Downlink MIMO-NOMA for Ultra-Reliable Low-Latency Communications

Chiyang Xiao, Jie Zeng, Wei Ni, Xin Su, Ren Ping Liu, Tiejun Lv and Jing Wang

***Abstract*—With the emergence of the mission-critical Internet of Things (IoT) applications, ultra-reliable low-latency communications (URLLC) are attracting a lot of attention. Non-orthogonal multiple access (NOMA) with multiple-input multiple-output (MIMO) is one of the promising candidates to enhance connectivity, reliability and latency performance of the emerging applications. In this paper, we derive a closed-form upper bound for the delay target violation probability in downlink MIMO-NOMA, by applying stochastic network calculus to the Mellin transforms of service processes. A key contribution is that we prove the infinite-length Mellin transforms resulting from the non-negligible interferences of NOMA, are Cauchy convergent, and can be asymptotically approached by a finite truncated binomial series in closed form. By exploiting the asymptotically accurate truncated binomial series, another important contribution is that we identify the critical condition for the optimal power allocation of MIMO-NOMA to achieve consistent latency and reliability between the receivers. The condition is employed to minimize the total transmit power, given a latency and reliability requirement of the receivers. It is also used to prove that the minimal total transmit power needs to change linearly with the path losses, to maintain latency and reliability at the receivers. This enables the power allocation for mobile MIMO-NOMA receivers to be effectively tracked. Extensive simulations corroborate the accuracy and effectiveness of the proposed model and the identified critical condition.**

***Index Terms*—URLLC, MIMO-NOMA, stochastic network calculus, delay violation probability, power allocation**

I. Introduction

Motivated by the explosive growth of mobile data requirement and number of communication devices boosted by the Internet of Things (IoT) [1], the fifth generation (5G) wireless system is anticipated to support wireless connectivity for both human centric and machine type services with guaranteed quality of service (QoS). Two usage scenarios in 5G, namely massive machine type communications (mMTC) and ultra reliable and low latency communications (URLLC) [2], are designed for IoT applications and distinguish 5G from previous generations. It is of crucial importance to achieve high reliability and low latency, while supporting a large number of connectivities for many IoT use cases, especially for mission critical tasks, such as factory automation, remote surgery, and intelligent transportation systems [3]. Typical emerging IoT applications require a latency from 0.25 ms to 10 ms and an outage probability (or packet loss rate) in the order of 10^{-3} to 10^{-9} [1]. It is also common for many IoT applications, such as unmanned aerial vehicle (UAV) communications and wireless sensors systems, to simultaneously provide services to a large number of devices, with limited bandwidth resources but extremely stringent statistical delay QoS. The demand on massive connectivity and low latency implicates the use of non-orthogonal multiple access (NOMA) [4], in coupling with multiple-input multiple-output (MIMO) [5], or "MIMO-NOMA" for short, due to its potential to enhance reliability [6] and latency [7]. Moreover, the striking overload factor of MIMO-NOMA can significantly improve spectral efficiency of wireless systems, hence remarkably increasing connectivities. NOMA is a promising access technique for the massive connectivity of 5G underlying different usage scenarios, including URLLC [4].

Both URLLC and NOMA are potentially the key components in future 5G networks. It is critical that they operate jointly and effectively to fulfill the potential of the networks. However, URLLC and NOMA have been studied in parallel so far. No work has jointly considered both, while the separately developed solutions for URLLC and NOMA provide little interoperability [1]. As a matter of fact, none of existing NOMA techniques have been designed to provide consistent reliability and low latency, due to the inter-user interference incumbent to NOMA. It is typically challenging to analyze the reliability and latency in the presence of interference [8]; leave alone optimizing them.

The authors in [3] defined the reliability of URLLC as the probability that the latency does not exceed a pre-described deadline. This definition emphasized on the importance of statistical delay QoS analysis and optimization for the transmission schemes in URLLC. Although there have been many studies on the physical layer power allocation to maximize system throughput or minimize outage probability for NOMA [9]-[10] and MIMO-NOMA [11], there have been few investigations on the network layer performance under statistical delay

Manuscript received June 22, 2018; revised December 10, 2018; accepted January 25, 2019.

This work was supported by the National Science and Technology Major Project (No. 2018ZX03001004-003), the Science and Technology Program of Beijing (No. D171100006317002), the Ministry of Education and China Mobile Joint Scientific Research Fund (Grant No. MCM20160105) and the Huawei Innovation Research Program (No. HIRPO20160105).

C. Xiao, X. Su and J. Wang are with the Department of Electronic Engineering, Tsinghua University, Beijing, China. (e-mail: xiao-cy16@mails.tsinghua.edu.cn, suxin@tsinghua.edu.cn, and wangj@tsinghua.edu.cn).

J. Zeng is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, and School of Electrical and Data Engineering, University of Technology Sydney, Sydney, Australia. e-mail: (zengjie@tsinghua.edu.cn).

W. Ni is with DATA61, CSIRO, Sydney, NSW 2122, Australia (e-mail: wei.ni@csiro.au).

R. P. Liu is with GBDTC, UTS, Sydney, NSW 2007, Australia (e-mail: renping.liu@uts.edu.au).

T. Lv is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. e-mail: (lvtiejun@bupt.edu.cn).

QoS constraint. Only a few works, e.g., [12] and [13], have jointly considered the power control in the physical layer and the statistical delay QoS constraint in the network layer. Both [12] and [13] exploited the concept of effective capacity [14] to characterize statistical delay QoS. The effective capacity of NOMA systems was maximized under delay QoS constraints. Therein, the requirement on delay QoS was parameterized by the exponential decay factor θ of the queue backlog, but the delay target violation probability was not adequately quantified. The authors of [15] analyzed the achievable link-layer rate in different signal-to-noise ratio (SNR) regions from an effective capacity perspective. It was revealed how the link-layer rate changes with the statistical delay QoS requirement of NOMA users and power allocation can be properly designed to improve the link-layer rate performance. Note that these existing studies [12]- [15] concerning statistical delay performance of NOMA have only assumed single-antenna scenarios. Research on stochastic delay QoS performance analysis and power control for MIMO-NOMA, where a base station (BS) with multiple antennas simultaneously serves multiple groups of users [5], has yet to be addressed in the literature.

This paper proposes a new model, which parameterizes the latency and reliability of service processes in downlink MIMO-NOMA. It uses stochastic network calculus (SNC) on the (min,$\times$) dioid algebra to translate the intractable delay profile (or more specifically, delay violation probability) of a pair of NOMA receivers to the deconvolution of the arrival and processes. An asymptotic upper bound of the delay violation probability is developed with the Mellin transform of the deconvolution, which is non-trivial though, due to the inter-user interference incumbent to NOMA. By using binomial expansion, we write the Mellin transforms as infinite series, and prove that the series are Cauchy convergent and can be asymptotically accurately approximated by their closed-form truncated version.

The key contributions of the paper can be summarized as follows.

- By carrying out the SNC analysis and Mellin transforms on the service process, we derive closed-form asymptotic upper bound of the delay violation probability for a pair of NOMA recivers. The analysis is non-trivial, as the Mellin transform of the service process is challenging in the presence of interference (which is incumbent to NOMA) and has not been addressed in the literature.
- Exploiting the upper bound, we identify the sufficient and necessary condition for the optimal transmit power under the delay QoS and user fairness, confirm the continuity and monotonicity of the delay violation probability over the powers, and qualify the use of simple bisection search for the optimal powers. In contrast, there has been little consideration on delay and reliability in existing power approaches for NOMA.
- Closed-form expressions for the optimal power allocation are derived in the case where the channel difference of the receiver pair is large. It is revealed that the optimal transmit powers for guaranteeing the delay violation probability of the pair are proportional to their path losses.

In a different yet relevant context, SNC has been increasingly used to describe the upper bound for queueing delays or backlogs, since the explicit queueing delay profiles are difficult to achieve due to the strict assumptions of the queueing theory and the randomness of fading channels. Network calculus emerged first as a theory that analyzes performance guarantees of queuing systems on a (min,+) dioid algebra in computer science. Network calculus can be divided into deterministic network calculus and SNC [16]. The deterministic network calculus models the arrival and service processes as deterministic envelop functions (also known as the arrival curve and the serive curve), and cannot capture the stochastic arrivals and services. SNC relaxes the deterministic envelops to be statistical ones, e.g., by introducing a pre-defined envelop violation probability [17]. SNC has been used for statistical delay analysis in fading channels, first in the bit domain [18], [19], where closed-form expressions were not tractable due to logarithm operation in the domain. In [20], a (min,$\times$) SNC was developed to present the fading channels in the *SNR domain*, where the SNR distribution at the receiver was used to describe the channel properties. Logarithm operations were suppressed, and closed-form results became possible.

The (min,$\times$) SNC represents the non-asymptotic probabilistic performance bounds in terms of the distribution of fading channels and arrival processes [21], relaxing the intractable delay target violation probability to the tractable upper bound. Based on the upper bounds obtained via (min,$\times$) SNC, a cross-layer power control framework was proposed in [22] for a single device in WirelessHART systems. The framework was further extended into a multi-hop version in [23] to minimize power consumption under statistical end-to-end delay constraints. Utilizing SNC, statistical delay QoS analysis was performed in [24] for millimeter-wave multi-hop systems with full-duplex buffered relays. But the delays involved in these works were in the range of tens to hundreds of milliseconds, far beyond the scope of URLLC. SNC was suggested by [8] and [25] to capture the “tail behavior”, i.e. queueing delay profile of URLLC transmissions. [26] investigated the network layer performance of multiple-input single-output (MISO) systems under statistical delay constraints. Probabilistic delay bounds were derived using SNC for URLLC. Distinctively different from these works, MIMO-NOMA undergoes strong interferences between receivers. The analysis of Mellin transforms, a critical step following the SNC, becomes non-trivial. To the best of our knowledge, none of the existing works have solved the Mellin transforms in the presence of non-negligible interferences, or can be extended to MIMO-NOMA.

The remainder of this paper is organized as follows. In section II, the system model for NOMA transmission is described. Section III introduces the fundamentals of the (min,$\times$) SNC and derives the upper bounds of of the delay target violation probabilities for downlink MIMO-NOMA. Section IV presents the cross-layer power control algorithm based on the derived upper bounds. Simulation results and analysis are presented in section V, and section VI concludes this paper. Notations used in the rest of the paper are listed in Table I.

TABLE I
IMPORTANT NOTATIONS

Notation	Description
p_m, q_m	Strong and weak receivers of the m-th receiver pair
$\eta_{p_m}^2$, $\eta_{q_m}^2$	Power allocation coefficients for receivers p_m and q_m
ρ_m	Total transmit power of the m-th receiver pair
$\mathbf{l}_m = [l_{p_m}, l_{q_m}]$	Receiver-to-BS distance pair
β	Path loss exponent
$\mathbf{H}_k$	Channel matrix between receiver k and the BS
$\mathbf{v}_{p_m}$, $\mathbf{v}_{q_m}$	Detection vectors applied at receivers p_m and q_m
$\mathbf{c}_m$	Precoding vector for the m-th receiver pair
$A_k(\tau,t)$, $S_k(\tau,t)$, $D_k(\tau,t)$	Cumulative arrival, service and departure processes in bit domain for receiver k
$\mathcal{A}_k(\tau,t)$, $\mathcal{S}_k(\tau,t)$	Cumulative arrival and service processes in SNR domain for receiver k
$a_k(t)$, $r_k(t)$	Instantaneous arrival and service in bit domain for receiver k
$\alpha_k(t)$, $\phi_k(t)$	Instantaneous arrival and service in SNR domain for receiver k
$\mathcal{M}_X(s,\tau,t)$	Mellin transform of $X(\tau,t)$ with parameter s
$\mathcal{K}_k(s,-w)$	The steady-state kernel
$\hat{B}_k(w)$	Upper bound of delay violation probability of receiver k with delay target w
ϵ	Target delay violation probability

II. SYSTEM MODEL

Consider a MIMO-NOMA system, where a BS with M antennas serves $2M$ randomly distributed receviers at the same time and frequency, as shown in Fig. 1. The $2M$ receivers are grouped into M pairs, according to their channel conditions. Each receiver pair consists of two receivers with different fading channel gains. The BS provides services to the M pairs of receivers by M beams. Each pair of receivers in a beam are multiplexed in a non-orthogonal fashion. Assume that in the m-th ($1 \le m \le M$) receviers pair, receiver p_m is closer to the BS than receiver q_m. Hence, receiver p_m has stronger channel condition and is referred to as the strong receiver, whereas receiver q_m is the weak receiver. The number of antennas equipped at each receiver is N. In this paper, $N > M/2$ is assumed to implement the transmission scheme based on signal alignment [5], where signals are superimposed (or aligned) in the desired signal space or direction by carefully designing the precoding and detection vectors for each receiver. With signal alignment, co-channel interference can be suppressed by exploiting the extra degrees of freedom provided by the multiple antennas transmitter and receivers. At the t-th time slot, the channel matrix between the BS and receiver k ($k \in \{p_1, \cdots, p_M\} \cup \{q_1, \cdots, q_M\}$) is denoted by $\mathbf{H}_k(t) = \frac{\mathbf{G}_k(t)}{l_k^{\beta/2}}$, where l_k is the distance between the BS and the receiver, β is the path loss exponent, and $\mathbf{G}_k(t) \in \mathbb{C}^{N\times M}$ denotes the small scale fading with independently and identically distributed (i.i.d.) circular symmetric complex Gaussian (CSCG) random variables. We further assume that the channels are block fading, i.e. $\mathbf{G}_k$ remains unchanged within a time slot, and changes independently between successive time slots.

This paper investigates the cross-layer power control under the general MIMO-NOMA framework proposed in [5]. Downlink MIMO-NOMA transmission is implemented by superimposing the signals destined for receiver p_m and q_m at the m-th transmit antenna port. As a result, the transmit signal vector at the BS at the t-th time slot is given by

$$\mathbf{s}(t) = \begin{bmatrix} \eta_{p_1} s_{p_1}(t) + \eta_{q_1} s_{q_1}(t) \\ \vdots \\ \eta_{p_M} s_{p_M}(t) + \eta_{q_M} s_{q_M}(t) \end{bmatrix}, \tag{1}$$

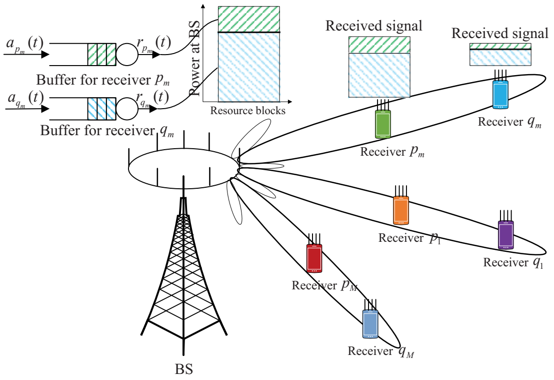


Fig. 1. Downlink MIMO-NOMA system. The signals of different receiver pairs are orthogonal in signal space. Receivers in the same pair are served in the NOMA fashion.

where $s_{p_m}(t)$ and $s_{q_m}(t)$ denote the signal intended for the receivers p_m and q_m, η_{p_m} and η_{q_m} are the square root of the power allocation coefficients for the receivers p_m and q_m, and $\eta_{p_m}^2 + \eta_{q_m}^2 = 1$. The BS precodes the signal vector with an $M \times M$ precoding matrix $\mathbf{P}(t)$, and then transmits the precoded signal to the M pairs of receivers. For a receiver k in the m-th receiver pair, i.e. $k \in \{p_m, q_m\}$, the received signal can be expressed as

$$\mathbf{y}_k(t) = \mathbf{H}_k(t)\mathbf{P}(t)\mathbf{s}(t) + \mathbf{n}_k(t), \tag{2}$$

where $\mathbf{n}_k(t) \in \mathbb{C}^{N\times 1}$ is the noise vector at the receiver. The receiver applies an $N\times 1$ detection vector $\mathbf{v}_k(t)$ to the received signal, leading to the following detection result

$$\begin{aligned} &\mathbf{v}_k^H(t)\mathbf{y}_k(t) = \mathbf{v}_k^H(t)\mathbf{H}_k(t)\mathbf{P}(t)\mathbf{s}(t) + \mathbf{v}_k^H(t)\mathbf{n}_k(t) \\ &= \frac{\mathbf{v}_k^H(t)\mathbf{G}_k(t)}{l_k^{\frac{\beta}{2}}}\mathbf{p}_m(t)(\eta_{p_m}s_{p_m}(t) + \eta_{q_m}s_{q_m}(t)) + \\ &\sum_{i\neq m} \frac{\mathbf{v}_k^H(t)\mathbf{G}_k(t)}{l_k^{\frac{\beta}{2}}}\mathbf{p}_i(t)(\eta_{p_i}s_{p_i}(t) + \eta_{q_i}s_{q_i}(t)) + \mathbf{v}_k^H(t)\mathbf{n}_k(t), \end{aligned} \tag{3}$$

where $\mathbf{p}_m(t)$ is the precoding vector for the m-th receiver pair, i.e. the m-th column of the precoding matrix $\mathbf{P}(t)$.

According to the signal alignment [5], the detection vector can be designed as follows:

$$\begin{bmatrix}\mathbf{v}_{p_m}(t)\\ \mathbf{v}_{q_m}(t)\end{bmatrix} = \mathbf{U}_m(t)\mathbf{x}_m(t), \tag{4}$$

where $\mathbf{U}_m(t) \in \mathbb{C}^{2N\times(2N-M)}$ is a matrix containing the right singular vectors of the matrix $\begin{bmatrix}\mathbf{G}_{p_m}^H(t) & -\mathbf{G}_{q_m}^H(t)\end{bmatrix}$ corresponding to its zero singular values, $\mathbf{x}_m(t)$ is a random $(2N-M)\times 1$ vector satisfying $|\mathbf{x}_m(t)| = 2$. With such detection vectors, we have $\mathbf{v}_{p_m}^H(t)\mathbf{G}_{p_m}(t) = \mathbf{v}_{q_m}^H(t)\mathbf{G}_{q_m}(t)$, i.e. the signals for receivers p_m and q_m are aligned in the same direction. Signals for different receiver pairs are aligned in different directions. The downlink multi-receiver-pair MIMO-NOMA channel is decomposed into M pairs of independent single antenna NOMA channels. Readers are refered to [5] for more details. $\mathbf{g}_m(t) = \mathbf{G}_{p_m}^H(t)\mathbf{v}_{p_m}(t)$ is the effective channel vector of the m-th receiver pair. To eliminate the inter-pair interference, the precoding matrix satisfies the following constraint

$$\mathbf{g}_m^H(t)\mathbf{p}_i(t) = 0, \quad \forall i \neq m. \tag{5}$$

This leads to a zero forcing based precoding design, as given by

$$\mathbf{P}(t) = \mathbf{G}^{-H}(t)\mathbf{D}, \tag{6}$$

where $\mathbf{G}(t) = [\mathbf{g}_1(t),\cdots,\mathbf{g}_m(t),\cdots,\mathbf{g}_M(t)]^H$, and $\mathbf{D}$ is a diagonal matrix which specifies the transmit power for each receiver pair. More precisely, $\mathbf{D}^2 = \mathrm{diag}\{\frac{\rho_1}{\mathbf{c}_1^H(t)\mathbf{c}_1(t)},\cdots,\frac{\rho_M}{\mathbf{c}_M^H(t)\mathbf{c}_M(t)}\}$, where $\mathbf{c}_m(t)$ is the m-th column of $\mathbf{G}^{-H}(t)$, and ρ_m denotes the total transmit power for the m-th receiver pair. The zero forcing based precoding eliminates the co-channel interference between different receiver pairs, while the interference between the same pair remains.

Using the detection and precoding design in (4) and (6), the received signal after detection for the m-th receiver pair at the t-th time slot can be expressed as follows:

$$\begin{aligned}\mathbf{v}_{p_m}^H(t)\mathbf{y}_{p_m}(t) = &\frac{\sqrt{\rho_m}\Big(\eta_{p_m}s_{p_m}(t)+\eta_{q_m}s_{q_m}(t)\Big)}{\sqrt{\mathbf{c}_m^H(t)\mathbf{c}_m(t)l_{p_m}^{\beta}}}\\ &+\mathbf{v}_{p_m}^H(t)\mathbf{n}_{p_m}(t);\end{aligned} \tag{7}$$

$$\begin{aligned}\mathbf{v}_{q_m}^H(t)\mathbf{y}_{q_m}(t) = &\frac{\sqrt{\rho_m}\Big(\eta_{p_m}s_{p_m}(t)+\eta_{q_m}s_{q_m}(t)\Big)}{\sqrt{\mathbf{c}_m^H(t)\mathbf{c}_m(t)l_{q_m}^{\beta}}}\\ &+\mathbf{v}_{q_m}^H(t)\mathbf{n}_{q_m}(t).\end{aligned} \tag{8}$$

Due to the signal alignment, the two receivers in the same receiver pair share the same small scale fading gain $\frac{1}{\mathbf{c}_m^H(t)\mathbf{c}_m(t)}$ while experiencing different large scale fadings. Without of generality, we focus on the m-th receiver pair to evaluate the physical layer information rate of the strong receiver p_m and the weak receiver q_m. For the weak receiver q_m, it decodes its message by treating the signal intended for receiver p_m as a noise. Hence, the service amount provided to receiver q_m at the t-th time slot can be written as follows:

$$\begin{aligned}r_{q_m}(t) &= W\log_2\left(1+\frac{\frac{\rho_m}{\mathbf{c}_m^H(t)\mathbf{c}_m(t)l_{q_m}^{\beta}}\eta_{q_m}^2}{\frac{\rho_m}{\mathbf{c}_m^H(t)\mathbf{c}_m(t)l_{q_m}^{\beta}}\eta_{p_m}^2+\left|\mathbf{v}_{q_m}^H(t)\right|^2\sigma^2}\right)\\ &= W\log_2\left(1+\frac{\eta_{q_m}^2}{\eta_{p_m}^2+\mathbf{c}_m^H(t)\mathbf{c}_m(t)\left|\mathbf{v}_{q_m}^H(t)\right|^2/\bar{\gamma}_{q_m}}\right),\end{aligned} \tag{9}$$

where W is the number of symbols used in one trasmission, σ^2 is the noise power, and $\bar{\gamma}_{q_m} = \frac{\rho_m}{\sigma^2 l_{q_m}^{\beta}}$ is SNR at receiver q_m.

The strong receiver p_m carries out successive interference cancellation (SIC) by first decoding the message intended for receiver q_m and then cancelling it from the received singal. Similar to (9), the service rate provided to receiver p_m can be written as

$$\begin{aligned}r_{p_m}(t) &= W\log_2\left(1+\frac{\rho_m\eta_{p_m}^2}{\mathbf{c}_m^H(t)\mathbf{c}_m(t)l_{p_m}^{\beta}\left|\mathbf{v}_{p_m}^H(t)\right|^2\sigma^2}\right)\\ &= W\log_2\left(1+\frac{\eta_{p_m}^2}{\mathbf{c}_m^H(t)\mathbf{c}_m(t)\left|\mathbf{v}_{p_m}^H(t)\right|^2/\bar{\gamma}_{p_m}}\right),\end{aligned} \tag{10}$$

where $\bar{\gamma}_{p_m} = \frac{\rho_m}{\sigma^2 l_{p_m}^{\beta}}$.

Consider the case that the BS delivers information to multiple receivers with statistical QoS requirements. The QoS requirements can be described by a predefined queueing delay target w and a delay target violation probability ϵ. This paper aims to minimize the transmit power which ensures $\Pr\{w_k(t) > w\} < \epsilon$, where $w_k(t)$ is the queueing delay of receiver $k \in \{p_m, q_m\}$ at any time slot t, i.e. the number of time slots it takes to successfully deliver the information bits that arrive at time slot t. In order to count for the probabilistic delay constraint, we resort to the newly developed (min,×) SNC which characterizes the statistical performance bounds (such as delay bound and queue backlog bound) via the distribution of the traffic arrivals and channel fading process.

III. SNC for Downlink MIMO-NOMA

Recall that the downlink MIMO-NOMA system of interest is a discrete-time, fluid-flow queuing system, the cumulative arrival, service and departure processes between time slots τ and $(t-1)$ can be defined by bivariate processes $A_k(\tau,t) = \sum_{i=\tau}^{t-1} a_k(i)$, $S_k(\tau,t) = \sum_{i=\tau}^{t-1} r_k(i)$ and $D_k(\tau,t) = \sum_{i=\tau}^{t-1} d_k(i)$, where $a_k(i)$, $r_k(i)$ and $d_k(i)$ are the instantaneous traffic arrival to receiver $k \in \{p_m, q_m\}$, service offered to receiver k, and the corresponding departure from the BS, respectively. Denote the queue backlog for receiver k at time slot i by $Q_k(i)$. Then, the queue evolves according to $Q_k(i+1) = Q_k(i)+a_k(i)-d_k(i)$. For receiver k, $D_k(\tau,t) = \sum_{i=\tau}^{t-1} d_k(i)$ defines the cumulative depature from time slot τ to time slot $t-1$. We have $d_k(i) = \min\{Q_k(i)+a_k(i), r_k(i)\}$.

For a work-conserving first come first served (FCFS) queueing system, $w_k(t)$ is expressed as

$$w_k(t) = \inf\{u > 0 : A_k(0,t) \leq D_k(0,t+u)\}. \tag{11}$$

By substituting into (11) the dynamic server property $D_k(0,t) \geq \inf_{0\leq\tau\leq t}\{A_k(0,\tau)+S_k(\tau,t)\}$ [27], we can obtain an upper bound for the delay. The cumulative processes, $A_k(\tau,t)$, $S_k(\tau,t)$ and $D_k(\tau,t)$ ($k \in \{p_m, q_m\}$), are defined in the so-called bit domain where the processes are measured in number of bits. Unfortunately, the logarithmic operator in $r_k(t)$ prevents expressing the statistics of the service process in a simple closed form, resulting in analytical intractability. We propose to use the (min,×) SNC to convert these processes from the bit domain to the SNR domain by taking exponent arithmetic. Denote the SNR-domain counterparts of the cumulative arrival and service processes of receiver k by $\mathcal{A}_k(\tau,t) = e^{A_k(\tau,t)}$ and $\mathcal{S}_k(\tau,t) = e^{S_m(\tau,t)}$, respectively. For a bit-domain process $X(\tau,t)$, we use $\mathcal{X}(\tau,t)$ to denote its SNR-domain counterpart. Then, $\mathcal{X}(\tau,t)-1$ represents the minimal required SNR if there are $X(\tau,t)$ bits to transmit.

The (min,×) SNC can characterize the input-output relationship of a queueing system with the following deconvolution operator defined on the (min,×)-algebra:

$$\mathcal{U} \oslash \mathcal{V}(\tau,t) = \sup_{u\leq\tau}\left\{\frac{\mathcal{U}(u,t)}{\mathcal{V}(u,\tau)}\right\}, \tag{12}$$

where $\oslash$ stands for deconvolution. Accordingly, the queueing delay of receiver k ($k \in \{p_m, q_m\}$) at time slot t can be rewritten as

$$w_k(t) = \inf\{u \geq 0 : \mathcal{A}_k \oslash \mathcal{S}_k(t+u,t) \leq 1\}. \tag{13}$$

Hence, the queueing delay can be upper bounded by [21]

$$\begin{aligned}\Pr\{w_k(t) > w\} &\leq \Pr\{\mathcal{A}_k \oslash \mathcal{S}_k(t+w,t) > 1\}\\ &\leq \mathcal{M}_{\mathcal{A}_k\oslash\mathcal{S}_k}(1+s,t+w,t),\end{aligned} \tag{14}$$

where the first inequality is based on [21], and the second inequality is based on the well-known Chernoff's bound (i.e., for an arbitrary bivariate stochastic process $X(\tau,t)$, $\Pr\{X(\tau,t) \geq a\} \leq a^{-s}\mathcal{M}_X(1+s,\tau,t)\ \forall a > 0, s > 0$) [28]. $\mathcal{M}_X(s,\tau,t) = \mathbb{E}\left[(X(\tau,t))^{s-1}\right]$ is the Mellin transform of any nonnegative stochastic process $X(\tau,t)$ for any $s \in \mathbb{R}$ whenever the expectation exists. According to the property of the Mellin transform of the deconvolution [21], (14) can be further upper bounded by $\Pr\{w_k(t) > w\} \leq \inf_{s>0}\{\mathcal{K}_k(s,-w)\}$, where $\mathcal{K}_k(s,-w)$ is the steady-state kernel with the following expression [21]:

$$\mathcal{K}_k(s,-w) = \lim_{t\to\infty}\sum_{u=0}^{t}\mathcal{M}_{\mathcal{A}_k}(1+s,u,t)\mathcal{M}_{\mathcal{S}_k}(1-s,u,t+w). \tag{15}$$

Therefore, the upper bound of the delay violation probability of receiver k is

$$B_k(w) = \inf_{s>0}\{\mathcal{K}_k(s,-w)\} \geq \Pr\{w_k(t) > w\}. \tag{16}$$

Given the signal alignment based precoding and detection scheme, the downlink multi-receiver-pair MIMO-NOMA channel is decomposed into M pairs of independent single antenna NOMA channels [5]. Without loss of generality, we focus on the m-th receiver pair. In what follows, we assume that for each receiver $k \in \{p_m, q_m\}$, the cumulative arrival $A_k(\tau,t)$ is i.i.d. incremental processes. It is also reasonable to assume that $S_k(\tau,t)$ is an i.i.d. incremental process. This is because for receiver $k \in \{p_m, q_m\}$, the increment of the cumulative service process $S_k(\tau,t)$ at time slot i is $s_k(i)$. Consider that each receiver experiences block fading channel. $s_k(i)$ is independent between time slots, and has the same distribution at different time slots.

Denote the increments of $A_k(\tau,t)$ and $S_k(\tau,t)$ by a_k and r_k, respectively. Then, the Mellin transform of $\mathcal{A}_k(\tau,t)$ can be expressed as the product of the Mellin transforms of $a_k(i)$ ($(\tau \leq i \leq t-1)$), i.e.,

$$\begin{aligned}\mathcal{M}_{\mathcal{A}_k}(s,\tau,t) &= \mathbb{E}\left[\left(\prod_{i=\tau}^{t-1}e^{a_k(i)}\right)^{s-1}\right] = \left(\mathbb{E}\left[e^{a_k(s-1)}\right]\right)^{t-\tau}\\ &= \left(\mathcal{M}_{\alpha_k}(s)\right)^{t-\tau},\end{aligned} \tag{17}$$

where $\alpha_k = e^{a_k}$. Likewise, the Mellin transform of the service processes in the SNR domain can be given by

$$\begin{aligned}\mathcal{M}_{\mathcal{S}_k}(s,\tau,t) &= \mathbb{E}\left[\left(\prod_{i=\tau}^{t-1}e^{r_k(i)}\right)^{s-1}\right] = \left(\mathbb{E}\left[e^{r_k(s-1)}\right]\right)^{t-\tau}\\ &= \left(\mathcal{M}_{\phi_k}(s)\right)^{t-\tau},\end{aligned} \tag{18}$$

where $\phi_k = e^{r_k}$. By substituting (17) and (18), (15) can be rewritten as

$$\mathcal{K}_k(s,-w) = \frac{\mathcal{M}^{w}_{\phi_k}(1-s)}{1-\mathcal{M}_{\alpha_k}(1+s)\mathcal{M}_{\phi_k}(1-s)}, \tag{19}$$

which is meaningful under the so-called "stability condition" $\mathcal{Z}(s) = \mathcal{M}_{\alpha_k}(1+s)\mathcal{M}_{\phi_k}(1-s) < 1$ [21], [22]; otherwise, the summation in (15) would be unbounded.

As shown in (16) and (19), the upper bound of the queueing delay violation probability is established on the Mellin transforms of the arrival and service processes in the SNR domain. Therefore, evaluating the upper bound requires deriving the Mellin transforms of α_k and ϕ_k. In this paper, we assume that the arrivals with low rates and low burstiness can be modeled by a Poisson process. That is to say, a_k in (17) is a Poisson random variable with an average of λ_k bits. In turn, the Mellin transform of α_k can be derived as

$$\mathcal{M}_{\alpha_k}(s) = \sum_{n=0}^{\infty}e^{n(s-1)}\frac{(\lambda_k)^n}{n!}e^{-\lambda_k} = e^{\lambda_k(e^{s-1}-1)}. \tag{20}$$

In order to obtain the Mellin transform of ϕ_k, we have to derive the probability density function (pdf) of the effective signal-to-interference-plus-noise (SINR) of receiver $k \in \{p_m, q_m\}$. The stochastic characteristic of ϕ_k is determined by two random terms, i.e. the term corresponding to the small scale fading $\mathbf{c}_k^H\mathbf{c}_k$ and the term corresponding to the detection gain on noise $|\mathbf{v}_k^H|^2$. Here, we suppress the time slot index in the brackets for notational simplicity.

For the m-th receiver pair, it has been proved in [5] that $\frac{1}{\mathbf{c}_m^H\mathbf{c}_m}$ is exponentially distributed, from which we can readily derive the distribution of $\mathbf{c}_m^H\mathbf{c}_m$. In contrast, the distribution of $|\mathbf{v}_k^H|^2$ is intractable for $k \in \{p_m, q_m\}$. One reason is that $|\mathbf{v}_k^H|^2$ is correlated with $\mathbf{c}_m^H\mathbf{c}_m$. The other reason is the uncertainty of $\mathbf{x}_m(t)$ in (4). We opt to use the upper bound of $|\mathbf{v}_k^H|^2$ instead of its instantaneous value. Noting that $|\mathbf{v}_{p_m}^H|^2 + |\mathbf{v}_{q_m}^H|^2 = 2$, we have $|\mathbf{v}_k^H|^2 \leq 2$. This leads to a lower bound of service rate $\hat{r}_k$, i.e. $\hat{r}_k \leq r_k$. Let $\hat{\phi}_k = e^{\hat{r}_k}$, then $\mathcal{M}_{\phi_k}(1-s) \leq \mathcal{M}_{\hat{\phi}_k}(1-s)$ holds since $s > 0$. According to the function monotonicity rule, $\mathcal{K}_k(s, -w)$ monotonically increases with $\mathcal{M}_{\phi_k}(1-s)$ whenever the stability condition holds. This monotonicity leads to an upper bound of the steady-state kernel, denoted by $\hat{\mathcal{K}}_k(s, -w)$, which can still be used to determine the upper bound of the delay target violation probability.

Theorem 1. *Given the power allocation coefficients η_{p_m} and η_{q_m} of the m-th receiver pair, the upper bounds of $\mathcal{M}_{\phi_{p_m}}(1-s)$ and $\mathcal{M}_{\phi_{q_m}}(1-s)$ are given by*

$$\mathcal{M}_{\hat{\phi}_{p_m}}(1-s) = \left(\frac{\eta_{p_m}^2 \bar{\gamma}_{p_m}}{2}\right)^{-\mathcal{W}s} e^{\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{p_m}}} \times \Gamma\left(1 - \mathcal{W}s, \frac{2}{\eta_{p_m}^2 \bar{\gamma}_{p_m}}\right) \tag{21}$$

and

$$\begin{aligned}&\mathcal{M}_{\hat{\phi}_{q_m}}(1-s) = \lim_{K\to\infty} \eta_{p_m}^{2\mathcal{W}s} \\ &\Bigg[1 - \mathcal{W}s\pi_m e^{\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}} E_i\left(-\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}\right) + \frac{\mathcal{W}s(\mathcal{W}s+1)\pi_m^2}{2} \\ &\times \left(e^{\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}} E_i\left(-\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}\right) + \frac{\eta_{p_m}^2 \bar{\gamma}_{q_m}}{2}\right) \\ &+ \sum_{n=3}^{K}\sum_{k=1}^{n-1} \frac{(k-1)!}{(n-1)!}\frac{(-\mathcal{W}s)^{\underline{n}}}{n!}\pi_m^n(-1)^{k+1}\left(\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}\right)^{-k} \\ &+ \sum_{n=3}^{K}\frac{(-\mathcal{W}s)^{\underline{n}}}{n!}\frac{e^{\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}}}{(n-1)!}E_i\left(-\frac{2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}\right)\Bigg],\end{aligned} \tag{22}$$

where $\mathcal{W} = W/\ln 2$, $\Gamma(x, a) = \int_a^\infty t^{x-1}e^{-t}dt$ is the upper incomplete Gamma function, $\pi_m = \frac{2\eta_{q_m}^2}{\eta_{p_m}^2 \bar{\gamma}_{q_m}}$, $E_i(x) = \int_{-x}^\infty \frac{e^{-t}}{t}dt$ is the exponential integral and $(x)^{\underline{n}}$ denotes the n-th falling factorial power of a real variable x, also known as the Pochhammer symbol [29], and is given by $(x)^{\underline{n}} = (x)(x-1)\cdots(x-n+1)$.

Proof: Please refer to Appendix A. ■

It is worth mentioning that the expression for $\mathcal{M}_{\hat{\phi}_{q_m}}(1-s)$ in (22) includes an infinite series as the result of the general binomial expansion. Nevertheless, we are able to prove that (22) is convergent by the following theorem.

Theorem 2. *Denote the summation of the first K terms in* (22) *by F_K, i.e., $F_K = \eta_{p_m}^{2\mathcal{W}s}\sum_{n=0}^{K-1} f_n$, where f_n is the n-th term in the square brackets of* (22)*. Then, $\{F_n\}_{n\geq 0}$ is a Cauchy sequence. As the limit of $\{F_n\}_{n\geq 0}$, $\mathcal{M}_{\hat{\phi}_{q_m}}(1-s) = \lim_{n\to\infty} F_n$ exists and can be asymptotically approached by F_K, provided K is sufficiently large.*

Proof: Please refer to Appendix B. ■

By substituting $\mathcal{M}_{\alpha_k}(s)$ and $\mathcal{M}_{\hat{\phi}_k}(1-s)$ ($k \in \{p_m, q_m\}$) into (15), the upper bounds of the delay violation probabilities for receivers p_m and q_m can be achieved.

From (22), the Mellin transform of the SNR-domain service process of the weak receiver leads to an infinite series. As stated in Theorem 2, the infinite series is Cauchy convergent and can be increasingly accurately approximated by its closed-form truncated version. As a matter of fact, a small number of terms, e.g., the first ten terms, are sufficient to approximate $\mathcal{M}_{\hat{\phi}_{q_m}}(1-s)$ with good accuracy. The delay QoS metric (i.e., delay violation probability) for the NOMA receivers can be efficiently evaluated with the truncated Mellin transform. This also facilitates designing the power control which takes the delay QoS of NOMA receivers into account.

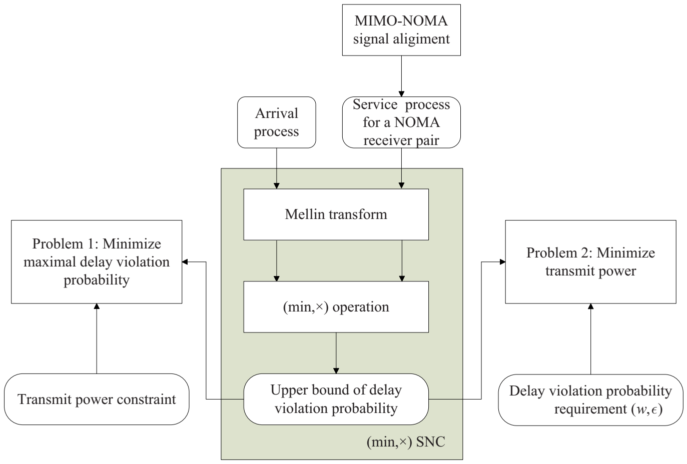

MIMO-NOMA signal aligiment
Arrival process
Service process for a NOMA receiver pair
Mellin transform
(min,×) operation
Upper bound of delay violation probability
(min,×) SNC
Problem 1: Minimize maximal delay violation probability
Transmit power constraint
Problem 2: Minimize transmit power
Delay violation probability requirement (w,ϵ)

Fig. 2. Illustration on the use of (min,×) SNC in the proposed power allocation for MIMO-NOMA with considerations on statistical delay.

IV. MIMO-NOMA Power Control Based on Probabilistic Delay Bounds

In this section, we use the asymptotic upper bound of the delay target violation probability, i.e., by substituting (21) and (22) into (16), as the delay QoS performance indicator, and optimize the power control problems of the MIMO-NOMA system under latency and reliability considerations:

- **Problem 1:** the optimal power allocation which minimizes the maximum of the delay target violation probabilities of the m-th receiver pair and achieves consistent delay and reliability within the pair, given the total transmit power ρ_m.
- **Problem 2:** the minimal required ρ_m which guarantees that the delay target violation probabilities of both receivers inviolate a prefined probability bound ϵ.

Fig. 2 shows the role of the (min,×) SNC in the MIMO-NOMA power allocation with the consideration of optimizing or guaranteeing statistical delay QoS for the NOMA receiver pairs.

A. ***Problem 1:*** *Optimal Power Allocation under Total Transmit Power Constraint*

The upper bound of the delay target violation probability is given by

$$\begin{aligned}\hat{B}_k(w) &= \inf_{s>0}\left\{\hat{\mathcal{K}}_k(s, -w)\right\} \\ &= \inf_{s>0}\left\{\frac{\mathcal{M}_{\hat{\phi}_k}^w(1-s)}{1 - \mathcal{M}_{\alpha_k}(1+s)\mathcal{M}_{\hat{\phi}_k}(1-s)}\right\}.\end{aligned} \tag{23}$$

It is clear that the larger $\eta_{p_m}^2$ is, the lower $\hat{B}_{p_m}(w)$ and the higher $\hat{B}_{q_m}(w)$. Hence, given the total transmit power ρ_m for the m-th receiver pair, it is impossible to minimize $\hat{B}_{p_m}(w)$ and $\hat{B}_{q_m}(w)$ at the same time. In order to strike a balance between the delay QoS of both receivers, i.e., to guarantee receiver fairness, the following min-max problem is formulated.

$$\begin{aligned} &\min_{\eta_{p_m},\eta_{q_m}} \max\left\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\right\} \\ &\text{s.t.} \quad \eta_{p_m}^2 + \eta_{q_m}^2 = 1. \end{aligned} \tag{24}$$

Before solving the problem, we put forth the following two lemmas which assert the monotonicity and identify the sufficient and necessary condition for the optimal solution to (24).

Lemma 1. *The upper bound of the delay target violation probability $\hat{B}_k(w)$ is continuous and monotonically decreasing with the receiver pair's total transmit power ρ_m and its power allocation cofficient η_k, $\forall k \in \{p_m, q_m\}$, when the stability condition holds.*

Proof: Please refer to Appendix C. ■

Lemma 1 shows that the upper bound of the delay violation probability for a receiver is continuous and monotonically decreasing with the total transmit power for the receiver pair and its power allocation cofficient. The continuity and monotonicity help derive the sufficient and necessary condition of the optimal solution for (24), as will be discussed in Lemma 2. When the pair of NOMA receivers have substantially different path losses, the sufficient and necessary condition further reveal the optimal power allocation coefficients are proportional to the path losses of the receivers, as will be revealed in Lemma 3.

Lemma 2. *The sufficient and necessary condition of the optimal solution for* (24) *is given by*

$$\hat{B}_{p_m}(w) = \hat{B}_{q_m}(w). \tag{25}$$

Proof: Please refer to Appendix D. ■

As a result of Lemma 1, $\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)}$ is continuous and monotonically decreases from ∞ to 0, as $\eta_{p_m}^2$ increases from 0 to 1. We can take a bisection search to solve (24) for the optimal power allocation. Hence, the value of $\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)}$ can be used as the metric for interval determination in the bisection search. According to Lemmas 1 and 2, $\eta_{p_m}^2$ is in the left interval if $\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)} < 1$, or in the right interval, otherwise. The details of the bisection search based optimal power allocation are summarized in Algorithm 1.

In Algorithm 1, a pair of power allocation coefficients are identified to satisfy $\left|\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)} - 1\right| \leq \delta_1$, where δ_1 is a predefined relative precision of the algorithm. Given the continuity of $\hat{B}_{p_m}(w)$ and $\hat{B}_{q_m}(w)$ over the power allocation coefficients η_{p_m} and η_{q_m}, there exists $\Lambda_{\delta_1} > 0$ such that $\left|\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)} - 1\right| \leq \delta_1$, $\forall \eta_{p_m} \in \{x | x \geq 0, |x - \eta_{p_m}^*| \leq \Lambda_{\delta_1}\}$. $\eta_{p_m}^*$ and $\eta_{q_m}^* = \sqrt{1 - (\eta_{p_m}^*)^2}$ denote the optimal power allocation coefficients of the m-th receiver pair. It is obvious that Λ_{δ_1} increases with δ_1. According to the continuity of $\left|\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)} - 1\right|$ over η_{p_m} and η_{q_m}, the supremum of Λ_{δ_1}, denoted by $\Lambda_{\delta_1}^{\text{alg1}}$, exists and is unique. Since the bisection search halves the search region each iteration, the total number of iterations is less than $\log_2(1/\Lambda_{\delta_1}^{\text{alg1}})$ in Algorithm 1.

Algorithm 1 Bisection search based power allocation for the m-th receiver pair in MIMO-NOMA

```
Require: $\eta_{p_m}^2 = 0$, $\eta_{q_m}^2 = 1$, interval lower bound $I_l = 0$,
    upper bound $I_u = 1$, tolerance $\delta_1$, targeted delay $w$, path
    loss $l_{p_m}^\beta$ and $l_{q_m}^\beta$
Ensure: Determine the optimal power allocation coefficients
    which satisfy $\hat{B}_{p_m}(w) = \hat{B}_{q_m}(w)$
compute $\hat{B}_{p_m}(w)$ and $\hat{B}_{q_m}(w)$ according to (17), (21),
    (22) and (23)
while $\left|\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)} - 1\right| > \delta_1$ do
    $\eta_{p_m}^2 = (I_l + I_u)/2$, $\eta_{q_m}^2 = 1 - \eta_{p_m}^2$
    Update $\hat{B}_{p_m}(w)$ and $\hat{B}_{q_m}(w)$ according to (17), (21),
    (22) and (23)
    if $\frac{\hat{B}_{p_m}(w)}{\hat{B}_{q_m}(w)} > 1$ then
        $I_l = (I_l + I_u)/2$
    else
        $I_u = (I_l + I_u)/2$
    end if
end while
return $\eta_{p_m}^2$ and $\eta_{q_m}^2$
```

Lemma 3. *When $l_{p_m}^\beta \gg l_{q_m}^\beta$ or $\bar{\gamma}_{q_m} \ll 1$, i.e. the difference of the large scale fadings between receivers p_m and q_m is significant or the SNR of receiver q_m is very small, then, under the same delay target w, the optimal power allocation given by Lemma 2 can be approximated by*

$$\eta_{p_m}^* \approx \left[1 + \left(\frac{l_{q_m}}{l_{p_m}}\right)^\beta\right]^{-\frac{1}{2}}, \quad \eta_{q_m}^* \approx \left[1 - \frac{1}{1 + \left(\frac{l_{q_m}}{l_{p_m}}\right)^\beta}\right]^{\frac{1}{2}}. \tag{26}$$

Proof: Please refer to Appendix E. ■

Although (26) is derived under the conditions that $l_{p_m}^\beta \gg l_{q_m}^\beta$ or $\bar{\gamma}_{q_m} \ll 1$, we will show via extensive simulations that (26) is accurate even if the conditions do not hold.

B. ***Problem 2:*** *Total Transmit Power Minimization under Delay Target Violation Probability Constraint*

We proceed to put a constraint ϵ on the delay target violation probability, by letting $\hat{B}_k(w) < \epsilon$ ($k \in \{p_m, q_m\}$) to ensure that $p_k(w)$ does not exceed ϵ. Since $\hat{B}_k(w)$ monotonically decreases with ρ_m (based on Lemma 1), there exists a minimal required value for ρ_m, which can be obtained by solving the following optimization problem

$$\begin{aligned} &\min_{\eta_{p_m},\eta_{q_m}} \rho_m \\ &\text{s.t.} \quad \max\left\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\right\} \leq \epsilon \\ &\qquad \eta_{p_m}^2 + \eta_{q_m}^2 = 1. \end{aligned} \tag{27}$$

Assume the optimal power allocation coefficients of (27) are $\eta^*_{p_m}$ and $\eta^*_{q_m}$. The corresponding upper bounds of the delay target violation probabilities are $\hat{B}^*_{p_m}(w)$ and $\hat{B}^*_{q_m}(w)$, respectively. We show that, when the minimal total transmit power ρ^*_m of the m-th receiver pair is attained, $\hat{B}^*_{p_m}(w) = \hat{B}^*_{q_m}(w)$ holds. Otherwise, according to Lemma 1, there would exist a new pair of power allocation coefficients $\eta^{**}_{p_m}$ and $\eta^{**}_{q_m}$ such that $\hat{B}^{**}_{p_m}(w) = \hat{B}^{**}_{q_m}(w) < \max\{\hat{B}^*_{p_m}(w), \hat{B}^*_{q_m}(w)\}$. This would allow for further power reduction without violating the first constraint condition in (27). This would contradict the minimality of ρ^*_m. Since $\hat{B}_k(w)$ monotonically decreases with ρ_m, we can use a bisection search to find the minimal total transmit power which guarantees the delay target violation probability bound for both receivers. The details of the bisection search process is summarized in Algorithm 2, where the function POWER_ALLOC in line 4 is the optimal power allocation specified in Algorithm 1.

Algorithm 2 Bisection search based total transmit power minimization for the m-th receiver pair in MIMO-NOMA

```
Require: Total power lower bound $\rho_l = 0$, upper bound $\rho_u = \rho_{\max}$, tolerance $\delta_2$, delay target $w$, delay target violation probability bound $\epsilon$, path loss $l^\beta_{p_m}$ and $l^\beta_{q_m}$
Ensure: Determine the minimal required total transmit power which satisfies $\hat{B}_{p_m}(w) = \hat{B}_{q_m}(w) \le \epsilon$
 1: $\epsilon' = 0$
 2: while $\left|\frac{\epsilon'}{\epsilon} - 1\right| > \delta_2$ do
 3:     $\rho_m = (\rho_l + \rho_u)/2$
 4:     $(\eta^2_{p_m}, \eta^2_{q_m}) =$ POWER_ALLOC$(\rho_m, w, l^\beta_{p_m}, l^\beta_{q_m})$
 5:     Calculate the corresponding delay target violation probability $\epsilon' = \hat{B}_{p_m}(w) = \hat{B}_{q_m}(w)$
 6:     if $\frac{\epsilon'}{\epsilon} > 1$ then
 7:         $\rho_l = (\rho_l + \rho_u)/2$
 8:     else
 9:         $\rho_u = (\rho_l + \rho_u)/2$
10:     end if
11: end while
12: return $\rho_m$
```

Algorithm 2 searches for ρ_m that $|\frac{\epsilon'}{\epsilon} - 1| \le \delta_2$, where $\epsilon' = \hat{B}_{p_m}(w) = \hat{B}_{q_m}(w)$, and δ_2 is a predefined relative precision of the algorithm. Let ρ^*_m denote the minimal required total transmit power in Problem 2. Recall that $\hat{B}_{p_m}(w)$ and $\hat{B}_{p_m}(w)$ are also functions of ρ_m. Given the continuity of $\hat{B}_{p_m}(w)$ and $\hat{B}_{q_m}(w)$ over ρ_m, there exists $\Lambda_{\delta_2} > 0$ such that $\left|\frac{\epsilon'}{\epsilon} - 1\right| \le \delta_2$, $\forall \rho_m \in \{x | x \ge 0, |x - \rho^*_m| \le \Lambda_{\delta_2}\}$. According to the continuity of $|\frac{\epsilon'}{\epsilon} - 1|$ over ρ_m, the supremum of Λ_{δ_2}, denoted by $\Lambda^{\text{alg2}}_{\delta_2}$, exists and is unique. The total number of iterations is less than $\log_2(\rho_{\max}/\Lambda^{\text{alg2}}_{\delta_2})$ in search of ρ_m in Algorithm 2. Since Algorithm 1 is nested in each iteration of Algorithm 2, the total complexity of Algorithm 2 for solving Problem 2 is $\mathcal{O}(\log_2(1/\Lambda^{\text{alg1}}_{\delta_1})\log_2(\rho_{\max}/\Lambda^{\text{alg2}}_{\delta_2}))$.

Lemma 3 can be exploited to efficiently implement the intra-pair power allocation; i.e., using (26), instead of Algorithm 1, to calculate η_{p_m} and η_{q_m} in each iteration of Algorithm 2. As a result, the complexity of each iteration can be reduced to $\mathcal{O}(1)$ in Algorithm 2 by eliminating the need for bisection search in Algorithm 1. The total complexity of Algorithm 2 can be reduced to $\mathcal{O}(\log_2(\rho_{\max}/\Lambda^{\text{alg2}}_{\delta_2}))$.

C. Fast-Track Power Allocation for Mobile or Nomadic Receivers

When the distance from of a receiver pair to the BS changes due to the movement of the receivers, the optimal power allcoation changes accordingly. The following lemma describes the relationship between the minimal required total transmit powers before and after the receivers change their locations.

Lemma 4. *Under the same delay target violation probability bound ϵ, assume that ρ_m and $\hat{\rho}_m$ are the minimal required total transmit powers for two different receiver-to-BS distance pairs $\mathbf{l}_m = [l_{p_m}, l_{q_m}]$ and $\hat{\mathbf{l}}_m = [\hat{l}_{p_m}, \hat{l}_{q_m}]$, respectively. When $l^\beta_{p_m} \gg l^\beta_{q_m}$ or $\bar{\gamma}_{q_m} \ll 1$, we have*

$$\frac{\rho_m}{\hat{\rho}_m} = \frac{l^\beta_{p_m} + l^\beta_{q_m}}{\hat{l}^\beta_{p_m} + \hat{l}^\beta_{q_m}} \tag{28}$$

i.e. ρ_m is in direct proportion to $l^\beta_{p_m} + l^\beta_{q_m}$.

Proof: Please refer to Appendix F. ■

Lemma 4 reveals that, if the receiver-to-BS distance pair changes from $\mathbf{l}$ to $\hat{\mathbf{l}}$ while λ, w and ϵ remain unchanged, the new optimal total transmit power $\hat{\rho}_m$ can be derived directly from the previous optimal total transmit power ρ_m according to $\hat{\rho}_m = \rho_m(\hat{l}^\beta_{p_m} + \hat{l}^\beta_{q_m})/(l^\beta_{p_m} + l^\beta_{q_m})$. As a result, the optimal total transmit power can be efficiently updated based on the large scale fadings of the receiver pairs, as opposed to re-performing the bisection search in Algorithm 2. This contributes to a significant computational complexity reduction of the power control. Algorithm 3 summarizes the fast-track power allocation for the mobile device pair m.

It is revealed in Lemma 4 that, given traffic arrival rate λ, the minimal power allocated for a pair of receivers to guarantee the delay target w with the violation probability ϵ is proportional to the sum of the path losses of the pair of receivers, when the channel difference between the strong and weak receivers is large. Sophisticated user pairing would not help further save the transmit power, and can be greatly simplified.

D. Extension to Inter-Pair Power Allocation

The proposed intra-pair power allocation described in Lemma 3 can be extended to inter-pair power allocation which allocates the finite transmit power of the BS for all the receiver pairs to minimize the maximal delay violation probability of all receivers. The MIMO-NOMA precoding/decoding scheme designed in [5] is adopted in this paper to extend our analysis and power control algorithms to multiple pairs of NOMA receivers, and hence helps the generalization of our analysis and algorithms. We confirm that the inter-pair power allocation is optimal if and only if all receivers have the same upper bound of delay violation probability. This can be proved in the same way as Lemma 2: one can always reduce the maximal delay violation probability by transfering part of the transmit power from the pair of receivers with the

Algorithm 3 Fast-track power allocation for mobile or nomadic receiver pair m

```
Require: Previous path loss $\mathbf{l} = [l_{p_m}^{\beta}, l_{q_m}^{\beta}]$, current pathloss $\hat{\mathbf{l}} = [\hat{l}_{p_m}^{\beta}, \hat{l}_{q_m}^{\beta}]$, previous requirement on latency and reliability $w$ and $\epsilon$, current requirement on latency and reliability $\hat{w}$ and $\hat{\epsilon}$, previous minimal required total transmit power $\rho_m$.
Ensure: Determine the minimal required total transmit power $\hat{\rho}_m$ which satisfies $\hat{B}_{p_m}(\hat{w}) = \hat{B}_{q_m}(\hat{w}) \le \hat{\epsilon}$
if ($w = \hat{w}$ & $\epsilon = \hat{\epsilon}$) then
    Update the total transmit power by $\hat{\rho}_m = \rho_m(\hat{l}_{p_m}^{\beta} + \hat{l}_{q_m}^{\beta})/(l_{p_m}^{\beta} + l_{q_m}^{\beta})$
else
    Run Algorithm 2 with $\hat{\mathbf{l}} = [\hat{l}_{p_m}^{\beta}, \hat{l}_{q_m}^{\beta}]$ to obtain $\hat{\rho}_m$
end if
Run Algorithm 1 or perform (26) with $\hat{\mathbf{l}} = [\hat{l}_{p_m}^{\beta}, \hat{l}_{q_m}^{\beta}]$ to obtain $\eta_{p_m}^2$ and $\eta_{q_m}^2$
return $\hat{\rho}_m$, $\eta_{p_m}^2$ and $\eta_{q_m}^2$
```

lowest delay violation probability to the pair with the highest. Consider the typical user pairing of MIMO-NOMA, where the channel difference between a selected pair of receivers is large. According to Lemma 4, the transmit power allocated for a receiver pair is proportional to the sum of the receivers' path losses to achieve the consistent delay violation probability among different receiver pairs. The transmit power allocated for the m-th pair is given by

$$\rho_m = \frac{l_{p_m}^{\beta} + l_{q_m}^{\beta}}{\sum_{i=1}^{M} l_{p_i}^{\beta} + l_{q_i}^{\beta}} \rho, \tag{29}$$

where $\rho = \sum_{i=1}^{M} \rho_i$ is the total transmit power of the BS. Once ρ_m is determined, the power allocation within each receiver pair can be achieved by conducting the proposed intra-pair power allocation, as described in Lemma 3.

V. Simulation Result

We present the numerical results of the proposed power control algorithms under different statistical QoS requirements and arrival rates in this section. The simulation assumes a homogenous statistical QoS provisioning for all MIMO-NOMA receivers. The time slot duration is set to be 1 ms, and the number of resource elements shared by each receiver pair is $W = 168$. The number of antennas at the BS and receivers are set to be $M = N = 4$, unless otherwise stated. The noise power is $\sigma^2 =$ -30 dBm, and the path loss exponent is $\beta = 3$. Before the performance of the proposed power control algorithms are presented, we first validate the effectiveness of the upper bound of the delay target violation probability, i.e., based on (15) and (22), in comparison to by Monte-Carlo simulations.

Fig. 3 compares the delay violation probability and its upper bound computed by (16), under different arrival rates ($\lambda = 15$ kbps or 3 kbps). We observe that the actual delay violation probability curve (by Monte-Carlo simulation) and the upper bound curve (by numerical calculation) have almost the same slope. This indicates that the upper bound can reasonably track the trend of the actual delay violation probability. We point out that under the same delay violation probability, the gaps between the corresponding delays of the simulation curve and upper bound curve are less than 1 ms in most cases. Although the gap is around 3 time slots for the weak receiver under large arrival rates, the upper bound manages to track the trend of the actual delay violation probability. This property lays the foundation of the power control algorithm based on the upper bound of the delay violation probability. It also endows the proposed power control algorithm certain robustness, since the upper bound of the delay violation probability exerts a guard interval of one time slot.

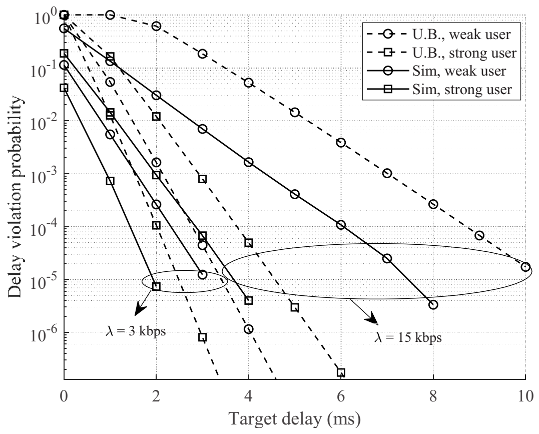


Fig. 3. Upper bound of delay target violation probability versus the delay target for the m-th receiver pair, compared to simulations under different arrival rates $\lambda = 15$ kbps and 3 kbps, with $l_{p_m} = 10$ m, $l_{q_m} = 20$ m, $\rho_m =$ 5 dBm, $\eta_{p_m}^2 = 0.25$.

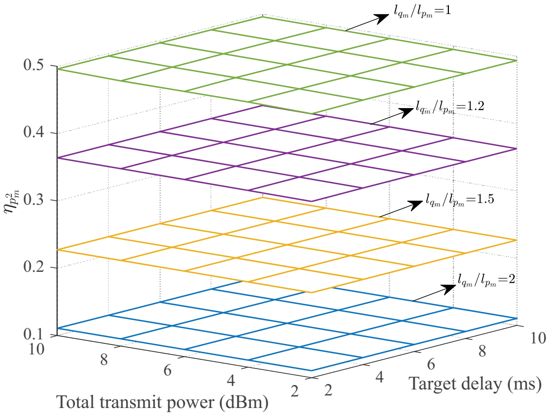


Fig. 4. Optimal power allocation coefficient obtained for receiver m versus total transmit power ρ_m and delay target w, under different distance pairs $[l_{p_m}, l_{q_m}] =$ [10 m, 10 m], [10 m, 12 m], [10 m, 15 m] and [10 m, 20 m], with arrival rate $\lambda = 10$ kbps.

We proceed to present the performance of the proposed power allocation algorithms. Fig. 4 depicts the optimal power allocation coefficients $\eta_{p_m}^2$ under different receiver-to-BS distance pairs, different total transmit powers, and different targeted delays. All the optimal power allocation coefficients are obtained by conducting the bisection search in Algorithm 1. The ratio of the path loss, $l_{q_m}^{\beta}/l_{p_m}^{\beta}$, varies from 1 to 8, and

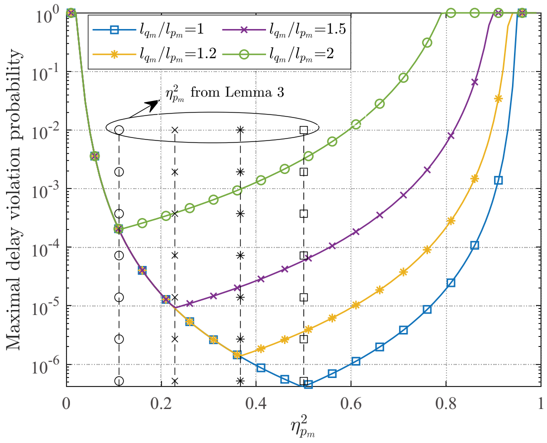


Fig. 5. Maximal delay target violation probability $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\}$ versus $\eta^2_{p_m}$ for the m-th receiver pair under different distance pairs $[l_{p_m}, l_{q_m}]$ = [10 m, 10 m], [10 m, 12 m], [10 m, 15 m] and [10 m, 20 m], with delay target $w = 3$ ms, arrival rate $\lambda = 5$ kbps, total transmit power $\rho_m = 5$ dBm.

the SNR $\bar{\gamma}_{q_m}$ varies from -7 dB to 10 dB. It can be seen that the optimal power allocation coefficient $\eta^2_{p_m}$ almost remains unchanged with ρ_m and w, i.e. $\eta^2_{p_m}$ is sololy dependent on by l_{q_m}/l_{p_m}. We see that under given l_{p_m}/l_{q_m}, the optimal value of $\eta^2_{p_m}$ that minimizes the maximal delay target violation probability can be accurately predicted by $[1+(l_{q_m}/l_{p_m})^{\beta}]^{-1}$. This indicates that Lemma 3 holds not only holds under the condition $l^{\beta}_{p_m} \gg l^{\beta}_{q_m}$ or $\bar{\gamma}_{q_m} \ll 1$, but also under various values of l_{q_m}/l_{p_m} and $\bar{\gamma}_{q_m}$. The application condition of Lemma 3 can be substantially relaxed.

Fig. 5 shows the maximal delay target violation probability $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\}$ under different power allocation coefficients and different receiver-to-BS distance pairs. It is obvious that $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\}$ first decreases with $\eta^2_{p_m}$, since under small $\eta^2_{p_m}$, the maximal delay target violation probability depends on $\hat{B}_{p_m}(w)$ which monotonically decreases with $\eta^2_{p_m}$, as proved in Lemma 1. When $\eta^2_{p_m}$ exceeds a certain value, $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\}$ begins to increase with $\eta^2_{p_m}$, because under large $\eta^2_{p_m}$, the maximal delay target violation probability is determined by $\hat{B}_{q_m}(w)$, which is monotonically increasing with $\eta^2_{p_m}$. Since the distance from the BS to the strong receiver does not change, the descending branches partially overlap with each other. We further mark the power allocation coefficients $(\eta^*_{p_m})^2$ obtained from Lemma 3 in Fig. 5. It can be seen that these values of $(\eta^*_{p_m})^2$ are exactly the ponits where the minimums of $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\}$ are obtained. This again verifies the effectiveness of Lemma 3.

Fig. 6 compares the maximal delay target violation probability between three different resource allocation schemes, namely, (a) MIMO-NOMA with the proposed power allocation scheme presented in Lemma 3, (b) MIMO-NOMA with fixed power allocation [30], where $\eta^2_{p_m} = \frac{1}{3}$ and $\eta^2_{q_m} = \frac{2}{3}$, and (c) MIMO-TDMA where each receiver in each receiver pair occupies half of the time resource. We can see that MIMO-NOMA with the proposed power allocation scheme outperforms the other two schmes under all receiver-to-BS distacne raitos. Specifically, when compared to the existing MIMO-NOMA [30] and MIMO-TDMA approaches with $l_{q_m}/l_{p_m} = 1.8$, the proposed MIMO-NOMA power allocation reduces the delay violation probability by 59.8% and 90.2%, respectively. We notice that as l_{q_m}/l_{p_m} increases, the maximal delay target violation probability also increases. This is because that as l_{q_m} increases, more transmit power should be allocated to the weak receiver to guarantee the identical delay QoS performance of both receivers.

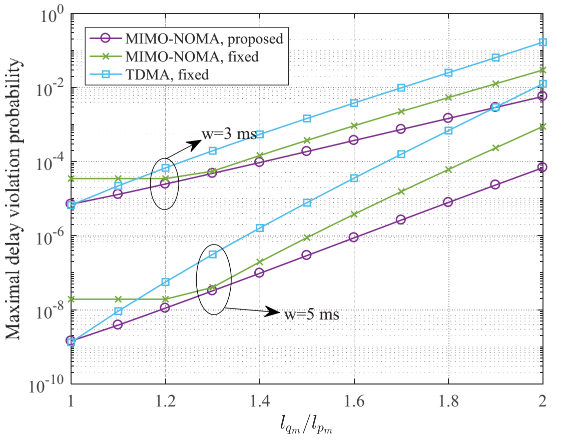


Fig. 6. Maximal delay target violation probability $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\}$ versus receiver-to-BS distance ratio l_{q_m}/l_{p_m}, with $l_{p_m} = 10$ m, delay target $w = 3$ ms and 5 ms, arrival rate $\lambda = 10$ kbps, total transmit power $\rho_m = 5$ dBm.

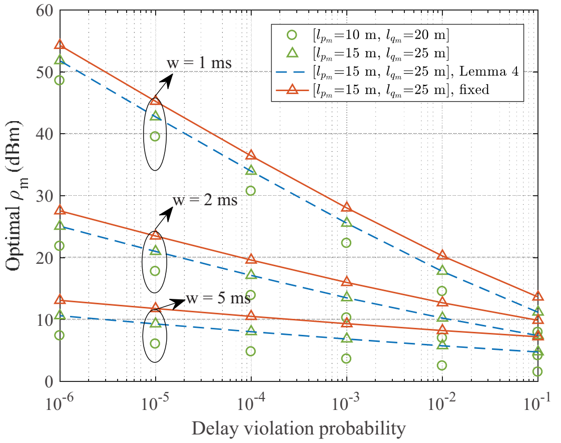


Fig. 7. Minimal total transmit power ρ_m versus delay violation probability ϵ for the m-th receiver pair under different UE-to-BS distance pairs $[l_{p_m}, l_{q_m}]$ = [10 m, 20 m] and [15 m, 25 m], with delay target $w = 1$ ms, 3 ms and 5 ms. Arrival rate $\lambda = 10$ kbps.

Fig. 7 compares the minimal total transmit power ρ_m required by the m-th receiver pair to ensure statistical delay QoS with different targeted delays and violation probabilities. In general, the case with larger receiver-to-BS distance requires higher ρ_m, since the larger distance incurs higher path loss. We also observe that both lower delay target and the violation probability, or in other words, more stringent delay QoS, can result in higher ρ_m. Given the receiver-to-BS distance pairs, a consistent gap between the minimal required ρ_m can be perceived. This is in line with Lemma 4, where the constant quotient between the minimal required total transmit powers

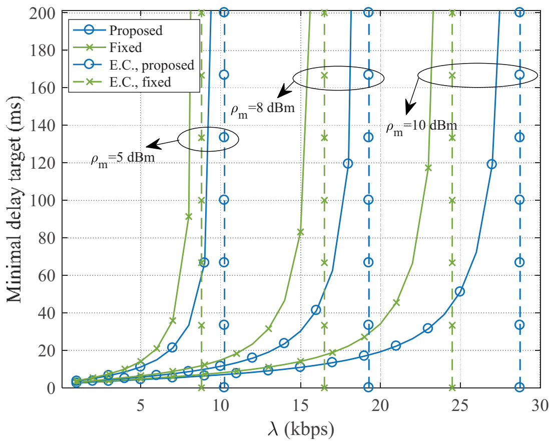


Fig. 8. Minimal delay target w that can be attained under different arrival rate λ and different total transmit power ρ_m. The UE-to-BS distance pair is $[l_{p_m}, l_{q_m}] =$ [20 m, 30 m] , delay target violation probability is $\epsilon = 10^{-5}$, which is typical for URLLC. E.C. is short for effective capacity, the dash lines represent the effective capacities of different power allocation schemes with the QoS exponent $\theta \to 0$.

depends on the ratio of the sum of the receiver pair's path losses. More specifically, we can use Lemma 4 to predict the minimal required $\hat{\rho}_m$ under $\hat{\mathbf{l}}_m =$ [15 m, 25 m] from the minimal required ρ_m under $\mathbf{l}_m =$ [10 m, 20 m]. The predicted powers are drawn in dashed curves. We find that the predicted transmit powers are almost the same with the optimal powers obtained from Algorithm 2. This confirms that the optimal total transmit power can be quickly updated, as opposed to conducting the bisection search in Algorithm 2, if the delay QoS requirements w and ϵ remain unchanged. The updating only requires the knowledge on the large scale fading of the receiver pair, and can dramatically reduce the computational complexity, as compared to Algorithm 2. We also show in Fig. 7 the minimal required total transmit power under the fixed power allocation scheme [30], where $\eta^2_{p_m} = \frac{1}{3}$ and $\eta^2_{q_m} = \frac{2}{3}$. It can be observed that the proposed power allocation scheme in Section IV-B outperforms the fixed power allocation scheme in terms of ρ_m to ensure the same statistical delay QoS requirement.

So far, we have demonstrated that the proposed power allocation scheme has better statistical delay performance than the classical fixed power allocation scheme. The premise on which we can apply the proposed bound based power allocation scheme in Algorithms 1 and 2 is that the stability condition $\mathcal{M}_{\alpha_k}(1+s)\mathcal{M}_{\phi_k}(1-s) < 1$ holds; see (19). This implies that the arrival rate λ should not exceed a certain value. In practice, one interesting question we care about is: given the statistical delay QoS constraint parameter w and ϵ and the total downlink transmit power ρ_m for the m-th receiver pair, what is the largest arrival rate that can be attained for both receivers? Here, the largest arrival rate under given QoS constraint is somewhat like the concept of effective capacity [14]. The difference is that the delay QoS constraint of effective capacity is described by the QoS exponent θ. Larger θ means more stringent delay guarantee, and vice versa. Alghouth effective capacity has an elegant mathematical expression in terms of θ and the distribution of the service process, it is hard to derive a closed-form expression for the largest tolerable arrival rate, when the delay QoS constraint is expressed in the form of delay target and violation probability. Hence, to answer the question, we show in Fig. 8 the minimal delay target w that can be attained with the constraint $\max\{\hat{B}_{p_m}(w), \hat{B}_{q_m}(w)\} \le \epsilon$, under different arrival rate λ.

When a delay target is selected, the corresponding λ in Fig. 8 is the maximal arrival that can be supported. We also plot the (w, λ) curve for the classical fixed power allocation scheme. It is obvious that the maximal tolerable arrival rate increases with w for both schemes. The proposed scheme can support a larger rate than the fixed power allocation scheme under the same delay target. When higher total transmit power is available, the maximal tolerable rate gets larger and so does the difference between the proposed scheme and the fixed scheme. We notice that there is a limit of the tolerable arrival rate when the delay target goes to infinity. The limit is exactly the effective capacity when the QoS exponent goes to zero. We show in Fig. 8 that the effective capacity follows the limit the maximal tolerable delay. The results reflected by the figure can be utilized to guide the design of system functionalities, such as traffic admission control or congestion control, whenever there is a requirement on statistical delay QoS.

VI. Conclusion

This paper investigates network layer performance bounds and cross-layer power control for downlink MIMO-NOMA in the context of URLLC. Closed-form upper bounds of the delay violation probabilities for MIMO-NOMA receivers are established, based on the (min,×) SNC and the Mellin transforms of the arrival and service processes. Based on the bounds, new algorithms are developed to achieve consistent latency and reliability within a MIMO-NOMA receiver pair, while minimizing the transmit power of the pair. It is revealed that the transmit power changes linearly with the path losses. Validated by simulations, the upper bounds of the delay violation probability and the actual probability have the same slope with a gap less than one time slot. The proposed MIMO-NOMA power allocation exhibits significant improvement over the existing MIMO-NOMA and MIMO-TDMA approaches, by reducing the delay violation probability by up to 59.8% and 90.2%, respectively.

Appendix A
Proof of Theorem 1

Since the service processes of receivers p_m and q_m have different stochastic behaviors, we characterize their Mellin transforms separately. For the strong reveiver p_m, we have $\hat{\phi}_{p_m} = e^{r_{p_m}} = \left(1 + \frac{\eta^2_{p_m}\bar{\gamma}_{p_m} z}{2}\right)^{\mathcal{W}}$, where $z = \frac{1}{\mathbf{c}^H_{p_m}\mathbf{c}_{p_m}}$ follows the exponential distribution with unit mean. Taking Mellin transformation on $\hat{\phi}_{p_m}$, the upper bound of $\mathcal{M}_{\phi_{p_m}}(1-$

s) can be derived as

$$\begin{aligned}\mathcal{M}_{\hat{\phi}_{p_m}}(1-s) &= \mathbb{E}\left[\left(1+\frac{\eta_{p_m}^2\bar{\gamma}_{p_m}z}{2}\right)^{-\mathcal{W}s}\right] \\ &= \int_0^\infty \left(1+\frac{\eta_{p_m}^2\bar{\gamma}_{p_m}z}{2}\right)^{-\mathcal{W}s} e^{-z}dz \\ &\overset{(a)}{=} \left(\frac{\eta_{p_m}^2\bar{\gamma}_{p_m}}{2}\right)^{-\mathcal{W}s} e^{\frac{2}{\eta_{p_m}^2\bar{\gamma}_{p_m}}} \int_{\frac{2}{\eta_{p_m}^2\bar{\gamma}_{p_m}}}^\infty v^{-\mathcal{W}s}e^{-v}dv, \end{aligned} \quad (30)$$

where (a) reads from the variable substitution $v = z+\frac{2}{\eta_{p_m}^2\bar{\gamma}_{p_m}}$, and (30) directly translates to (21) with the definition of the upper incomplete Gamma function.

For the weak receiver q_m, we have $\hat{\phi}_{q_m} = e^{r_{q_m}} = \left(1+\frac{\eta_{q_m}^2\bar{\gamma}_{q_m}z}{2+\eta_m^2\bar{\gamma}_{q_m}z}\right)^{\mathcal{W}}$. By taking the Mellin transform on $\hat{\phi}_{q_m}$, we can obtain the upper bound for $\mathcal{M}_{\phi_{q_m}}(1-s)$ as

$$\begin{aligned}\mathcal{M}_{\hat{\phi}_{q_m}}(1-s) &= \mathbb{E}\left[\left(1+\frac{\eta_{q_m}^2\bar{\gamma}_{q_m}z}{2+\eta_{p_m}^2\bar{\gamma}_{q_m}z}\right)^{-\mathcal{W}s}\right] \\ &\overset{(a)}{=} \int_0^\infty \left(1+\frac{\eta_{q_m}^2}{\eta_{p_m}^2} - \frac{2\frac{\eta_{q_m}^2}{\eta_{p_m}^2}}{2+\eta_{p_m}^2\bar{\gamma}_{q_m}z}\right)^{-\mathcal{W}s} e^{-z}dz \\ &\overset{(b)}{=} \eta_{p_m}^{2\mathcal{W}s}\int_0^\infty \left(1-\frac{2\frac{\eta_{q_m}^2}{\eta_{p_m}^4\bar{\gamma}_{q_m}}\Big/\left(1+\frac{\eta_{q_m}^2}{\eta_{p_m}^2}\right)}{\frac{2}{\eta_{p_m}^2\bar{\gamma}_{q_m}}+z}\right)^{-\mathcal{W}s} e^{-z}dz \\ &\overset{(c)}{=} \eta_{p_m}^{2\mathcal{W}s}\int_0^\infty \sum_{n=0}^\infty \frac{(-\mathcal{W}s)^{\underline{n}}}{n!}\left(-\frac{\pi_m}{z+\frac{2}{\eta_{p_m}^2\bar{\gamma}_{q_m}}}\right)^n e^{-z}dz, \end{aligned} \quad (31)$$

where (a) is obtained by rewritting $\frac{\eta_{q_m}^2\bar{\gamma}_{q_m}z}{2+\eta_{p_m}^2\bar{\gamma}_{q_m}z} = \frac{\eta_{q_m}^2}{\eta_{p_m}^2} - \frac{2\eta_{q_m}^2/\eta_{p_m}^2}{2+\eta_{p_m}^2\bar{\gamma}_{q_m}z}$ and using the fact that z has the exponential distribution; (b) is due to $\eta_{p_m}^2+\eta_{q_m}^2 = 1$; and (c) is based on the general binomial theorem.

The integral in the last equality of (31) can be rewritten in the form of exponential integral $E_i(\cdot)$ using the following identity integrals [31].

$$\begin{aligned}&\int_0^\infty \frac{e^{-\mu x}}{x+b}dx = -e^{b\mu}E_i(-b\mu), \quad (|\arg b| < \pi, \Re(\mu) > 0), \\ &\int_0^\infty \frac{e^{-px}}{(x+a)^2}dx = pe^{ap}E_i(-ap)+\frac{1}{a}, \quad (p > 0, a > 0), \\ &\int_0^\infty \frac{e^{-\mu x}}{(x+b)^n}dx = \frac{1}{(n-1)!}\sum_{k=1}^{n-1}(k-1)!(-\mu)^{n-k-1}b^{-k} \\ &-\frac{(-\mu)^{n-1}}{(n-1)!}e^{b\mu}E_i(-b\mu), (n > 2, |\arg b| < \pi, \Re(\mu) > 0),\end{aligned}$$

By applying the three identity integrals to (31), (22) is obtained. This completes the proof.

Appendix B
Proof of Theorem 2

According to the last equality of (31), we have

$$f_n = \int_0^\infty \frac{(-\mathcal{W}s)^{\underline{n}}}{n!}\left(\frac{-\pi_m}{z+\frac{2}{\eta_{p_m}^2\bar{\gamma}_{q_m}}}\right)^n e^{-z}dz. \quad (32)$$

To prove the convergence of $\{F_n\}_{n\geq 0}$, we have the following inequality between f_{n+1} and f_n:

$$\begin{aligned}f_{n+1} &= -\frac{\mathcal{W}s+n}{n+1}\int_0^\infty \frac{(-\mathcal{W}s)^{\underline{n}}}{n!}\left(\frac{-\pi_m}{z+\frac{2}{\eta_{p_m}^2\bar{\gamma}_{q_m}}}\right)^{n+1} e^{-z}dz \\ &< \frac{\mathcal{W}s+n}{n+1}\frac{\pi_m}{\frac{2}{\eta_{p_m}^2\bar{\gamma}_{q_m}}}\int_0^\infty \frac{(-\mathcal{W}s)^{\underline{n}}}{n!}\left(\frac{-\pi_m}{z+\frac{2}{\eta_m^2\bar{\gamma}_{q_m}}}\right)^n e^{-z}dz \\ &= \frac{\mathcal{W}s+n}{n+1}\eta_{q_m}^2 f_n. \end{aligned} \quad (33)$$

If $\mathcal{W}s < 1$ or $n \to \infty$, (33) translates to $f_{n+1} < \eta_{q_m}^2 f_n$, i.e. $\exists K > 0$, $\forall n > K$, $f_{n+1} < \eta_{q_m}^2 f_n$ always holds. Therefore, $\forall \varepsilon > 0$, $\exists K_\varepsilon = K + \lceil \ln\frac{\varepsilon}{f_K}/\ln\eta_{q_m}^2 \rceil$ such that $\forall n \geq K_\varepsilon$, $f_n < f_K\eta_{q_m}^{2(n-K)} < \varepsilon$. Hence, the sequence $\{f_n\}_{n\geq 0}$ converges to zero. Then, $\forall i > j \geq K_\varepsilon$, $|F_i - F_j| = \eta_{p_m}^{2\mathcal{W}s}\sum_{n=j}^i f_n < \eta_{p_m}^{2\mathcal{W}s}f_{K_\varepsilon}\eta_{p_m}^{2(j-K_\varepsilon)}/\eta_{p_m}^2 < \varepsilon\eta_{q_m}^{2(j-K_\varepsilon)}/\eta_{p_m}^{2(1-\mathcal{W}s)} \to 0$. As a result, the sequence $\{F_n\}_{n\geq 0}$ is a Cauchy sequence. According to the completeness of the real numbers, the limit of $\{F_n\}_{n\geq 0}$, i.e. $\mathcal{M}_{\hat{\phi}_{q_m}}(1-s) = \lim_{n\to\infty}F_n$, exists. Due to the properties of Cauchy sequence, $\mathcal{M}_{\hat{\phi}_{q_m}}(1-s)$ can be accurately approximated by F_K if K is sufficiently large.

Appendix C
Proof of Lemma 1

Since $\hat{B}_k(w)$ is monotonically increasing with $\hat{\mathcal{K}}_k(s,-w)$ and $\hat{\mathcal{K}}_k(s,-w)$ is monotonically increasing with $\mathcal{M}_{\hat{\phi}_k}(1-s)$, the monotonicity of $\hat{B}_k(w)$ with respect to ρ_m and η_k is identical with that of $\mathcal{M}_{\hat{\phi}_k}(1-s)$. It can be readily verified that $\frac{\partial\mathcal{M}_{\hat{\phi}_k}(1-s)}{\partial\rho_m} \leq 0$ and $\frac{\partial\mathcal{M}_{\hat{\phi}_k}(1-s)}{\partial\eta_k^2} \leq 0$. Hence, the monotonicity of $\hat{B}_k(w)$ is confirmed. In order to prove the continuity of $\hat{B}_k(w)$, we first show that $\inf_s\{f(s,x)\}$ is continuous with x if $f(s,x)$ is continuous with x. According to the definition of continuity, we have, $\forall\varepsilon > 0, \forall s$, there always exists $\delta > 0$ such that $\forall x_2 \in \{x : |x_1 - x| < \delta\}$, $|f(s,x_1) - f(s,x_2)| < \varepsilon$ holds. Then, $\forall s$, we have $f(s,x_2) - \varepsilon < f(s,x_1) < f(s,x_2) + \varepsilon$, which translates to $|\inf_s\{f(s,x_1)\} - \inf_s\{f(s,x_2)\}| < \varepsilon$ by taking infimum on both sides. Hence, $\inf_s\{f(s,x)\}$ is continuous with x. Since $\hat{\mathcal{K}}_k(s,-w)$ is continuous with respect to ρ_m and η_k in the stability region, $\inf_s\{\hat{\mathcal{K}}_k(s,-w)\}$ is continuous with respect to ρ_m and η_k.

Appendix D
Proof of Lemma 2

Denote the optimal power allocation coefficients of the m-th receiver pair by $\eta_{p_m}^*$ and $\eta_{q_m}^*$. The corresponding upper bounds of the delay target violation probabilities of the strong and weak receivers are $\hat{B}_{p_m}^*(w)$ and $\hat{B}_{q_m}^*(w)$, respectively. Without loss of generality, we hypothetically assume that

$\hat{B}^*_{p_m}(w) > \hat{B}^*_{q_m}(w)$. According to Lemma 1, $\forall\varepsilon > 0$, there exists $\delta > 0$, when the following new power allocation coefficients, $\eta^{**}_{p_m} = \eta^*_{p_m} + \delta/2$ and $\eta^{**}_{q_m} = \sqrt{1 - (\eta^{**}_{p_m})^2}$, are adopted, the inequalities $\hat{B}^*_{p_m}(w) > \hat{B}^{**}_{p_m}(w) > \hat{B}^*_{p_m}(w) - \varepsilon$ and $\hat{B}^*_{q_m}(w) + \varepsilon > \hat{B}^{**}_{q_m}(w) > \hat{B}^*_{q_m}(w)$ hold. In other words, $\hat{B}^{**}_{p_m}(w)$ and $\hat{B}^{**}_{q_m}(w)$ are the upper bounds of the delay target violation probabilities under the new power allocation coefficients $\eta^{**}_{p_m}$ and $\eta^{**}_{q_m}$. For $0 < \varepsilon < \hat{B}^*_{p_m}(w) - \hat{B}^*_{q_m}(w)$, we have $\max\{\hat{B}^{**}_{p_m}(w), \hat{B}^{**}_{q_m}(w)\} < \max\{\hat{B}^*_{p_m}(w), \hat{B}^*_{q_m}(w)\}$. This contradicts the hypothesis of optimality of $\eta^*_{p_m}$ and $\eta^*_{q_m}$, and therefore concludes the proof.

Appendix E
Proof of Lemma 3

Note that, for a given w, $\hat{B}_k(w)$ ($k \in \{p_m, q_m\}$) is determined solely by $\mathcal{M}_{\hat{\phi}_k}(1-s)$, since the Poisson arrival rates are the same for the two receivers. Hence, a sufficient condition of (25) is $\mathcal{M}_{\hat{\phi}_{p_m}}(1-s) = \mathcal{M}_{\hat{\phi}_{q_m}}(1-s)$, which holds if the two receivers have identically distributed SINRs, i.e. $\forall z \geq 0$,

$$1 + \frac{\eta^2_{p_m}\bar{\gamma}_{p_m}z}{2} = 1 + \frac{\eta^2_{q_m}\bar{\gamma}_{q_m}z}{2 + \eta^2_{p_m}\bar{\gamma}_{q_m}z}, \tag{34}$$

which equals to

$$1 + \frac{1}{\frac{2}{\eta^2_{p_m}\bar{\gamma}_{p_m}z}} = 1 + \frac{1}{\frac{2}{\eta^2_{q_m}\bar{\gamma}_{q_m}z} + \frac{\eta^2_{p_m}}{\eta^2_{q_m}}}. \tag{35}$$

This is reasonable since when $l^\beta_{p_m} \gg l^\beta_{q_m}$, most of the transmit power is allocated to the weak receiver to ensure the same statistical delay performance. This leads to $\frac{\eta^2_{p_m}}{\eta^2_{q_m}} \ll 1$, which translates to $\eta^2_{p_m} \ll \eta^2_{q_m} < 1$. The ratio of the two terms in the denominator on the right-hand side (RHS) of (35) is $r_d = \frac{2}{\bar{\gamma}_{q_m}\eta^2_{p_m}z}$. Since z is exponentially distributed, the probability that r_d exceeds a large threshold T is given by $\Pr\{r_d > T\} = 1 - e^{\frac{-2}{\bar{\gamma}_{q_m}\eta^2_{p_m}T}}$. If $\bar{\gamma}_{q_m} \ll 1$ or $\eta^2_{p_m} \ll 1$, we have $\Pr\{r_d > T\} \to 1$. In other words, $\frac{\eta^2_{p_m}}{\eta^2_{q_m}}$ is negligible with high probability, as compared to $\frac{2}{\eta^2_{q_m}\bar{\gamma}_{q_m}z}$. Hence, (35) can be approximated with high accuracy by

$$1 + \frac{1}{\frac{2}{\eta^2_{p_m}\bar{\gamma}_{p_m}z}} = 1 + \frac{1}{\frac{2}{\eta^2_{q_m}\bar{\gamma}_{q_m}z}}, \tag{36}$$

from which $\eta^2_{p_m}\bar{\gamma}_{p_m} = \eta^2_{q_m}\bar{\gamma}_{q_m}$ can be obtained. Together with the constraint $\eta^2_{p_m} + \eta^2_{q_m} = 1$, we can achieve the power allocation in (26). In turn, the result verifies the condition that if $l^\beta_{p_m} \gg l^\beta_{q_m}$ then $\eta^2_{p_m} \ll 1$.

Appendix F
Proof of Lemma 4

Let $\hat{B}_k(w, \mathbf{l}_m, \rho_m)$ denote the upper bound of the delay target violation probability with the receiver-to-BS distance pair $\mathbf{l}_m$ and the total transmit power ρ_m. According to the analysis in Section IV-A, we have $\hat{B}_{p_m}(w, \mathbf{l}_m, \rho_m) = \hat{B}_{q_m}(w, \mathbf{l}_m, \rho_m) = \hat{B}_{p_m}(w, \hat{\mathbf{l}}_m, \hat{\rho}_m) = \hat{B}_{q_m}(w, \hat{\mathbf{l}}_m, \hat{\rho}_m) = \epsilon$. Since the Poisson arrival rate does not change, $\hat{B}_{p_m}(w, \mathbf{l}_m, \rho_m) = \hat{B}_{p_m}(w, \hat{\mathbf{l}}_m, \hat{\rho}_m)$ translates to

$$\mathbb{E}\left[\left(1 + \frac{\eta^2_{p_m}\rho_m z}{2\sigma^2 l^\beta_{p_m}}\right)^{-\mathcal{W}s}\right] = \mathbb{E}\left[\left(1 + \frac{\hat{\eta}^2_{p_m}\hat{\rho}_m z}{2\sigma^2 \hat{l}^\beta_{p_m}}\right)^{-\mathcal{W}s}\right] \tag{37}$$

which implies

$$\frac{\rho_m}{\hat{\rho}_m} = \frac{\hat{\eta}^2_{p_m} l^\beta_{p_m}}{\eta^2_{p_m}\hat{l}^\beta_{p_m}} \tag{38}$$

Substituting $\eta^2_{p_m} = 1/(1 + l^\beta_{q_m}/l^\beta_{p_m})$ and $\hat{\eta}^2_{p_m} = 1/(1 + \hat{l}^\beta_{q_m}/\hat{l}^\beta_{p_m})$ into (38), we can obtain (28).

References

[1] P. Schulz *et al.*, "Latency critical iot applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[2] "TR 38.913, v14.3.0," 3GPP, Tech. Rep., Jun. 2017.

[3] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom, K. F. Trillingsgaard, A. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, Mar. 2018.

[4] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, thirdquarter 2018.

[5] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[6] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, Secondquarter 2017.

[7] Z. D. P. F. Z. Ma, Z. Zhang and H. Li, "Key techniques for 5g wireless communications: Network architecture, physical layer, and MAC layer perspectives," *Science China Information Sciences*, vol. 58, no. 4, pp. 1–20, Feb. 2015.

[8] D. M. B. Mehdi and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, p. 18341853, Oct. 2018.

[9] Z. Zhang and R. Q. Hu, "Uplink non-orthogonal multiple access with fractional power control," in *Proc. 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2017, pp. 1–6.

[10] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191–5202, Oct. 2017.

[11] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.

[12] ——, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849–1858, Apr. 2017.

[13] G. Liu, Z. Ma, X. Chen, Z. Ding, F. R. Yu, and P. Fan, "Cross-layer power allocation in nonorthogonal multiple access systems for statistical QoS provisioning," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 388–11 393, Dec. 2017.

[14] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[15] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, Oct. 2018.

[16] Y. Jiang and Y. Liu, *Stochastic network calculus*. Springer, 2008, vol. 1.

[17] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, Firstquarter 2015.

[18] L. Lei, J. Lu, Y. Jiang, X. S. Shen, Y. Li, Z. Zhong, and C. Lin, "Stochastic delay analysis for train control services in next-generation high-speed railway communications system," *IEEE Intell. Transp. Syst. Mag.*, vol. 17, no. 1, pp. 48–64, Jan. 2016.

[19] G. Yang, M. Xiao, and Z. Pang, "Delay analysis of traffic dispersion with Nakagami-m fading in millimeter-wave bands," in *Proc. 2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018, pp. 1–6.

[20] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "A (min,) network calculus for multi-hop fading channels," in *2013 Proceedings IEEE INFOCOM*, Apr. 2013, pp. 1833–1841.

[21] ——, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[22] N. Petreska, H. Al-Zubaidy, and J. Gross, "Power minimization for industrial wireless networks under statistical delay constraints," in *Proc. 2014 26th International Teletraffic Congress (ITC)*, Sep. 2014, pp. 1–9.

[23] R. K. N.Petreska, H. Al-Zubaidy and J. Gross, "Bound-based power optimization for multi-hop heterogeneous wireless industrial networks under statistical delay constraints," 2017. [Online]. Available: https://arxiv.org/abs/1608.02191

[24] G. Yang, M. Xiao, H. Al-Zubaidy, Y. Huang, and J. Gross, "Analysis of millimeter-wave multi-hop networks with full-duplex buffered relays," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 576–590, Feb. 2018.

[25] 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.

[26] J. Arnau and M. Kountouris, "Delay performance of MISO wireless communications," in *Proc. 2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2018, pp. 1–8.

[27] C. S. Chang, *Performance guarantees in communication networks*. Springer, 2000.

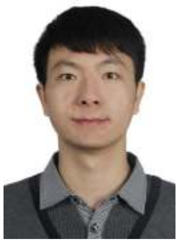
[28] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. New York, NY, USA: ACM, 2015, pp. 13–22.

[29] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. New York: Dover, 1965.

[30] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[31] I. Gradshteyn and I. Ryzhik, *Table of integrals, series, and products*, 6th ed. New York: Academic Press, 2000.

Chiyang Xiao (S'18) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University. His research interests are in the area of novel multiple access, massive MIMO, machine learning, and signal processing.



Jie Zeng (M'09–SM'16) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University in 2006 and 2009, respectively. His research interests include novel network architecture, ultra-dense networks, and novel multiple access. He has published 3 books and over 100 journal and conference papers. He holds more than 30 Chinese and 7 international patents. He obtained the science and technology award of Beijing in 2015 and the best cooperation award of Samsung Electronics in 2016.

Wei Ni (M'09-SM'15) received the B.E. and Ph.D. degrees in Electronic Engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. Currently he is a Team Leader at CSIRO, Sydney, Australia, and an adjunct professor at the University of Technology Sydney (UTS). He also holds honorary positions at the University of New South Wales (UNSW) and Macquarie University (MQ). Prior to this, he was a postdoctoral research fellow at Shanghai Jiaotong University from 2005 – 2008; Deputy Project Manager at the Bell Labs R&I Center, Alcatel/Alcatel-Lucent from 2005 – 2008; and Senior Researcher at Devices R&D, Nokia from 2008 – 2009. His research interests include stochastic optimization, game theory, graph theory, as well as their applications to network and security.

Dr Ni has been serving as Vice Chair of IEEE NSW VTS Chapter and Editor of IEEE Transactions on Wireless Communications since 2018, secretary of IEEE NSW VTS Chapter from 2015 – 2018, Track Chair for VTC-Spring 2017, Track Co-chair for IEEE VTC-Spring 2016, and Publication Chair for BodyNet 2015. He also served as Student Travel Grant Chair for WPMC 2014, a Program Committee Member of CHINACOM 2014, a TPC member of IEEE ICC'14, ICCC'15, EICE'14, and WCNC'10.

Xin Su (M'03–SM'15) received the M.S. and Ph.D. degrees in Electronic Engineering from UESTC (University of Electronic Science and Technology of China) in 1996 and 1999, respectively. Currently he is a full professor in the Research Institute of Information Technology at Tsinghua University. He is also the chairman of IMT-2020(5G) wireless technology work group in MIIT (Ministry of Industry and Information Technology of Peoples Republic of China) and vice chairman of the Innovative Wireless Technology Work Group of CCSA (China Communications Standards Association). His research interests include broadband wireless access, wireless and mobile network architecture, self-organizing network, software defined radio, and cooperative communications. Dr. Su has published over 100 papers in the core journals and important conferences, and owned more than 30 patents.

Ren Ping Liu (M'09–SM'14) received his B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, China, and the Ph.D. degree from the University of Newcastle, Australia. He is currently a Professor and Head of Discipline of Network & Cybersecurity at University of Technology Sydney. Professor Liu is also the co-founder and CTO of Ultimo Digital Technologies Pty Ltd, developing IoT and Blockchain. Prior to that he was a Principal Scientist and Research Leader at CSIRO, where he led wireless networking research activities. He specialises in system design and modelling and has delivered networking solutions to a number of government agencies and industry customers. His research interests include wireless networking, Cybersecurity, and Blockchain.

Professor Liu was the founding chair of IEEE NSW VTS Chapter and a Senior Member of IEEE. He served as Technical Program Committee chairs and Organising Committee chairs in a number of IEEE Conferences. Prof Liu was the winner of Australian Engineering Innovation Award and CSIRO Chairman medal. He has over 150 research publications, and has supervised over 30 PhD students.

Tiejun Lv (M'08–SM'12) received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1997 and 2000, respectively. From January 2001 to January 2003, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. In 2005, he was promoted to a Full Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT). From September 2008 to March 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He is the author of 2 books, more than 60 published IEEE journal papers and 170 conference papers on the physical layer of wireless mobile communications. His current research interests include signal processing, communications theory and networking. He was the recipient of the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006. He received the Nature Science Award in the Ministry of Education of China for the hierarchical cooperative communication theory and technologies in 2015.

Jing Wang (M'99) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University in 1983 and 1986, respectively. He has served on the Faculty of Tsinghua University since 1986, where he is currently a Professor with the School of Information Science and Technology. He serves as the Vice Director of the Tsinghua National Laboratory for Information Science and Technology. His research interests are in the area of wireless digital communications, including modulation, multiuser detection, and vehicle-tovehicle communication. He has published over 150 conference and journal papers.