# Joint Resource Allocation and Link Adaptation for Ultra-Reliable and Low-Latency Services

Md Arman Hossen*, Thang X. Vu*, Van-Dinh Nguyen‡, Symeon Chatzinotas*, and Björn Ottersten*
* *Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg*
‡ *College of Engineering and Computer Science, VinUniversity, Hanoi 100000, Vietnam*
E-mails: {arman.hossen, thang.vu, symeon.chatzinotas, bjorn.ottersten}@uni.lu, dinh.nv2@vinuni.edu.vn

***Abstract*—With the emergence of ultra-reliable and low latency communication (URLLC) services, link adaptation (LA) plays a pivotal role in improving the robustness and reliability of communication networks via appropriate modulation and coding schemes (MCS). LA-based resource management schemes in both physical and medium access control layers can significantly enhance the system performance in terms of throughput, latency, reliability, and quality of service. Increasing the number of retransmissions will achieve higher reliability and increase transmission latency. In order to balance this trade-off with improved link performance for URLLC services, we study a joint subcarrier and power allocation problem to maximize the achievable sum-rate under an appropriate MCS. The formulated problem is mixed-integer nonconvex programming which is challenging to solve optimally. In addition, a direct application of standard optimization techniques is no longer applicable due to the complication of the effective signal-to-noise ratio (SNR) function. To overcome this challenge, we first relax the binary variables to continuous ones and introduce additional variables to convert the relaxed problem into a more tractable form. By leveraging the successive convex approximation method, we develop a low-complexity iterative algorithm that guarantees to achieve at least a locally optimal solution. Simulation results are provided to show the fast convergence of the proposed iterative algorithm and demonstrate the significant performance improvement in terms of the achievable sum-rate, compared with the conventional LA approach and existing retransmission policy.**

***Index Terms*—Hybrid automatic repeat request (HARQ), link adaptation, modulation and coding scheme (MCS), resource allocation, ultra-reliable and low latency communication (URLLC).**

## I. Introduction

Link adaptation (LA) is a fundamental functionality to deal with deep fading and wireless link failures, providing suggestions for the optimal transmitting parameters. The accurate downlink LA has been a major challenge due to the diverse performance requirements of the three major 5G service categories: enhanced mobile broadband (eMBB) with high data rate and spectral efficiency, massive machine-type communications (mMTC) with access to numerous machine-type devices, and ultra-reliable and low latency communication services (URLLC) with stringent latency and reliability requirements [1]. Among them, URLLC has been considered a key technological enabler to support immersive critical services (e.g., industrial automation, remote surgery, and intelligent transportation, etc.). The stringent requirement for ultra-low latency (nearly or less than 1ms) and ultra-high reliability ($10^{-5} - 10^{-7}$) on URLLC services will certainly pose a great challenge in the system design due to the path-loss, interference, deep fades and blockage, which demands the short packet communication [2]. As per 3GPP Release-15, the transmission of a 32-Byte payload should maintain the ultra-low latency of below 1 ms and $99.999\%$ reliability), whereas 3GPP Release-16 has made further enhancements in latency and reliability bounds [3].

In the downlink, the base station (BS) utilizes the LA scheme to select an appropriate modulation and coding schemes (MCS) based on the channel quality indicator (CQI) feedback information from the user equipment (UE). UE measures the signal-to-noise ratio (SNR) at subcarriers and maps an effective SNR using physical layer abstraction (PLA) techniques, where the feedback CQI has the information of the most suitable MCS for data transmission. LA selects the best MCS whereas Automatic Repeat Request (ARQ) and hybrid automatic repeat request (HARQ) help to alleviate the error occurred in the data transmission which leads to an efficient link quality. In ARQ schemes, the receiver uses an error detection code such as a cyclic redundancy check to send a positive acknowledgment (ACK) or a negative Acknowledgment (NACK) to the transmitter. HARQ mechanism is a combination of ARQ with channel coding, with the objective of reducing the number of transmissions by adding Forward Error Correction (FEC) bits to the existing error detection bits. In particular, the receiver decodes a retransmitted packet with previously transmitted erroneous packets through soft combining methods like Chase Combining (CC) and Incremental Redundancy (IR) [4]. In CC-HARQ, the same packet is being retransmitted to the receiver in case of error, while in IR-HARQ, transmitted packets contain a different mixture of information bits and parity bits whereas packets are being identified by the version of messsages. Every retransmitted packet in CC-HARQ can be considered as extra information in addition to the previously received packet. Given its simplicity and tractability, we will consider the CC-HARQ in this work.

The current literature on the combination of LA and HARQ to maximize the system performance is still sparse and isolated. For example, the estimation of worst-case channel degradation for URLLC services was studied in [5], where BS uses

This work is funded by the Luxembourg National Research Fund (FNR) as part of AFR individual PhD Grant "Design And Optimization Of Intelligent Tactile Internet Systems Over 5g And Beyond Wireless Networks"

the received CQI report to calculate the channel degradation over a time window. Join optimization of LA and HARQ retransmission was analyzed in [6], in which the transmission policy is developed by adding some waiting time to reduce the number of retransmissions. PLA techniques for multi-connectivity (MC) network were investigated in [7], whereas MC-based link adaptation for URLLC services was studied in [8]. In [9], joint link adaptation and dynamic resource allocation for multiplexed EMBB and URLLC services was analyzed along with an attractive CQI measurement procedure. The resource allocation of short packet transmission to enable mission critical services was studied in [10], with the main aim of minimizing the decoding error probability while optimizing the power and blocklength. Recently, the short transmission time interval (TTI) and round trip time (RTT) with restricting the fixed number of HARQ retransmissions and increasing the subcarrier spacing-based LA techniques were studied to substantially enhance the latency and reliability performance for URLLC services [8], [11], [12]. The introduction of network intelligence can bring a substantial impact through minimizing the heuristically tuned parameters and achieving the optimal performance. However, obtaining the optimal LA is challenging, especially for downlink transmission due to limited feedback channels. Notably, an outer loop link adaptation (OLLA) can be deployed to dynamically adjust MCS at BS [13], by using HARQ statistics on ACK/NACK and adding some offsets to SNR estimation. Nevertheless, this method exhibits slow convergence, making it inapplicable to URLLC scenarios.

Most of the works mentioned above-mentioned works aim to optimize rate, while, the main goal of URLLC applications is to precisely control the packet error rate (PER) for every single packet transmission in each coherence under the ultra-reliability constraint. Increasing the number of retransmissions can achieve higher reliability, which also increases the overall transmission latency. To balance this latency reliability trade-off, it is necessary to minimize the number of retransmissions while meeting the reliability requirement. With this in mind, we consider a joint optimization of LA and HARQ retransmissions to maximize the total link throughput, where the number of retransmissions is taken into account depending on the latency budget. In addition, we adopt the short TTI with a finite blocklength to meet the stringent latency-reliability requirement. Our main contributions are summarized as follows:

- We develop an optimization framework to jointly optimize the power allocation and subcarrier assignment in multiuser URLLC systems. Our aim is to maximize the achievable data rate in URLLC services while satisfying stringent latency requirements and balancing latency-reliability trade-offs.
- We optimize the HARQ mechanism via the fixed number of transmissions per packet and the latency-reliability trade-off. An efficient iterative algorithm based on successive convex approximation (SCA) method is developed to solve the formulated optimization problem. We demonstrate the effectiveness of the proposed framework via superior simulation performance compared to the reference baselines.

The rest of this paper is organized as follows. The system model is described in Section II, while the optimization problem is given in Section II-B. We provide the proposed solution in Section III. Section IV is dedicated to numerical results and relevant discussions. We finally conclude the paper in Section V.

## II. System Model and Problem Formulation

### A. System Model

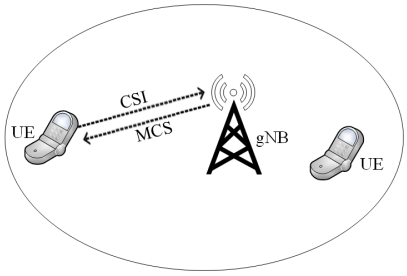


Fig. 1: Simplified network with gNB and UE.

We consider a downlink communication system as shown in Fig. 1, where a BS equipped with a single antenna serves the set $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$ of $K$ UEs via the frequency division multiple access (FDMA). Each user, say UE $k$, has a decoding error rate target of $\epsilon_k$ and a transmission latency target of $d_k$ ms. Under FDMA, the system bandwidth $B$ is divided into the set $\mathcal{N} = \{1, 2, \ldots, N\}$ of $N$ orthogonal subcarriers.

Under FDMA, we assume each subcarrier is occupied by one UE only to eliminate multiuser interference. However, one UE allows to access multiple subcarriers. To establish this relationship, we introduce new binary variables $\mathbf{a} \triangleq \{a_{ki}\}_{\forall k,i}$, satisfying:

$$a_{ki} = \begin{cases} 1, & \text{if UE } k \text{ is assigned to subcarrier } i \in \mathcal{N} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The SNR of UE $k$ at subcarrier $i$ can be written as

$$\gamma_{ki}(a_{ki}, p_{ki}) = \frac{a_{ki} p_{ki} |h_{ki}|^2}{\sigma_k^2} \tag{2}$$

where $p_{ki}$ is the transmit power coefficient that BS allocates to UE $k$ at subcarrier $i$, $h_{ki}$ is the channel between BS to UE $k$ at subcarrier $i$ and $\sigma_k^2$ is the Additive white Gaussian noise (AWGN) varience.

The effective SNR is then computed using the exponential effective SNR mapping (EESM) technique. The basic idea of EESM is to find a compression function to map the set of SNRs to a single value, which results in a good predictor of the actual block error rate (BLER). In this work, we have adjusted the number of HARQ retransmissions to achieve the rigorous reliability requirement and also considered the same HARQ process, MCS and subcarrier during retransmission. Moreover, we consider that SNRs experienced on each subcarrier may vary through retransmissions. For this reason, the effective

SNR of each UE are summed over the retransmissions that is used for the EESM. The computation process of the effective SNR with multiple retransmissions using the EESM-PLA technique has been studied [14]. Accounting for the latency budget, we consider the maximum number of transmissions which can be allowed for a packet is $n$. As we utilize the same MCS and subcarrier for $n$ retransmissions of each packet, the effective SNR after $n$ retransmissions can be computed by [14]

$$\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}) = -\beta \ln\Big(\frac{1}{N_k}\sum_{i\in\mathcal{N}_k}\exp\big(-\frac{n}{\beta}\gamma_{ki}(a_{ki},p_{ki})\big)\Big) \quad (3)$$

where $\mathbf{p} \triangleq \{p_{ki}\}_{\forall k,i}$, $N_k$ is the total pilot signals (or subcarrier) transmitted by BS to UE $k$, $\beta$ is an MCS and channel dependent fitting parameter obtained from real channel realization or simulations. The optimal value of $\beta$ can be found by minimizing the mean square error (MSE) between the AWGN-based SNR and the effective SNR at the same decoding error probability. In addition, $\beta$ needs to be optimized for each modulation scheme and channel profile. Linear curve fitting ($\beta = ak+b$) and exponential curve fitting ($\beta = a + b(1-e^{-ck})$) have been studied in [14] to capture the variation in $\beta$, whereas the authors in [7] proposed a calibration technique based on different channel realizations and noise variance.

In this work, we assume that retransmission is done immediately after finishing the first one. Joint optimization of power and subcarrier allocation along with the controlled number of HARQ retransmissions is the main goal of this paper while maintaining the stringent URLLC service requirements. Let us denote by $T$ and $\epsilon_k$ the blocklength and the packet error rate of UE $k$, respectively. The achievable rate with short packet communication is different from Shannon's capacity, which is expressed as [15]

$$R_k(\mathbf{a},\mathbf{p}) = C_k(\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p})) - \sqrt{\frac{V_k(\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))}{T}}Q^{-1}(\epsilon_k) \quad (4)$$

with

$$C_k(\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p})) = \log_2(1+\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))$$
$$V_k(\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p})) = (\log_2 e)^2\Big(1-\frac{1}{(1+\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))^2}\Big)$$
$$Q(x) = \frac{1}{\sqrt{2}}\int_x^\infty e^{-\frac{u^2}{2}}\,du$$

where $C_k(\gamma_k^{\text{eff}})$, $V_k(\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))$ and $Q^{-1}(x)$ are the traditional Shannon capacity, channel dispersion function and inverse of the Gaussian Q-function, respectively. The primary role of link adaptation is to dynamically adjust the data rate of transmitted signal to equalize the achievable rate of each user. Considering the latency budget $d_k$, the achievable rate of UE $k$ after $n$ transmissions is

$$R_k(\mathbf{a},\mathbf{p}) = \Big\lfloor\frac{d_k}{nT}\Big\rfloor\Big(\log_2\big(1+\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p})\big) - \sqrt{\frac{(\log_2 e)^2}{T}(1-\frac{1}{(1+\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))^2})}Q^{-1}(\epsilon_k)\Big) \quad (5)$$

where $\lfloor . \rfloor$ denotes the floor function due to the fixed latency budget in our system and retransmission process. The complicated form of the channel dispersion function $V_k(\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))$ makes the achievable rate $R_k(\mathbf{a},\mathbf{p})$ in (5) intractable to be analyzed and optimized. To bypass this issue, we assume that the effective SNR of each user is larger than 5 dB, so that $1-\frac{1}{(1+\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}))^2} \approx 1$. This assumption is reasonable to guarantee high-reliable and low-latency communication services. As a result, the achievable rate in (5) can be simplified to

$$R_k(\mathbf{a},\mathbf{p}) = \Big\lfloor\frac{d_k}{nT}\Big\rfloor\Big(\log_2\big(1+\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p})\big) - \sqrt{\frac{1}{0.48T}}Q^{-1}(\epsilon_k)\Big). \quad (6)$$

### *B. Problem Formulation*

The main objective of this work is to maximize the total achievable rate of all UEs by jointly optimizing power and subcarrier allocation, which is mathematically formulated as

$$\max_{\mathbf{a},\mathbf{p}} \quad \sum_{k\in\mathcal{K}} R_k(\mathbf{a},\mathbf{p}) \quad (7a)$$
$$\text{s.t.} \quad a_{ki} = \{0,1\},\ \forall k,i \quad (7b)$$
$$\sum_{k\in\mathcal{K}} a_{ki} \le 1,\ \forall i \quad (7c)$$
$$N_k = \sum_{i\in\mathcal{N}} a_{ki} \le N,\ \forall i \quad (7d)$$
$$\sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}} p_{ki}a_{ki} \le P_{\max},\ \forall k,i \quad (7e)$$
$$\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}) \ge \bar{\gamma},\ \forall k \quad (7f)$$

where $P_{\max}$ and $\bar{\gamma}$ are the total transmit power at the BS and the threshold SNR for all UEs, respectively. Constraint (7b) refers to the binary selection of subcarriers, while (7c) ensures that one subcarrier is occupied by one user at a time. Constraint (7e) is the power constraint at BS with the power budget of $P_{\max}$.

**Challenges of solving problem** (7): We first observe that the objective function (7b) is nonconcave in $(\mathbf{a},\mathbf{p})$, leading to a mixed-integer nonconvex programming due to the binary nature of subcarrier allocation variables (7b). In addition, the product of $a_{ki}$ and $p_{ki}$ in (7e) makes the problem even more challenging to solve directly. Although the existing MILP solvers (e.g. Gurobi or MOSEK) can be used to directly solve a mixed-integer problem, the resulting computation complexity increases exponentially, especially when the number of UEs is large.

## III. Proposed Algorithms

To bypass the product of $\mathbf{a}$ and $\mathbf{p}$ in constraint (7e), we further decompose it into two constraints and rewrite (7) equivalently as

$$\max_{\mathbf{a},\mathbf{p}} \quad \sum_{k\in\mathcal{K}} R_k(\mathbf{a},\mathbf{p}) \quad (8a)$$
$$\text{s.t.} \quad a_{ki} = \{0,1\},\ \forall k,i \quad (8b)$$
$$\sum_{k\in\mathcal{K}} a_{ki} \le 1,\ \forall i \quad (8c)$$
$$N_k = \sum_{i\in\mathcal{N}} a_{ki} \le N,\ \forall i \quad (8d)$$

$$\sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}} p_{ki} \le P_{\max},\ \forall k,i \tag{8e}$$

$$p_{ki} \le a_{ki}P_{\max},\ \forall k,i \tag{8f}$$

$$\gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}) \ge \bar{\gamma},\ \forall k. \tag{8g}$$

In constraint (8f), we note that $p_{ki} = 0$ if $a_{ki} = 0$; otherwise, $p_{ki} \le P_{\max}$. As a result, the SNR in (2) becomes $\gamma_{ki}(a_{ki}, p_{ki}) = p_{ki}|h_{ki}|^2/\sigma_k^2$.

It is very challenging to solve problem (8) due to the binary nature of $\mathbf{a}$. To overcome this issue, we relax binary variables $\mathbf{a}$ into continuous one, such as $0 \le a_{ki} \le 1$. The relaxed problem of (8) is expressed as

$$\max_{\mathbf{a},\mathbf{p}} \quad \sum_{k\in\mathcal{K}} R_k(\mathbf{a},\mathbf{p}) \tag{9a}$$

$$\text{s.t.} \quad 0 \le a_{ki} \le 1,\ \forall k,i \tag{9b}$$

$$(8c),(8d),(8e),(8f),(8g). \tag{9c}$$

We can observe that all constrains in (9) are linear. To tackle the non-concavity of $R_k(\mathbf{a},\mathbf{p})$, we introduce new variables $\boldsymbol{\phi} \triangleq \{\phi\}_{k\in\mathcal{K}}$ which satisfy $\phi_k \le \gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p})$. In addition, the optimal solution $\mathbf{a}^*$ of the relaxed problem (9) is often not exact binary at convergence, making the original problem (7) (or (8)) infeasible. Therefore, we incorporate a penalty function $\Psi(\mathbf{a}) \triangleq \sum_{i\in\mathcal{N},k\in\mathcal{K}}(a_{ki}^2 - a_{ki})$ [16] into the objective (9a). The function $\Psi(\mathbf{a})$ is always non-positive for $\mathbf{a} \in [0,1]$, which is useful to penalize the uncertainty of relaxed variables. As a result, problem (9) is equivalently rewritten as

$$\max_{\mathbf{a},\mathbf{p}} \quad \sum_{k\in\mathcal{K}} R_k(\phi_k) + \mu\Psi(\mathbf{a}) \tag{10a}$$

$$\text{s.t.} \quad \phi_k \le \gamma_k^{\text{eff}}(\mathbf{a},\mathbf{p}),\ \forall k \tag{10b}$$

$$\phi_k \ge \bar{\gamma},\ \forall k \tag{10c}$$

$$(8c),(8d),(8e),(8f),(8g),(9b) \tag{10d}$$

where $\mu$ is a penalty parameter and

$$R_k(\phi_k) = \Big\lfloor \frac{d_k}{nT} \Big\rfloor \Big( \log_2(1+\phi_k) - \sqrt{\frac{1}{0.48T}} Q^{-1}(\epsilon_k) \Big)$$

which is a concave function. It is noted that with an appropriate and sufficiently large value of $\mu$, problems (9) and (10) are equivalent. We are now in position to convexify the objective (10a) and constraint (10b). For the objective (10a), the function $\Psi(\mathbf{a})$ is convex that is useful to apply the inner approximation method to linearize it [17]. At iteration $\kappa$ of an iterative algorithm presented shortly, $\Psi(\mathbf{a})$ is iteratively replaced by

$$\Psi^{(\kappa)}(\mathbf{a}) := \sum_{i\in\mathcal{N},k\in\mathcal{K}} \big(2a_{ki}^{(\kappa)}a_{ki} - a_{ki} - (a_{ki}^{(\kappa)})^2\big) \tag{11}$$

where $a_{ki}^{(\kappa)}$ is a feasible point of $a_{ki}$. It is clear that $\Psi^{(\kappa)}(\mathbf{a}) \le \Psi(\mathbf{a})$ and $\Psi^{(\kappa)}(\mathbf{a}^{(\kappa)}) = \Psi(\mathbf{a}^{(\kappa)})$.

For constraint (10b), we first rewrite it as

$$\frac{1}{N}\sum_{i\in\mathcal{N}_k} \exp\Big(-\frac{2p_{ki}|h_{ki}|^2}{\beta\sigma_k^2}\Big) - \exp\Big(-\frac{1}{\beta}\phi_k\Big) \le 0. \tag{12}$$

As $\psi_k(\phi_k) \triangleq \exp\big(-\frac{1}{\beta}\phi_k\big)$ is convex in $\phi_k$, we apply the inner approximation method to approximate it around the feasible point $\phi_k^{(\kappa)}$ as

$$\psi_k^{(\kappa)}(\phi_k) = \psi_k(\phi_k^{(\kappa)}) + \nabla_{\phi_k}\phi_k^{(\kappa)}\big(\phi_k - \phi_k^{(\kappa)}\big) \tag{13}$$

**Algorithm 1** Iterative Algorithm to Solve Problem (7)

1: **Initialize**: Set $\kappa = 1$ and generate an initial feasible point for $\boldsymbol{\phi}^{(0)}$ and $\mathbf{a}^{(0)}$ for all constraints in (15);
2: **repeat**
3: Solve (15) to obtain the optimal solution $(\mathbf{a}^*, \mathbf{p}^*, \boldsymbol{\phi}^*)$;
4: Update $\mathbf{a}^{(\kappa)} := \mathbf{a}^*$ and $\boldsymbol{\phi}^{(\kappa)} := \boldsymbol{\phi}^*$;
5: Set $\kappa = \kappa + 1$;
6: **until** Convergence
7: **Output:** $(\mathbf{a}^*, \mathbf{p}^*, \boldsymbol{\phi}^*)$;

**Algorithm 2** Exhaustive Search Algorithm to Select the best MCS

1: **Initialization:** $\beta = \{\beta_1, \beta_2, \ldots, \beta_q\}$, $\max_\phi = 0$, $\beta_{best} = 0$, and $\phi_k^\beta = [\,]$
2: **for** $m = 1, 2, \ldots, q$ **do**
3: Solve **Algorithm 1** and append $\phi_k^*$ in $\phi_k^\beta$
4: $\max_\phi \leftarrow \phi_k^{\beta_1}$ and $\beta_{best} \leftarrow \beta_1$
5: **if** $\phi_k^{\beta_q} > \max_\phi$ **then**
6: $\max_\phi \leftarrow \phi_k^{\beta_{best}}$ and $\beta_{best} \leftarrow \beta_q$
7: **end if**
8: **end for**
9: **Output:** $\beta_{best}$

$$:= \exp\big(-\frac{1}{\beta}\phi_k^{(\kappa)}\big) + \frac{1}{\beta}(\phi_k^{(\kappa)})^2 \exp\big(-\frac{1}{\beta}\phi_k^{(\kappa)}\big) - \frac{1}{\beta}\phi_k^{(\kappa)} \exp\big(-\frac{1}{\beta}\phi_k^{(\kappa)}\big)\phi_k \tag{14}$$

which is a global lower bound of $\psi_k(\phi_k)$, satisfying $\psi_k^{(\kappa)}(\phi_k) \le \psi_k(\phi_k)$ and $\psi_k^{(\kappa)}(\phi_k^{(\kappa)}) = \psi_k(\phi_k^{(\kappa)})$.

Summing up, we solve the following approximate convex program of (10) at iteration $\kappa$

$$\max_{\mathbf{a},\mathbf{p},\boldsymbol{\phi}} \quad \sum_{k\in\mathcal{K}} R_k(\phi_k) + \mu\Psi^{(\kappa)}(\mathbf{a}) \tag{15a}$$

$$\text{s.t.} \quad \frac{1}{N}\sum_{i\in\mathcal{N}_k} \exp\Big(-\frac{2p_{ki}|h_{ki}|^2}{\beta\sigma_k^2}\Big) \le \psi_k^{(\kappa)}(\phi_k), \forall k \tag{15b}$$

$$(8c),(8d),(8e),(8f),(8g),(9b),(10c). \tag{15c}$$

The proposed iterative for solving (7) is summarized in Algorithm 1. Under an initial feasible point $\boldsymbol{\phi}^{(0)}$ and $\mathbf{a}^{(0)}$, Algorithm 1 generates a sequence of better points and a non-decreasing sequence of the objective values, which arrives to at least a locally optimal solution [17]. A large value of $\mu \gg 1$ penalizes the objective function for values of $a$ that are not 0 or 1. The per-iteration computational complexity of solving (15) in Step 3 of Algorithm 1 is $\mathcal{O}(\sqrt{4KN + 2N + 3K}\big(2KN + K)^3\big)$.

Finally, to select the best MCS, BS examines Algorithm 1 for all available values of $\beta$ and chooses the MCS which avails the maximum total achievable rate over all UEs. Algorithm 2 describes the key steps based on the exhaustive search method. For every $\beta$ value, we solve Algorithm 1 to obtain the optimal solution $\boldsymbol{\phi}^*$ and store it into $\boldsymbol{\phi}^\beta$. The optimal value of $\beta_m$ is found, corresponding to the maximum total achievable rate considered as the reference for the MCS selection.

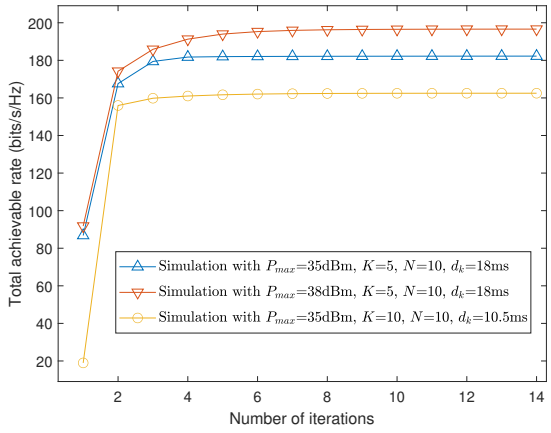


Fig. 2: Convergence behavior of Algorithm 1.

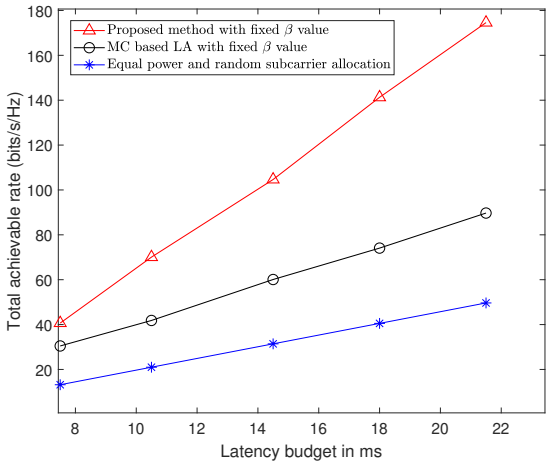


Fig. 3: Total achievable rate versus the latency budget.

## IV. Numerical Analysis and Discussions

We consider a downlink FDMA-based URLLC system with one BS and 5 URLLC users. We consider Rayleigh fading channel with the total number of subcarriers as $\frac{B}{N} = 10$. If the BPSK modulation scheme with subcarrier spacing 30 KHz and TTT is 0.5 ms is considered, we can transmit 30 kHz $\times 10 \times 0.5$ ms $= 150$ bits within one TTI. To simplify, one TTI is equivalent to one channel use. The total latency budget for the system is considered as $2d_k$ channel uses. As we consider the HARQ retransmission process with two retransmissions, then the latency budget should be at least twice of the transmission time required for transmitting the entire blocklength $T$, which is $\lfloor \frac{d_k}{2T} \rfloor$. The total power budget of the BS is varied between 28 dBm to 40 dBm, and the The blocklength is set to 7 channel uses. The target packet error rate is set in the range of $[10^{-2}, 10^{-8}]$. The AWGN power density at each subcarrier is assumed to be $10^{-15}$ W. The simulation is averaged over 1000 channel realizations.

Fig. 2 plots the convergence behavior of the proposed Algorithm 1. Different parameter (BS power, number of UEs, number of subcarriers, and latency budget in channel uses) has been selected to justify the convergence of our proposed algorithm. In most cases, Algorithm 1 converges within a few iterations, i.e. less than 10 iterations. As expected, we see that the higher transmission power budget at the BS and higher latency budget can offer significant better total achievable rate. For comparison purpose, we consider two benchmark schemes: $i$) "MC based LA with fixed $\beta$ value", where LA with 5G URLLC configuration was studied for MC-based networks [8]; and $ii$) "Equal power and random subcarrier allocation", where all UEs are allocated the same power and random subcarrier allocation is considered. To simplify the first benchmark, we consider a single connectivity-based network with a fixed $\beta$ value. In Fig. 3, we show the total achievable rate of different resource allocation schemes versus the latency budget. The same finite blocklength has been considered in both transmission and retransmission. As mentioned previously, the same packets will be transmitted twice to maintain higher reliability. We observe that the proposed algorithm offers the best achievable rate compared with two benchmark schemes, which demonstrates the effectiveness of joint optimization of power and carrier allocation. On the other hand, as the latency budget of the system increases, the total achievable rate also increases. This is because for a fixed number of bits available to transmit, more retransmissions can take place to increase the achievable rate. Fig. 4 plots

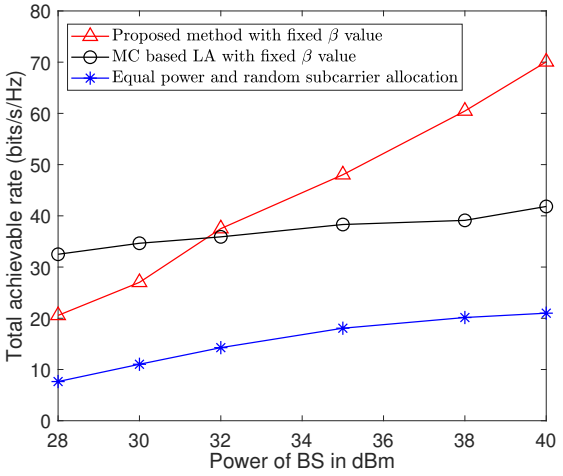


Fig. 4: Total achievable rate versus the total power budget of BS.

the total achievable rate versus the BS's power budget for different resource allocation schemes. As expected, the total achievable rate of all the considered schemes increases when $P_{\max}$ increases. Algorithm 1 dramatically outperforms the equal power and random subcarrier allocation scheme and MC based LA with fixed $\beta$ for high value of $P_{\max}$. In Fig. 5, we plot the total achievable rate versus PER. It is clear that when PER decreases, the effective SNR of all UEs increases. In addition, the inverse Q-function value increases with PER

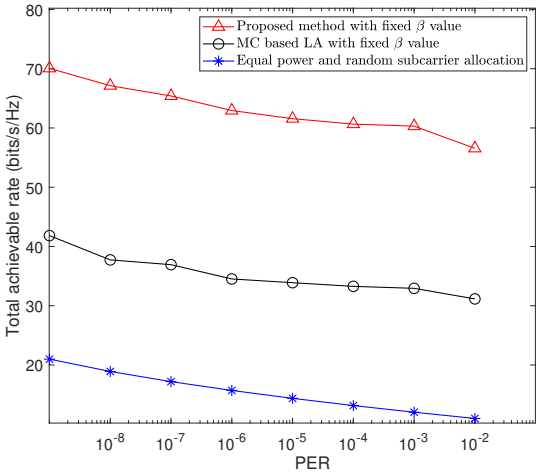


Fig. 5: Total achievable rate versus PER.

decreases, and it is negated by the AWGN rate. The proposed algorithm with joint optimization of both power and subcarrier provides better efficient estimated SNR as well as data rate.

Finally, Fig. 6 illustrates the impact of different values of $\beta$ on the system performance. As observed from the figure, the achievable rate first increases and then decreases when $\beta$ increases. This phenomenon suggests that there exists an appropriate value of $\beta$ to select the best MCS.

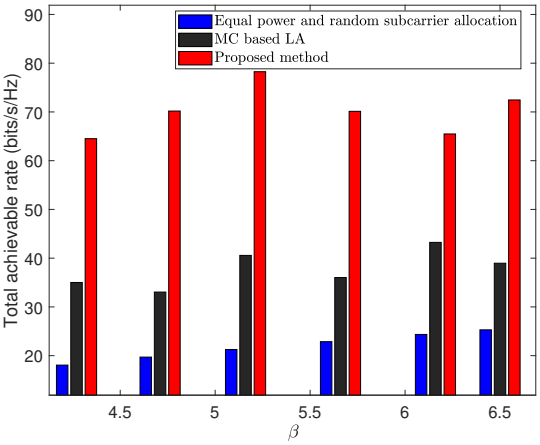


Fig. 6: Total achievable rate for different $\beta$.

## V. Conclusion

In this work, we have studied a joint subcarrier and power allocation problem to maximize the achievable sum rate of all users under an appropriate MCS. The formulated problem is a mixed-integer nonconvex programming, where the existing approaches are not applicable to solve it directly. Alternatively, we first transformed it into a more tractable form by relaxing binary variables and introducing new variables to facilitate the optimization. An efficient iterative based on the inner approximation framework was then proposed, which achieves at least a locally optimal solution. Extensive numerical results provided to demonstrate the merits of the proposed algorithms. In this work, we considered perfect channel state information available at the BS. In extension of this work, we would focus on the impact of imperfect channel state information in resource optimization for stringent URLLC services.

## References


[1] 3GPP, "Study on scenarios and requirements for next generation access technologies," *TR 38.913 v14.2.0, Tech. Rep.*, March 2017.

[2] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124–130, 2018.

[3] H. V. K. Mendis and F. Y. Li, "Achieving ultra reliable communication in 5g networks: A dependability perspective availability analysis in the space domain," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2057–2060, 2017.

[4] G. Qiu, M.-M. Zhao, M. Lei, and M.-j. Zhao, "Throughput maximization for polar coded ir-harq using deep reinforcement learning," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, 2020, pp. 1–6.

[5] A. Belogaev, E. Khorov, A. Krasilov, D. Shmelkin, and S. Tang, "Conservative link adaptation for ultra reliable low latency communications," in *2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, 2019, pp. 1–5.

[6] M. Deghel, S. E. Elayoubi, A. Galindo-Serrano, and R. Visoz, "Joint optimization of link adaptation and harq retransmissions for urllc services," in *2018 25th International Conference on Telecommunications (ICT)*, 2018, pp. 21–26.

[7] W. Anwar, K. Kulkarni, N. Franchi, and G. Fettweis, "Physical layer abstraction for ultra-reliable communications in 5g multi-connectivity networks," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1–6.

[8] J. Khan and L. Jacob, "Link adaptation for multi-connectivity enabled 5g urllc: Challenges and solutions," in *2021 International Conference on COMmunication Systems NETworkS (COMSNETS)*, 2021, pp. 148–152.

[9] 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. 28 912–28 922, 2018.

[10] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint power and blocklength optimization for urllc in a factory automation scenario," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1786–1801, 2020.

[11] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over nakagami fading channels," *Wireless Personal Communications*, vol. 13, p. 119–143, 2000.

[12] M. P. Mota, D. C. Araujo, F. H. Costa Neto, A. L. F. de Almeida, and F. R. Cavalcanti, "Adaptive modulation and coding based on reinforcement learning for 5g networks," in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[13] R. A. Delgado, K. Lau, R. Middleton, R. S. Karlsson, T. Wigren, and Y. Sun, "Fast convergence outer loop link adaptation with infrequent updates in steady state," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017, pp. 1–5.

[14] S. Lagen, K. Wanuga, H. Elkotby, S. Goyal, N. Patriciello, and L. Giupponi, "New radio physical layer abstraction for system-level simulations of 5g networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

[15] C. Li, N. Yang, and S. Yan, "Optimal transmission of short-packet communications in multiple-input single-output systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7199–7203, 2019.

[16] H. V. Nguyen, V.-D. Nguyen, O. A. Dobre, D. N. Nguyen, E. Dutkiewicz, and O.-S. Shin, "Joint power control and user association for NOMA-based full-duplex systems," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 8037–8055, 2019.

[17] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, May 2010.