau_{\max}$ and $\tau_{\min}$ are the pre-defined maximum and minimum allowable bounds of the HoLD ratio $\overline{\tau}_{[t,c]}$. The intermediate $s^{(2,l)}_{i,t}, \forall i \in \mathfrak{I}_{2,l}$ states are the favorable state set of Algorithm-2 RML sub-networks, in order to preserve the minimum possible; though, balanced DL and UL HoLD performance.

The action space $A^{(2,l)}$ is built to present all the possible DL and UL symbol structures of the l^{th} pattern sub-book as

$$A^{(2,l)}_t = \left\{a^{(2,l)}_{1,t},\ a^{(2,l)}_{2,t},\ \ldots,\ a^{(2,l)}_{\mathbf{Card}\left(A^{(2,l)}\right),t}\right\}, \tag{26}$$

where $a^{(2,l)}_{j,t} \equiv \hat{w}_j, \forall j \in \mathbf{Card}\left(A^{(2,l)}\right)$, with $A^{(2,l)}$ as the set of all radio structures in the l^{th} sub-book. Accordingly,

the immediate environment return $\Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right), \forall i \in \mathfrak{J}_{2,l}$, $j \in$ **Card** $\left(A^{(2,l)}\right)$ is defined by how much average HoLD $\overline{\tau}_{[t,c]}$ deviation is observed from its balanced mean $\mathfrak{m}^{(2,l)}$ as follows:

$$\Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right) = \left|\overline{\tau}_{[t,c]} - \mathfrak{m}^{(2,l)}\right|, \tag{27}$$

where the mean value of the HoLD ratio $\mathfrak{m}^{(2,l)}$ is adopted as the optimization target of the secondary Q-RML sub-networks, as it ensures a balanced DL and UL HoLD performance. Then, the secondary RML instances adopt the action, i.e, symbol structure $\hat{w}_j$, which offers the minimum variance of the relative HoLD performance as given by

$$\left(a^{(2,l)}_{j,t}\right)^* = \underset{a_j \in A^{(2,l)}}{\arg\min}\ \Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right). \tag{28}$$

Similarly to eq. (19), the value function entries $Q^{(2,l)}_{[t,c]}$ of Algorithm-2 are iteratively updated to reflect the learning experiences, as expressed by

$$Q^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right) \leftarrow \left(1-\alpha^{(2)}\right)\ Q^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right) + \alpha^{(2)} \left[\Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right) + \gamma^{(2)} \underset{a_j \in A^{(2,l)}}{\arg\min} Q^{(2,l)}_{[t+1,c]}\left(s^{(2,l)}_{i,t+1}, a^{(2,l)}_{j}\right)\right], \tag{29}$$

where $\alpha^{(2)}$ and $\gamma^{(2)}$ are the learning rate and discount factor of the secondary sub-network, respectively. The detailed steps of secondary RML instance is described by Algorithm 2.

Algorithm 2 Algorithm-2 for Outage Latency Minimization

Initialize:
for each $s^{(2,l)} \in S^{(2,l)}$ and $a^{(2,l)} \in A^{(2,l)}$ **do**
  Initialize the Q-value $Q^{(2,l)}_{[t_0,c]}\left(s^{(2,l)}_{i,t_0}, a^{(2,l)}_{j,t_0}\right)$
end for
top:
At the next pattern update time epoch t:
  Activate the $Q-2-l$, to selected $\frac{d^l}{u^l}$ from $Q-1$
  Generate a random number $\varrho^{(2,l)} \in \mathcal{U}\ (0, 1)$
  if $\left(\varrho^{(2,l)} \leq \epsilon^{(2,l)}\right)$, **then**
    Apply a random action $a^{(2,l)}_{j,t} \in A^{(2,l)}$
  else
    Apply the action $a^{(2,l)}_{j,t} \in A^{(2,l)}$, *accord. to eq. (28)*
  end if
  Observe DL and UL HoLD statistics $\overline{\tau}_{[t,c]}$
  Get the cost $\Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right)$, *accord. to eq. (27)*
  Determine system next state $s^{(2,l)}_{t+1}$, *accord. to eq. (24)*
  Update Q-value $Q^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right)$, *accord. to eq. (29)*
  Move time indexer: $t = t+1$, $s^{(2,l)}_{t} = s^{(2,l)}_{t+1}$, ***goto*** top.

VI. STATE-OF-THE-ART DUPLEXING SCHEMES

We compare the performance of the proposed solution against the most widely adopted duplexing schemes, for different directional traffic offered loads. The proposed solution is evaluated under two main variants, i.e., when either Algorithm-1 learning is solely adopted or both Algorithm-1 and Algorithm-2 are activated. For the former case, a default DL/UL evenly-distributed pattern structure is employed, following the estimated $\frac{d}{u}$ from Algorithm-1. The duplexing deployments under investigation are as follows:

Frequency division duplexing (FDD): for a comprehensive URLLC latency analysis, FDD is considered as the reference case. The FDD DL and UL bandwidth allocations are configured equivalently to the TDD cases, such that the total bandwidth is fixed. That is, the bandwidth allocation for each of the UL and DL direction is half of the TDD bandwidth.

Dynamic TDD (dTDD) [5]: neighboring BSs independently and dynamically in time select the radio patterns which better satisfy their link selection criteria. Herein, for the sake of cross-scheme fairness, we adopt the same buffered traffic criterion of Algorithm-1 as per eq. (12). The structure of the selected radio pattern, in terms of the placement of the DL and UL symbols, is presumed to be always evenly distributed, and with a symbol block size of 4 symbols. For example, a 14-symbol slot with $\frac{d}{u} = \frac{2}{1}$ is configured as [DDDDFUUUUDDDDF]. Such strategy allows for distributed DL and UL transmission opportunities across the pattern duration. Herein, no inter-BS coordination is assumed, hence, BS-BS and UE-UE CLI can be inflicted.

Static TDD (sTDD): a pre-defined global radio pattern is configured for all neighboring BSs, that meets the average traffic demands of the cluster. We assume a perfect knowledge of the average offered traffic ratio $\frac{\Omega^{\text{dl}}}{\Omega^{\text{ul}}}$, thus, configuring the global radio pattern with a perfect-matching $\frac{d}{u}$. Although sTDD requires the simplest implementation complexity, without CLI infliction, it offers no pattern adaptation to the BS-specific varying traffic and latency demands.

Semi-static TDD (Semi-sTDD) [7]: it is built on top of the sTDD scheme in order to offer an extended TDD adaptation flexibility. Basically, Semi-sTDD follows the same setup as the sTDD scheme; however, the common radio pattern is periodically updated to meet the varying cross-BS traffic demands, and accordingly re-used by all coordinated BSs. In that regard, neighboring BSs continuously exchange indications to their respective traffic needs over the *Xn-interface*.

VII. PERFORMANCE EVALUATION

A. SIMULATION METHODOLOGY

We evaluate the performance of the proposed solution using extensive dynamic system level simulations, where the main modeling assumptions are listed in Table 1. The simulations follow the system model described in Section II, and are in line with agreed 3GPP system level simulation methodology. The simulated scenario is the Urban Macro (UMa) with three sector base station sites placed in a regular hexagonal grid and UEs randomly positioned, following a spatial uniform distribution. Time-variant dynamic traffic is simulated for each UE as per the description in Section II.A. Each UE is

TABLE 1. Simulation setup and major parameters.

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance, 3D SCM channel [32]
Carrier configuration	10 MHz carrier bandwidth at 3.5 GHz; synchronous TDD
PHY numerology	30 kHz subcarrier spacing; 12 subcarriers per PRB; TTI of 4-OFDM symbol duration
Max transmit power	BS: 43 dBm; UE: 23 dBm
Control channel	Error-free control signalling with dynamic link adaptation
DL data channel	QPSK to 64QAM modulation
UL data channel	QPSK modulation
Channel state information	Channel quality indication and precoding matrix indication, reported every 5 ms; Sub-band size: 8 PRBs
Antenna configuration	$N = 8$, $M = 2$; LMMSE-IRC DL/UL receiver
UL power control	Open-loop, $\alpha = 1$, $P0 = -96$ dBm, variable power offset for low and high pathloss UEs, $\hat{c}$ = -110 dB
Packet scheduler	DL: proportional fair, UL: configured grant
BLER target	DL: 1 percent with dynamic MCS selection, UL: pre-defined MCS configuration [31]
HARQ	Asynchronous HARQ with Chase combining; Maximum number of re-transmissions: 6, UL HARQ power boost: +3 dB
Average UE load	$K^{\text{dl}} = K^{\text{ul}} = 40$, and 118; Uniformly distributed
Traffic composition	FTP3, $f^{dl} = f^{ul}$ = 256 bits; λ^{dl} = 50 packets/sec; λ^{ul} = 50 packets/sec
Offered load $\Omega^{\text{dl}}/\Omega^{\text{ul}}$	1 : 1 [0.5 : 0.5 Mbps]; 1 : 2 [1 : 2 Mbps]; 2 : 1 [2 : 1 Mbps]
Processing time	PDSCH preperation: 2.5-OFDM symbols PDSCH decoding: 4.5-OFDM symbols PUSCH preperation: 5.5-OFDM symbols PUSCH decoding: 3-OFDM symbols
TDD pattern periodicity	10 ms
Default symbol structure	DL/UL evenly distributed in blocks of four symbols with same link direction
Proposed solution setup	$L = 9$ $\mathfrak{J}_1 = \mathfrak{J}_2 = 3$ $\mu_{\text{max}} = 0.75$ $\mu_{\text{min}} = \tau_{\text{min}} = 0.2$ $\mathfrak{y}^{(1)} = \mathfrak{y}^{(2)} = \alpha^{(1)} = \gamma^{(1)} = 0.5$ $\epsilon^{(1)} = \epsilon^{(2)} = 0.25$ $\tau_{\text{max}} = 0.8$ **Card** $A_t^{(2)} = \{11, 19, 27, 34, 61, 48, 31, 10, 7\}$ $\alpha^{(2)} = \gamma^{(2)} = 0.7$

served by the cell corresponding to the highest received reference signal received power. The advanced three-dimensional 3GPP UMa radio propagation model is assumed [37]. The simulator includes explicit modeling of all the major MAC and PHY layer functionalities, and related RRM functionalities. For each transmission, the per subcarrier symbol SINR is calculated. Such SINR calculations assume LMMSE-IRC and include both the effect of the co-channel and potential CLI into account in line with the SINR calculations in (4) and (5). Based on all the subcarrier symbols SINR for the transmission, the combined mean mutual information per coded bit (MMIB) mapping [38] is applied for calculation of the effective SINR level. The respective transmission packet error probability (PEP) is calculated based on look-up tables, obtained from extensive link level simulations. Based on the calculated PEP, the corresponding packet is determined as either successful or failed. During the DL TTIs, DL UEs are dynamically scheduled based on the proportional fair criterion, assuming also dynamic link adaptation with adaptive selection of the MCS based on the most recent received CQI reports, including also outer loop link adaptation. UL UEs are served using the CG baseline as outlined in Section II.A. HARQ re-transmissions are always prioritized over new packet transmissions. For each frame periodicity (10 ms), the proposed learning framework in Section V runs in a distributed manner for each cell to determine the next radio pattern configuration.

The simulator is validated via so-called calibration exercises, where baseline statistics are reported and compared between 3GPP partners [39]. Simulations are run for a sufficiently long-time period to ensure statistical reliable results, and thereby a solid basis for drawing mature conclusions. In line with [30], the default simulation length is 5 million successfully decoded URLLC payloads. Thus, assuming that the URLLC packets are fully uncorrelated, the target 99.999% percentile of the URLLC latency distribution is calculated with a maximum error margin of $\pm 5\%$, and therefore, with a 95% statistical confidence level [40].

Due to the nature of the simulations where the UEs are created at the start, traffic is dynamic (i.e. payloads are generated according to Poisson point processes), and the various control loops (e.g. for link adaptation, TDD frame adaptation, etc.), we apply a so-called warm-up time. Only after the warm-up time, the performance statistics are collected from

the simulations. By default, the warm-up time is configured to equal 1 second as this is found to be enough time for the network performance to stabilize.

B. BASELINE PERFORMANCE COMPARISON

Fig. 5 depicts the complementary cumulative distribution function (CCDF) of the combined DL and UL achievable latency in ms, under the proposed scheme, FDD, dTDD, and the sTDD deployments. Clearly, the FDD scheme always outperforms the dTDD scheme. This is mainly attributed to the absence of the CLI as well as the concurrent availability of the DL and UL transmission opportunities. The sTDD is configured with the assumption of the optimal knowledge of the directional offered load. Hence, it is configured with a perfect-matching pattern configuration, i.e., $\Omega^{\rm dl}/\Omega^{\rm ul} = 1 \rightarrow d/u = 1$. Looking particularly at the outage URLLC latency at the 10^{-5} probability, the proposed Algorithm-2 clearly offers a significant outage latency improvement. That is, 70% and 53% outage latency reduction compared to dTDD and sTDD, respectively. Although, it inflicts $\sim$ 51% outage latency increase compared to the FDD case. The performance merits of the proposed solution are mainly due to the sufficient learning gain to compensate for the directional HoLD in designing the radio pattern configuration. The sTDD, with the optimal knowledge of $/\Omega^{\rm dl}/\Omega^{\rm ul}$, offers a slight latency enhancement than the proposed Algorithm-1, due to the non-existent CLI. Though, it exhibits a clear performance loss compared to the proposed Algorithm-2, as the latter introduces an additional latency-aware RML layer.

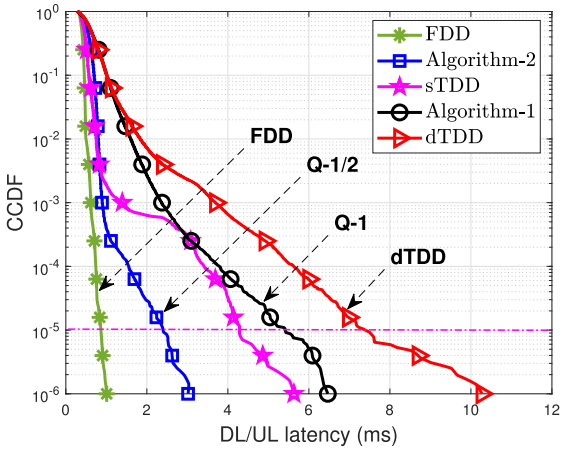


FIGURE 5. Achievable latency, with $\Omega = 1$ Mbps, and $\Omega^{\rm dl}/\Omega^{\rm ul} = 1$.

Fig. 6 shows the empirical CDF (ECDF) of the traffic ratio $\overline{\mu}$ for all schemes under evaluation. Clearly, the larger the $\overline{\mu}$, the larger the buffered DL traffic compared to that is of the UL direction. The dTDD scheme obviously inflicts the lowest $\overline{\mu}$, with $\overline{\mu} = 0.15$ at the 50%ile, indicating that the UL traffic is consistently blocked by the BS-BS CLI, i.e., the buffered UL traffic is 5.6x times the corresponding DL traffic, despite the configured $\Omega^{\rm dl}/\Omega^{\rm ul} = 1$. The sTDD provides a marginally improved UL buffering performance, compared to dTDD, due to the CLI-free UL. However, it does

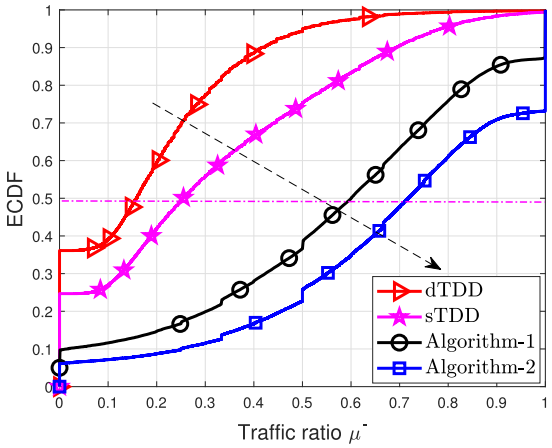


FIGURE 6. Buffering performance, with $\Omega = 1$ Mbps, and $\Omega^{\rm dl}/\Omega^{\rm ul} = 1$.

not account for the DL and UL traffic variations. The proposed Algorithm-1 and Algorithm-2 solutions offer a smooth traffic buffering performance, clearly without the UL traffic accumulation problem, i.e., $\overline{\mu} = 0.57$ and 0.71, respectively. This denotes the buffered UL traffic size is 0.75x and 0.48x times the buffered DL traffic, respectively. Accordingly, the proposed learning solution dynamically compensates for the degraded UL capacity by assigning more UL transmission opportunities across the radio pattern, leading to a faster UL traffic recovery. Though, this comes at the expense of an additional DL traffic buffering, i.e., 25% more buffered DL traffic with Algorithm-2.

Fig. 7 shows the CCDF of the achievable URLLC latency under the proposed algorithm and the Semi-sTDD scheme, respectively, for both light and high offered load cases. With $\Omega = 0.25$ Mbps, the proposed learning algorithm clearly achieves a significant enhancement of the DL/UL URLLC outage latency, offering 1.06 ms at the 10^{-5} probability, with 60% outage latency reduction, compared to Semi-sTDD. For such a lightly-loaded case, the URLLC outage latency is dominated by the structure of the DL and UL symbols across the radio pattern, rather than the CLI intensity. The proposed learning solution autonomously optimizes the pattern structure to provide a faster and BS-specific DL and UL link switching design. Though, the Semi-sTDD inflicts a

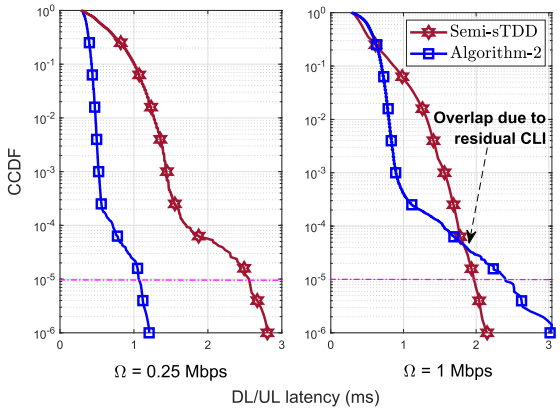


FIGURE 7. Latency comparison to Semi-sTDD, with $\Omega^{\rm dl}/\Omega^{\rm ul} = 1$.

clear URLLC outage degradation due to the high DL and UL traffic fluctuations across neighboring BSs, thus, adopting a common radio pattern offers limited TDD adaptability. For the high-load region $\Omega = 1$ Mbps, the CLI becomes vital to control because of the increased DL traffic size, and thus, the critical BS-BS CLI. The proposed solution therefore exhibits limited degrees of freedom in designing the sufficient DL and UL switching structure, in order to control the severe CLI accordingly. The Semi-sTDD scheme offers 21% latency reduction, compared the proposed solution, mainly due to the absence of the CLI. This case, unlike the lightly-loaded setup, the cross-BS traffic statistics converge to the same average, hence, the Semi-sTDD with a global radio pattern becomes more efficient to achieve a decent URLLC outage latency.

C. Q-RML CONVERGENCE PERFORMANCE

Achieving a robust convergence performance of the RL-based solutions is demonstrated to be a challenging task [25] mainly due to the sparse reward function observed from surrounding environment. Furthermore, since the system-model adopted in this work incorporates time-variant channel conditions with sporadic and UE-specific packet arrivals, analyzing the convergence performance of the proposed learning approach becomes vital. We performed a large set of the system level simulations with various warm-up periods in order to obtain the best possible RL settings which offer the best achievable URLLC outage latency. As described in Section VII.A, the warm-up duration implies the starting period of the simulation until the system gets loaded. We also utilize such time as the convergence delay of the proposed QRL framework where the action exploration is prioritized to stabilize all corresponding Q-value functions during the warm-up. That is, we adopted warm-up periods from 0.25 to 1.5 second alongside with adopting different action exploration-exploitation probabilities from 0 to 0.7 for both Algorithm 1 and 2, respectively. Therefore, based on our extensive sensitivity analysis, we adopt $\sim$1 second of warm-up time over which the action exploration probability for both the primary and secondary learning instances is set to $\epsilon^{(1)} = \epsilon^{(2)} = 0.25$. During the actual simulation time, i.e,. QRL inference time, the actions which offer the lowest possible cost functions are always utilized, i.e., $\epsilon^{(1)} = \epsilon^{(2)} = 0.0$ during inference (no action exploration). This setting offers the shortest convergence delay and accordingly, the best achievable URLLC outage performance for the considered system configurations.

To monitor the actual convergence performance of the proposed QRL framework, we calculate the learning temporal difference (TD). The TD reflects how well the Q-learning is converging towards the optimal policy in time. In particular, it captures the difference among the current learning samples and the former learning experiences as

$$\begin{aligned}\mathbf{TD}_{Q_1} &= \Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)+\gamma^{(1)}\underset{a_l\in A^{(1)}}{\arg\min}\; Q^{(1)}_{[t+1,c]}\left(s^{(1)}_{i,t+1}, a^{(1)}_{l}\right)\\ &\quad -Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right). \end{aligned} \tag{30}$$

$$\begin{aligned}\mathbf{TD}_{Q_{2,l}} &= \Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right)+\gamma^{(2)}\\ &\quad \underset{a_j\in A^{(2,l)}}{\arg\min}\; Q^{(2,l)}_{[t+1,c]}\left(s^{(2,l)}_{i,t+1}, a^{(2,l)}_{j}\right)-Q^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right). \end{aligned} \tag{31}$$

As depicted by Fig. 8, the TD distribution of both Algorithm-1 and Algorithm-2 is quite compressed, where Algorithm-2 tends to experience a faster learning convergence than Algorithm-1, due to the already refined learning of the symbol ratio d/u. Upon convergence, the new learning observations do not significantly change the applied actions, leading to a slower transition rate over the *state-action* pairs. That is, at the 50%-ile of the TD distribution, the secondary learning exhibits a normalized TD of 0.08. This denotes that, upon convergence, the cost values of the proposed Algorithm-2 are fluctuating in time by only $\pm 8\%$, due to the sufficient learning of the pattern structure. Such convergence performance is obtained with the baseline system setting as indicated by Table 1. That is, an offered traffic load of 1 Mbps/cell and equal DL and UL traffic load split, where the action exploration probabilities are set as: $\epsilon^{(1)} = \epsilon^{(2)} = 0.25$ during the warm-up time.

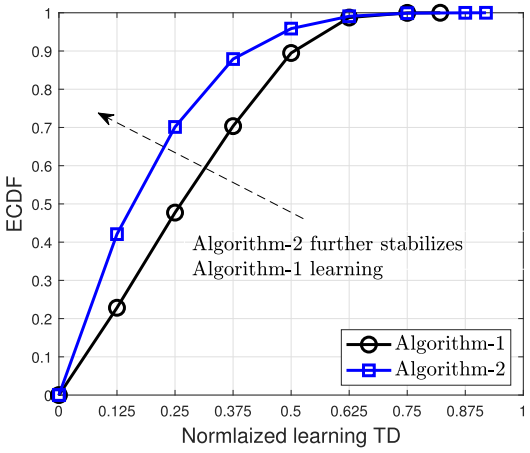


FIGURE 8. TD performance, with $\Omega = 1$ Mbps, and $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1$.

In particular, the modeling of the learning objectives, i.e., learning targets, learning inputs and outputs are shown to significantly impact the achievable convergence performance. As the main learning objective of the primary Q-RML is the aggregated buffered traffic, it imposes partial stationarity due to the several active users at the same time. That is, an abrupt change of the aggregate buffered traffic is not highly likely. For the secondary Q-RML networks, the Kaiser-window filtered delay statistics of the buffered users are considered instead of the actual latency values, as the latter could potentially rapidly change, disturbing the learning convergence. Thus, the convergence of the proposed approach has a quick time cycle. Furthermore, as the learnable action set are the set of all possible TDD radio frame configurations, the complexity of the proposed solution scales mainly with the number of possible TDD radio patterns.

That is typically limited by couple of hundreds, allowing for a further quicker convergence delay, i.e., the complexity for calculating and updating the Q-values of each possible action (TDD pattern).

D. CROSS LINK INTERFERENCE PERFORMANCE

As a consequence to the achievable radio frame learning potential, the proposed solution tends to realize an autonomous trade-off between the DL and UL symbol switching periodicity versus the subsequent CLI performance. In particular, a faster DL and UL switching periodicity during the radio pattern is favored; though, it is likely to result in frequent CLI occurrences, due to the higher probability of adjacent BSs adopting opposite link directions. Accordingly, the latency merits, obtained from the fast link switching, are completely wiped out, and reverted into an outage latency loss due to the severe CLI. Therefore, as shown by Fig. 9, the proposed solution clearly offers a substantial reduction of the BS-BS CLI, compared to the dTDD scheme.

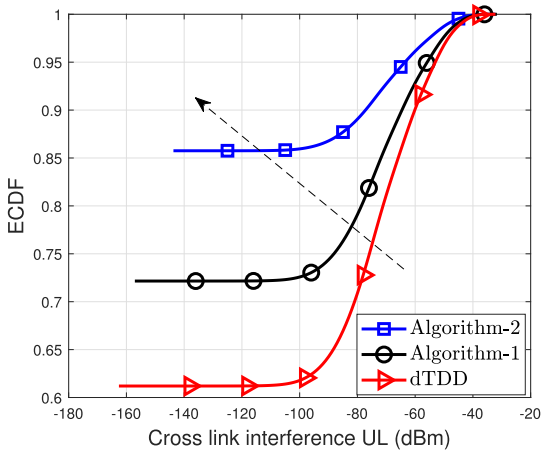


FIGURE 9. BS-BS CLI performance, with $\Omega = 1$ Mbps, and $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 1$.

E. PERFORMANCE EVALUATION WITH DIFFERENT OFFERED TRAFFIC RATIOS

Examining the proposed solution under different offered load ratios, Fig. 10 presents the achievable latency performance with $\Omega = 3$ Mbps, and $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 1/2$ and $2/1$, respectively. Particularly, with $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 2/1$, the URLLC outage latency becomes dictated by the severe CLI, and especially the BS-BS CLI, due to the larger DL traffic portion. The proposed solution dynamically compensates for the highly-degraded UL PRB capacity by allocating more UL transmission opportunities, leading to 67% outage UL latency reduction, compared to dTDD. However, it comes at the expense of further increased DL traffic buffering, i.e., 49% outage DL latency increase. With $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 1/2$, where the BS-BS CLI is negligible, proposed solution achieves a reliable outage latency improvement for both link directions.

To explore how the schemes under evaluation re-act to the directional traffic variations, we define the symbol ratio

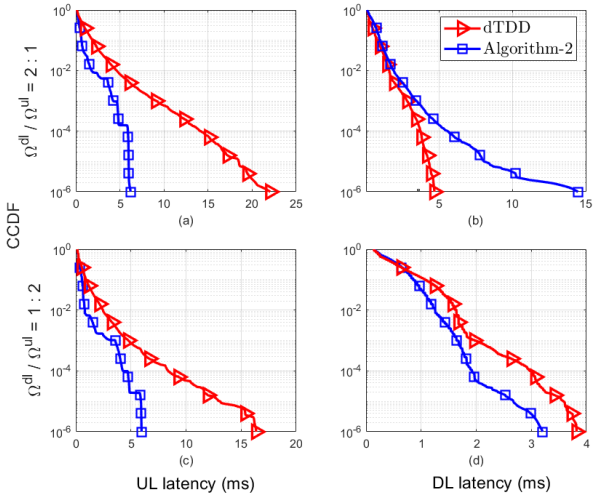


FIGURE 10. Latency performance (ms), with $\Omega = 3$ Mbps, and $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}}$.

$\eta^c \rightarrow [0, 1]$ as given by

$$\eta^c = \frac{d^c}{d^c + u^c}. \tag{32}$$

Accordingly, Fig. 11 shows the average symbol ratio η^c of the proposed solution, sTDD, dTDD, and Semi-sTDD schemes, respectively, for different $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}}$ ratios. Clearly, the sTDD scheme always adopts a linear mapping from $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}}$ to d^c/u^c due to the fixed pattern configuration. That is, $\eta^c = 0.33$, 0.5, and 0.66 for $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = d^c/u^c = 1/2$, $1/1$, and $2/1$, respectively. The Semi-sTDD scheme follows the sTDD in terms of the dynamically configured average symbol ratio η^c; however, with moderate variations due to the additional TDD pattern adaptation gain, e.g., adopting +12% UL symbols on average than the sTDD scheme with $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 2/1$. The dTDD scheme performs quite efficiently under light CLI intensity. That is, with $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 1/2$, an almost-balanced DL and UL adaptation is achieved, where an average $\eta^c = 0.29$ is observed. It implies that the $u^c = 2.4\,d^c$ symbol configuration is favored by the dTDD pattern adaptation process, to allow for the degraded UL capacity due to the residual CLI. However, the dTDD scheme obviously inflicts an UL capacity blocking under high CLI intensity conditions, i.e., $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 2/1$, where $\eta^c = 0.34$

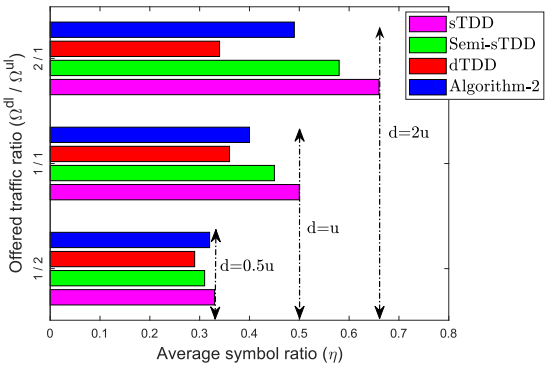


FIGURE 11. Symbol configuration, with $\Omega = 1$ Mbps, and $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}}$.

is exhibited. That denotes the $u^c = 1.9\,d^c$ configuration is adopted on average, and subsequently, the DL capacity inflicts a starvation of the transmission opportunities across the configured radio patterns.

Moreover, Fig. 11 shows that the proposed solution preserves a balanced symbol configuration performance under all considered directional load cases. Unlike the sTDD and Semi-sTDD schemes, proposed learning solution tends to bias the pattern configuration towards even more UL transmission opportunities to compensate for the severe BS-BS CLI. Although, unlike the dTDD solution, proposed solution does not exhibit the UL capacity blocking issue, even under severe CLI conditions, i.e., $\eta^c = 0.49$ with $\Omega^{\text{dl}}/\Omega^{\text{ul}} = \frac{2}{1}$. This is mainly attributed to the well-learned trade-off among the residual CLI and the link switching periodicity.

Finally, Fig. 12 depicts the achievable per-TTI UL throughput performance in Mbps of the proposed solution and dTDD case, respectively. The proposed solution achieves a considerable capacity improvement due to the faster traffic transmissions. Obviously, the major capacity gain of the proposed is realized at the lower percentiles, i.e., BS-edge UL UEs, since those are the most impacted by the obtained CLI enhancement and the faster UL transmissions accordingly.

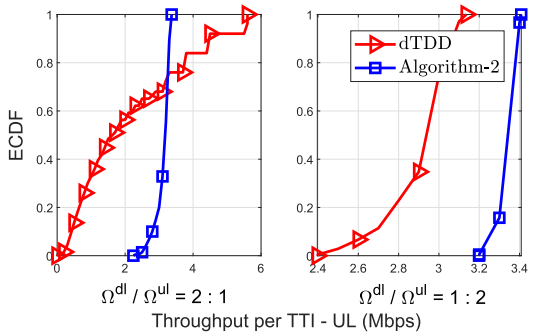


FIGURE 12. Throughput performance, with $\Omega = 3$ Mbps.

VIII. CONCLUDING REMARKS

In this paper, a radio pattern optimization scheme has been proposed for 5G new radio TDD systems. The proposed solution encompasses dual reinforcement Q-reinforcement-learning (QRL) instances for online optimization of the achievable URLLC outage latency, tackling a *min-max* URLLC problem. The primary QRL-network seeks to estimate the number of the DL and UL symbols across the next radio pattern, which best satisfies a faster; but, balanced downlink and uplink traffic handling. The secondary QRL-sub-networks select the corresponding pattern structure to achieve a decent URLLC outage latency accordingly.

Through extensive system-level simulations, the proposed solution demonstrates a significant URLLC outage latency improvement compared to state-of-the-art dynamic TDD proposals. As an example, the URLLC outage latency is reduced by 70% and 53% compared to the fully dynamic and static TDD solutions, respectively, when assessed at high offered loads. The proposed solution achieves URLLC outage latency of 1 ms at the modest offered load of 250 kbps, while the semi-static TDD solution with inter-cell coordination achieves 2.7 ms latency, i.e. a latency reduction of 60%. Such impressive gain is achieved while the proposed ML solution runs independently for each cell. The semi-static TDD solution utilizes explicit inter-cell coordination. However, at high offered load, where the outage latency is in orders of magnitude higher than the 1 ms URLLC target, the semi-static TDD with explicit inter-cell coordination to avoid any CLI displays as good performance as the proposed solution.

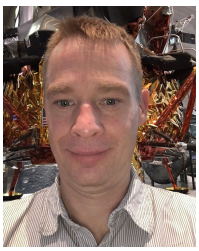
The main insights brought by this paper are as follows: (1) URLLC latency and reliability performance is highly challenged in dynamic TDD deployments, due to the non-concurrent downlink and uplink transmission opportunities, and the additional cross-link interference (CLI), (2) thus, the real-time optimization of the radio pattern structure becomes vital towards a decent URLLC outage performance, (3) accordingly, machine learning techniques can be efficiently utilized to offer a proactive pattern estimation learning gain, (4) in this regard, reinforcement Q-learning has been adopted due to its online (real-time) learning capabilities, and simple implementation complexity under the adopted system model, and (5) proposed solution demonstrates a flexible and dynamic radio pattern selection strategy to autonomously trade-off the CLI intensity with the URLLC outage performance; however, the achievable gain is shown to be load-dependent. As a future extension of this study, various learning approaches such as the state-action-reward-state-action (SARSA) shall be considered in order to learn and further optimize the selection of TDD radio patterns. Furthermore, extending the ML-driven solution for TDD pattern optimization to include explicit inter-cell coordination may offer further performance benefits; including also faster learning convergence and robustness.

REFERENCES

[1] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.
[2] *Service requirements for 5G Systerm*, document TS 22.261, V16.6.0, 3GPP, Dec. 2018.
[3] J. Lee, "Spectrum for 5G: Global status, challenges, and enabling techs," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12–18, Mar. 2018.
[4] K. I. Pedersen, G. Berardinelli, F. Frederiksen, and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 122–128, Apr. 2017.
[5] A. A. Esswie and K. I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," in *Proc. IEEE 17th Annu. Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2020, pp. 1–6.
[6] *NR; Physics Layer Procedures for Control*, document V16.0.0, TS 38.213, 3GPP, Dec. 2019.
[7] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, "Semi-static radio frame configuration for URLLC deployments in 5G macro TDD networks," in *Proc. IEEE WCNC*, Apr. 2020, pp. 1–5.
[8] A. A. Esswie and K. I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, Kuala Lumpur, Malaysia, Apr. 2019, pp. 1–6.
[9] L. Binyong and C. Gang, "Dynamic TDD DL/UL reconfiguration based on shift," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Oct. 2016, pp. 1561–1568.

[10] Z. Huo, N. Ma, and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," in *Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2017, pp. 962–967.
[11] A. Lukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, Apr. 2016, pp. 1–7.
[12] A. Lukowa and V. Venkatasubramanian, "Coordinated user scheduling in 5G dynamic TDD systems with beamforming," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Bologna, Spain, Sep. 2018, pp. 596–597.
[13] A. A. Esswie and K. I. Pedersen, "Cross-link interference suppression by orthogonal projector for 5G dynamic TDD URLLC systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–7.
[14] E. de Olivindo Cavalcante, G. Fodor, Y. C. B. Silva, and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with BS to BS interference constraints," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 788–791, Oct. 2018.
[15] A. A. Esswie, K. I. Pedersen, and P. E. Mogensen, "Quasi-dynamic frame coordination for ultra-reliability and low-latency in 5G TDD systems," in *Proc. IEEE Int. Conf. Commun. Workshops*, Shanghai, China, May 2019, pp. 1–6.
[16] J. W. Lee, C. G. Kang, and M. J. Rim, "SINR-ordered cross link interference control scheme for dynamic TDD in 5G system," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Chiang, Mai, Jan. 2018, pp. 359–361.
[17] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions," *IEEE Access*, vol. 7, pp. 137184–137206, 2019.
[18] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.
[19] I.-S. Comsa, S. Zhang, M. E. Aydin, P. Kuonen, Y. Lu, R. Trestian, and G. Ghinea, "Towards 5G: A reinforcement learning-based scheduling solution for data traffic management," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 4, pp. 1661–1675, Dec. 2018.
[20] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42–48, Mar. 2019.
[21] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning radio resource management in RANs: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138–145, Sep. 2018.
[22] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.
[23] C. She, R. Dong, Z. Gu, Z. Hou, Y. Li, W. Hardjawana, C. Yang, L. Song, and B. Vucetic, "Deep learning for ultra-reliable and low-latency communications in 6G networks," *IEEE Netw.*, pp. 1–7, Feb. 2020.
[24] P. Louridas and C. Ebert, "Machine learning," *IEEE Softw.*, vol. 33, no. 5, pp. 110–115, May 2016.
[25] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[26] M. Assaouy, O. Zytoune, and D. Aboutajdine, "Policy iteration vs Q-Sarsa approach optimization for embedded system communications with energy harvesting," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2017, pp. 1–6.
[27] *Study on New Radio Access Technology Physical Layer Aspects*, document V14.2.0, TR 38.802, 3GPP, Sep. 2017.
[28] *NR; Physical Channels Modulation*, document V16.1.0, TS 38.211, 3GPP, Mar. 2020.
[29] *NR; Physical Layer Procedures for Data*, document V16.0.0, TS 38.214, 3GPP, Dec. 2019.
[30] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28912–28922, 2018.
[31] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovcs, and P. Mogensen, "Joint resource configuration and MCS selection scheme for uplink grant-free URLLC," in *Proc. Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.
[32] *Study 3D Channel Model for LTE*, document V12.7.0, TR 36.873, 3GPP, Dec. 2014.
[33] Y. Ohwatari, N. Miki, Y. Sagae, and Y. Okumura, "Investigation on interference rejection combining receiver for space frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191–203, Jan. 2014.
[34] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653–133667, 2019.
[35] S. Calisir and M. K. Pehlivanoglu, "Model-free reinforcement learning algorithms: A survey," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.
[36] M. Abu Alsheikh, D. T. Hoang, D. Niyato, H. Tan, and S. Lin, "Markov decision processes with applications in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1239–1267, Oct. 2015.
[37] *Study Channel Model for Frequencies From 0.5 To 100 GHz*, document V16.1.0, TR 38.901, 3GPP, Dec. 2019.
[38] D. G. Popescu, M. Varga, and V. Bota, "Comparison between measured and computed values of the mean mutual information per coded bits in OFDM based wireless transmissions," in *Proc. 36th Int. Conf. Telecommun. Signal Process. (TSP)*, Rome, Italy, Jul. 2013, pp. 380–384.
[39] *Study Physical Layer Enhancements for NR Ultra-Reliable Low Latency Case (URLLC)*, document TR 38.824, 3GPP, Mar. 2019.
[40] L. D. Brown, T. T. Cai, and A. Dasgupta, "Confidence intervals for a binomial proportion and asymptotic expansions," *Annu. Statist.*, vol. 30, no. 1, pp. 160–201, 2002.

ALI A. ESSWIE (Member, IEEE) received the M.Sc. degree in electrical and computer engineering from Memorial University, Canada, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Systems, Aalborg University. From 2013 to 2016, he acted as a Wireless Standards Engineer with Intel Labs and Huawei Technologies. He is also with Nokia Bell Labs, Aalborg. He has authored more than 25 publications and holds ten patent filings. His main research interests include 5G new radio, wireless machine learning/AI, radio resource management, ultra-reliable and low-latency communications, massive MIMO, and channel estimation.



KLAUS I. PEDERSEN (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. He is currently leading the radio Access Systems Research Team, Nokia Bell Labs, Aalborg, and a part-time Professor at the Wireless Communications Network (WCN) Section, Aalborg University. He has authored/coauthored approximately 200 peer-reviewed publications on a wide range of topics, as well as an inventor on several patents. His current research interest includes 5G new radio evolution, including radio resource management aspects to enable new use cases with a special emphasis on mechanisms that offer improved end-to-end (E2E) performance delivery. Recently, he was also a part of the EU funded research project ONE5G that focused on E2E-aware optimizations and advancements for the Network Edge of 5G New Radio that was successfully concluded, in Summer 2019.

PREBEN E. MOGENSEN (Member, IEEE) received the M.Sc. and Ph.D. degrees from Aalborg University, in 1988 and 1996, respectively. He is currently a Principal Scientist at the Nokia Bell Labs Aalborg Department, Denmark, and a Bell Labs Fellow. He is also a Professor at Aalborg University and the Head of the Wireless Communication Networks (WCN) Section. He is currently engaged in research and standardization for vertical use cases for LTE and 5G, including the LPWA IoT, URLLC, I.4.0, V2X, UAV, and train communication. He has published more than 400 articles within wireless communication. He has over 19 000 Google Scholar citations.