Stable Throughput Region of Downlink NOMA Transmissions with Limited CSI

Yong Zhou and Vincent W.S. Wong
Department of Electrical and Computer Engineering
The University of British Columbia, Vancouver, Canada
e-mail: {zhou, vincentw}@ece.ubc.ca

Abstract—Non-orthogonal multiple access (NOMA) has recently been proposed as a key enabling technology for the fifth generation (5G) wireless networks. Different from the existing works which focus on the performance analysis of NOMA with backlogged traffic, in this paper, we analyze the stable throughput region of downlink NOMA transmission with dynamic traffic arrival for users with different priorities. By utilizing limited instantaneous channel state information (CSI) at the base station, we propose an opportunistic NOMA scheme to enhance the network performance. Considering both NOMA and dynamic traffic arrival leads to interacting queues, which complicate the performance analysis. By using tools from stochastic geometry and queueing theory, we decouple the interacting queues and characterize the stable throughput region of the proposed opportunistic NOMA scheme in terms of the threshold to trigger NOMA and transmission power allocation coefficients. Numerical results show that, compared to the orthogonal multiple access scheme, the proposed opportunistic NOMA scheme can significantly enhance the stable throughput region when the design parameters are appropriately selected.

I. INTRODUCTION

To meet the increasing traffic demand due to the proliferation of smart devices and data hungry applications, non-orthogonal multiple access (NOMA) has recently been proposed as a promising multiple access technique to enhance the spectrum efficiency of the fifth generation (5G) wireless networks [1], [2]. The base station using NOMA can serve multiple users simultaneously by exploiting the power domain rather than the time/frequency/code domain in orthogonal multiple access (OMA). By appropriately allocating the transmission power of the base station to multiple users with diverse channel conditions, NOMA can also achieve a balance between network throughput and user fairness.

The research on NOMA has recently received considerable attention [3]–[9]. The system-level performance of downlink NOMA transmission is evaluated in [3], which shows that user pairing and transmission power allocation are important design aspects of NOMA. The outage probabilities of NOMA with randomly deployed users and cooperation among users are analyzed in [4] and [5], respectively. The authors in [6] study the performance of NOMA with multiple-input multiple-output (MIMO) for both downlink and uplink transmission, in which signal alignment is utilized to mitigate the co-channel interference among different user pairs. The impact of user pairing on the performance of NOMA is analytically investigated in [7], which shows that NOMA achieves better performance when the paired users have more diverse channel conditions. The applications of NOMA in Internet of Things and cognitive radio networks are studied in [8] and [9], respectively. However, all the aforementioned studies focus on the performance analysis of NOMA with backlogged traffic, which cannot be directly extended to the scenario with dynamic traffic arrival.

This work is motivated by the following three aspects. First, with dynamic traffic arrival, queue stability is an important quality of service (QoS) requirement. To guarantee the stability of a queue, NOMA cannot always be performed as its average service rate can be degraded due to the sharing of frequency channel and transmission power with other users. Second, considering dynamic traffic arrival together with NOMA complicates the performance analysis by introducing interacting queues. In particular, the service process of a queue depends on the status of other queues, which determines whether NOMA or OMA should be enabled. Third, channel state information (CSI) plays an important role in designing user pairing and transmission power allocation strategies, which have significant impact on the performance of NOMA. As full CSI is difficult to obtain in practice, the impact of limited CSI on the performance of NOMA should be investigated.

In this paper, we investigate the performance of downlink NOMA transmission with dynamic traffic arrival for all users. In such a scenario, the stable throughput region [10], [11] is an important performance metric, which is defined as the set of maximum achievable packet arrival rates given that all queues are stable. We propose an opportunistic NOMA scheme to enhance the stable throughput region, where NOMA for users with different priorities is enabled only if the channel gain between the high-priority user and the base station does not fall below a certain threshold. The main contributions of this paper are three-fold:

1) We develop a theoretical performance analysis framework for downlink NOMA transmission with dynamic traffic arrival and spatially random users. This framework provides a better understanding of the benefits and limitations of NOMA.

2) By using limited instantaneous CSI at the base station, we propose an opportunistic NOMA scheme to serve users with different priorities. We characterize the stable throughput region of the proposed opportunistic NOMA scheme by utilizing tools from stochastic geometry and queueing theory.

3) Numerical results show that the stable throughput re-
nation of opportunistic NOMA is significantly larger than that of OMA. The impact of important design parameters (e.g., threshold to trigger NOMA and transmission power allocation coefficients) on the performance of NOMA is also illustrated.

The remainder of this paper is organized as follows. We describe the network topology, queueing model, and signal reception model in Section II. Section III presents an opportunistic NOMA scheme and characterizes the corresponding stable throughput region. Numerical results are illustrated in Section IV. Finally, Section V concludes this paper.

II. SYSTEM MODEL

A. Network Topology and Queueing Model

Consider a downlink communication scenario consisting of one base station and $M + 1$ users, as shown in Fig. 1. Base station $S$ locates at the center of the circular network coverage area with radius $r$. Users are categorized into two groups with different priorities. User $D_0$ has a high priority to be served, while other users (i.e., $\{D_m, m \in M = \{1, \ldots, M\}\}$) have the same low priority. Over a single frequency channel, the time is slotted into constant durations. The locations of low-priority users are assumed to follow a binomial point process (BPP). Specifically, $M$ low-priority users at each time slot are independently and uniformly distributed within a circle centered at base station $S$ (i.e., origin) with radius $r_L < r$. On the other hand, the distance between base station $S$ and high-priority user $D_0$ is fixed and denoted as $r_0 \in (r_L, r]$. Extension to multiple high-priority users and random distances between the base station and the high-priority users is possible at the expense of complicating the derived expressions.

Base station $S$ is equipped with two queues of infinite size, denoted as $Q_H$ and $Q_L$, which store the packets to be transmitted to high-priority user $D_0$ and $M$ low-priority users, respectively, as shown in Fig. 2. The packet arrival at base station $S$ for user $D_m$ follows an independent and identically distributed (i.i.d.) Bernoulli process with an average arrival rate of $\lambda_m$ (packets/time slot). Hence, the average arrival rate of queue $Q_L$ is $\lambda_L = \sum_{m=1}^M \lambda_m$. Base station $S$ and all users have a single antenna. All packets have equal length and each packet is transmitted in one time slot. The packets of the same priority are served in a first-in first-out (FIFO) manner.

At the end of each time slot $t \in \mathbb{Z}^+$, the locations of low-priority users are changed according to a high mobility random walk model within the circle with radius $r_L$ as in [12], [13]. The channel between any two transceivers suffers from path loss and Rayleigh fading. The fading coefficients are assumed to remain invariant during one time slot and vary independently over different time slots and among different links, as in [10], [12]. Due to channel impairments and interference, a packet can be successfully decoded if the received signal-to-interference-plus-noise ratio (SINR) is not smaller than the required reception threshold. Upon successfully or erroneously receiving a packet from base station $S$, the corresponding receiver sends an acknowledgement (ACK) or negative ACK (NAK) frame via an error-free and delay-free control channel. After receiving the ACK frame, the packet is removed from the queue at base station $S$. Otherwise, base station $S$ retransmits the packet until it is successfully decoded. The protocol overhead due to ACK and NAK feedback is much smaller than the packet size and is not considered in this paper.

We denote $Q_H(t)$ and $Q_L(t)$ as the queue length of $Q_H$ and $Q_L$ at time slot $t$, respectively. A queue is said to be stable if its queue length has a limiting distribution as time goes to infinity [14]. If the arrival and service processes of a queue are jointly stationary and ergodic, by Lloyes’ theorem [15], the sufficient condition for the stability of queue $Q_H$ is that $\lambda_H < \mu_H$, where $\lambda_H = \lambda_0$ and $\mu_H$ (packets/time slot) denotes the average service rate of queue $Q_H$. The network is stable when both queues $Q_H$ and $Q_L$ are stable. The stable throughput region is defined as the set of maximum arrival rates $\{\lambda_H, \lambda_L\}$, which can stabilize the network.

B. Signal Reception Model

NOMA has the potential to enhance the spectrum efficiency by exploiting the power domain to simultaneously serve multiple users. To reduce the implementation complexity, we consider the case that two users are paired to perform NOMA. Such a two-user NOMA scheme is specified in Long Term Evolution Advanced (LTE-A) [16] and considered in [6], [7]. We pair high-priority user $D_0$ with low-priority user $D_L$, which is the intended receiver of the first packet from queue $Q_L$. When NOMA is performed to transmit the packets from both queues $Q_H$ and $Q_L$ at time slot $t$, the superpositioned signal transmitted by base station $S$ can be expressed as $\alpha_H \sqrt{P_S s_0(t)} + \alpha_L \sqrt{P_S s_L(t)}$, where $P_S$ denotes the total transmission power of base station $S$, $\alpha_H$ and $\alpha_L$ denote the power allocation coefficients for high- and low-priority users,
respectively, and \( s_k(t) \) denotes the signal intended for user \( D_k \) at time slot \( t \). Without loss of generality, \( s_k(t) \)'s are assumed to be i.i.d. Gaussian random variables with zero mean and unit variance. As \( r_m < r_0, \forall m \in M \), according to the design principle of NOMA, we have \( \alpha_{m} > \alpha_{0} \) and \( \alpha_{m}^2 + \alpha_{0}^2 = 1 \). Over the block fading channel, the signal received by user \( D_m \) at time slot \( t \) is given by

\[
y_m(t) = (\alpha_m s_0(t) + \alpha_m s_k(t)) \sqrt{P_S h_m(t)} \sqrt{\ell(x_m) + n_m(t)},
\]

where \( h_m(t) \) denotes the Rayleigh fading channel gain between base station \( S \) and user \( D_m \) with zero mean and unit variance at time slot \( t \). \( n_m(t) \) denotes the additive white Gaussian noise at user \( D_m \) with zero mean and variance \( \sigma^2 \) at time slot \( t \). \( \ell(x_m) = (1 + r_m^2)^{-1} \) denotes the non-singular path loss between base station \( S \) and user \( D_m \). \( x_m \) denotes the location coordinate of user \( D_m \), and \( \beta \) denotes the path loss exponent.

After receiving the signal from base station \( S \), user \( D_0 \) treats the signal intended for user \( D_k \) as co-channel interference and decodes its own signal based on the SINR given by

\[
\Gamma_{0|L}(t, \alpha_H) = \frac{\alpha_H^2 P_S |h_0(t)|^2 \ell(x_0)}{\alpha_L^2 P_S |h_0(t)|^2 \ell(x_0) + \sigma^2},
\]

where \( \Gamma_{0|L}(t, \alpha_H) \) denotes the SINR of signal \( s_0(t) \) observed by user \( D_0 \) when pairing with a low-priority user at time slot \( t \).

On the other hand, user \( D_k \) first tries to decode the signal intended for user \( D_0 \) with the SINR given by

\[
\Gamma_{0\rightarrow k}(t, \alpha_H) = \frac{\alpha_H^2 P_S |h_k(t)|^2 \ell(x_k)}{\alpha_L^2 P_S |h_k(t)|^2 \ell(x_k) + \sigma^2},
\]

where \( \Gamma_{0\rightarrow k}(t, \alpha_H) \) denotes the SINR of signal \( s_0(t) \) observed by user \( D_k \) at time slot \( t \).

Let \( \Gamma_{th} \) denote the threshold for successful packet reception. If user \( D_k \) successfully decodes signal \( s_0(t) \), i.e., \( \Gamma_{0\rightarrow k}(t, \alpha_H) \geq \Gamma_{th} \), user \( D_k \) removes signal \( s_0(t) \) from received signal \( y_k(t) \) by applying successive interference cancellation (SIC), and then decodes its own signal with the signal-to-noise ratio (SNR) given by

\[
\Gamma_k(t, \alpha_L) = \frac{\alpha_L^2 P_S |h_k(t)|^2 \ell(x_k)}{\sigma^2},
\]

where \( \Gamma_k(t, \alpha_L) \) denotes the SNR of signal \( s_k(t) \) observed by user \( D_k \) at time slot \( t \).

Based on the above discussions, by using NOMA, users \( D_0 \) and \( D_k \) can successfully decode their own signals if the events \( \{ \Gamma_{0|L}(t, \alpha_H) \geq \Gamma_{th} \} \) and \( \{ \Gamma_{0\rightarrow k}(t, \alpha_H) \geq \Gamma_{th} \cap \Gamma_k(t, \alpha_L) \geq \Gamma_{th} \} \) occur, respectively. On the other hand, by using OMA (e.g., time division multiple access (TDMA)), user \( D_k \) can successfully decode its own signal if event \( \{ \Gamma_k(t, 1) \geq \Gamma_{th} \} \) occurs. By using NOMA, base station \( S \) can serve users \( D_0 \) and \( D_k \) simultaneously, at the cost of reducing the probability of successful packet reception at user \( D_0 \). Specifically, by sharing the frequency channel and splitting the transmission power, the received SINR at user \( D_0 \) decreases, i.e., \( \Gamma_{0|L}(t, \alpha_H) < \Gamma_0(t, 1) = P_S |h_0(t)|^2 \ell(x_0) / \sigma^2 \).

As a result, to guarantee the stability of queue \( Q_H \), NOMA cannot always be enabled, especially when the average arrival rate \( \lambda_H \) is large.

### III. STABLE THROUGHPUT REGION

In this section, we present an opportunistic NOMA scheme by utilizing limited instantaneous CSI at base station \( S \) and a baseline OMA scheme, and derive their stable throughput regions.

#### A. Opportunistic NOMA

We consider that limited instantaneous CSI is available at base station \( S \). First, when queue \( Q_H \) is non-empty at time slot \( t \), one-bit information is fed back from user \( D_0 \) to base station \( S \). In particular, user \( D_0 \) feeds back 1 to base station \( S \) if the instantaneous channel gain, \( |h_0(t)|^2 \ell(x_0) \), is not less than a threshold, \( \theta \), and feeds back 0 to base station \( S \) otherwise. Second, when queue \( Q_H \) is empty at time slot \( t \), the intended receivers of the first two packets from queue \( Q_L \) feed back their distance information to base station \( S \). Based on limited instantaneous CSI, NOMA can be opportunistically enabled by base station \( S \) to enhance the stable throughput region.

The opportunistic NOMA system, denoted as \( \Phi^{ON} \), is described as follows. As user \( D_0 \) has a high priority to be served, base station \( S \) transmits a packet from queue \( Q_H \) whenever it is non-empty. Without loss of generality, the intended receiver of the second packet from queue \( Q_L \), at time slot \( t \), when available, is denoted as user \( D_k \). Depending on the status of queues \( Q_H \) and \( Q_L \) at time slot \( t \), the packet transmissions in opportunistic NOMA system \( \Phi^{ON} \) can be categorized into the following three cases:

**Case 1:** If \( Q_H(t) > 0 \) and \( Q_L(t) > 0 \), then base station \( S \) transmits the first packet from queue \( Q_H \) and the first packet from queue \( Q_L \) to users \( D_0 \) and \( D_k \), respectively, using NOMA with fixed power allocation coefficients \( (\alpha_H^2, \alpha_L^2) \) when \( |h_0(t)|^2 \ell(x_0) \geq \theta \), and transmits the first packet from queue \( Q_H \) to user \( D_0 \) using OMA with power \( P_S \) when \( |h_0(t)|^2 \ell(x_0) < \theta \).

**Case 2:** If \( Q_H(t) > 0 \) and \( Q_L(t) = 0 \), then base station \( S \) transmits the first packet from queue \( Q_H \) to user \( D_0 \) using OMA with power \( P_S \).

**Case 3:** If \( Q_H(t) = 0 \) and \( Q_L(t) > 0 \), then base station \( S \) transmits the first and second packets from queue \( Q_L \) to users \( D_k \) and \( D_k \), respectively, using NOMA when the first two packets are intended for different users (i.e., \( D_k \neq D_k \)), and transmits the first packet from queue \( Q_L \) to user \( D_k \) using OMA with power \( P_S \) when \( D_k = D_k \) or \( Q_L(t) = 1 \).

Based on the opportunistic NOMA system described above, the average service rate of queue \( Q_H \) depends on the status of queue \( Q_L \). In particular, when queue \( Q_L \) is empty, base station \( S \) transmits a packet from queue \( Q_H \) to user \( D_0 \) using OMA. On the other hand, when queue \( Q_L \) is non-empty, base station \( S \) transmits the first packet from queue \( Q_H \) and the first packet from queue \( Q_L \) to users \( D_0 \) and \( D_k \) using NOMA with probability \( \Psi \left( |h_0(t)|^2 \ell(x_0) \geq \theta \right) = \exp \left( -\theta (1 + r_0^2) \right) \). Note that the probabilities of successful packet reception at user \( D_0 \)
using OMA and NOMA are different. Similarly, the average service rate of queue $Q_L$ also depends on the status of queue $Q_H$. As a result, queues $Q_H$ and $Q_L$ are interacting with each other and their average service rates cannot be directly calculated. Stochastic dominance [14] can be used to decouple the interacting queues and to facilitate the characterization of the stable throughput region.

By using stochastic dominance, we construct two dominant systems $\Phi_1^{ON}$ and $\Phi_2^{ON}$ based on the original opportunistic NOMA system $\Phi^{ON}$. The dominant systems, as a modification of the original system (i.e., $\Phi^{ON}$), ensure that their queue lengths are always not less than those in the original system by enabling the empty queues to transmit dummy packets. The transmission of dummy packets reduces the probability of successful packet reception by generating co-channel interference, but does not contribute to the throughput. Hence, the stability condition of dominant systems is sufficient for the stability of the original system. The stable throughput regions of the constructed two dominant systems are discussed as follows.

1) Stable throughput region in dominant system $\Phi^{ON}_1$. In dominant system $\Phi^{ON}_1$, if queue $Q_L$ is empty, then queue $Q_L$ contributes a dummy packet when user $D_0$ feeds back 1 to base station $S$, while queue $Q_H$ acts the same as in the original system $\Phi^{ON}$.

In this case, the service process of queue $Q_H$ can be divided into two cases: a) base station $S$ transmits one packet to user $D_0$ using OMA when $|h_0(t)|^2\ell(x_0) < \theta$; b) base station $S$ transmits one packet to user $D_0$ using NOMA when $|h_0(t)|^2\ell(x_0) \geq \theta$. As a result, the average service rate of queue $Q_H$ in dominant system $\Phi^{ON}_1$, denoted as $\mu^{ON}_H$, can be expressed as

$$
\mu^{ON}_H = P\left(\Gamma_0(t, 1) \geq \Gamma_{th}, |h_0(t)|^2\ell(x_0) < \theta\right) + P\left(\Gamma_0(t, \alpha_H) \geq \Gamma_{th}, |h_0(t)|^2\ell(x_0) \geq \theta\right).
$$

(5)

The probability of successful packet reception at user $D_0$ using OMA (i.e., the first term of the right-hand side of (5)) is denoted as $q^{OMA}_H(\theta)$. For simplicity of notation, we denote $\rho = \Gamma_{th}\sigma^2_t/P_S$. If $\theta \leq \rho$, we have $q^{OMA}_H(\theta) = 0$. Otherwise, we have

$$
q^{OMA}_H(\theta) = P\left(\frac{\rho}{\ell(x_0)} \leq |h_0(t)|^2 < \frac{\theta}{\ell(x_0)}\right) = \exp\left(-\rho\left(1 + r_0^{\beta}\right)\right) - \exp\left(-\theta\left(1 + r_0^{\beta}\right)\right),
$$

(6)

where $(a)$ follows from the Rayleigh fading channel.

The probability of successful packet reception at user $D_0$ using NOMA (i.e., the second term of the right-hand side of (5)), denoted as $q^{ON}_{H|L}(\alpha_H, \theta)$, can be expressed as

$$
q^{ON}_{H|L}(\alpha_H, \theta) = P\left(|h_0(t)|^2 \geq \max\left\{\frac{\rho}{\alpha_H - \Gamma_{th}\alpha_L}, \theta\right\}\ell(x_0)\right) = \exp\left(-\max\left\{\frac{\rho}{\alpha_H - \Gamma_{th}\alpha_L}, \theta\right\}\left(1 + r_0^{\beta}\right)\right),
$$

(7)

where $\alpha_H^2 > \Gamma_{th}\alpha_L^2$. Otherwise, we have $q^{ON}_{H|L}(\alpha_H, \theta) = 0$.

After deriving the average service rate of queue $Q_H$, by Loynes’ theorem, queue $Q_H$ is stable if $\lambda_H < \mu^{ON}_H$ where

$$
\lambda_H < \mu^{ON}_H = \begin{cases} 
q^{ON}_{H|L}(\alpha_H, \theta), & \text{if } \theta \leq \rho, \\
q^{OMA}_H(\theta) + q^{ON}_{H|L}(\alpha_H, \theta), & \text{if } \theta > \rho.
\end{cases}
$$

(8)

On the other hand, the service process of queue $Q_L$, can also be categorized into two cases: a) if queue $Q_H$ is non-empty, base station $S$ transmits one packet to user $D_0$ using NOMA when $|h_0(t)|^2\ell(x_0) \geq \theta$, b) if queue $Q_H$ is empty, base station $S$ transmits two packets to users $D_k$ and $D_L$ using NOMA when $D_k \neq D_l$ (which occurs with probability $1 - \frac{1}{M}$) and transmits one packet to user $D_0$ using OMA when $D_k = D_l$ (which occurs with probability $\frac{1}{M}$). Note that, for ease of presentation, the average arrival rates of low-priority users are set to be the same, i.e., $\lambda_m = \lambda_L/M$, $\forall m \in M$, but the analysis can be easily extended to a general scenario with diverse average arrival rates. As all low-priority users follow the same location distribution, the average probability of successful packet reception at each low-priority user is the same. Hence, the average service rate of queue $Q_L$ in dominant system $\Phi^{ON}_1$, denoted as $\mu^{ON}_L$, can be expressed as

$$
\mu^{ON}_L = \mathbb{P}(Q_H(t) > 0)\mathbb{P}(|h_0(t)|^2\ell(x_0) \geq \theta)\mu^{ON}_H(\alpha_L) + \mathbb{P}(Q_H(t) = 0)\left(1 - \frac{1}{M}\right)q^{ON}_L + \frac{1}{M}q^{OMA}_L,
$$

(9)

where the probability of queue $Q_L$ being non-empty is $\mathbb{P}(Q_H(t) > 0) = \lambda_H/\mu^{ON}_H$, $\mu^{ON}_H(\alpha_L)$ denotes the probability of successful packet reception at user $D_k$ with power allocation coefficient $\alpha_L$, when pairing with user $D_0$, $q^{ON}_L$ is the summation of the probabilities of successful packet reception at users $D_k$ and $D_l$ using NOMA, and $q^{OMA}_L$ is the probability of successful packet reception at user $D_0$ using OMA.

The probability of successful packet reception at user $D_k$ when pairing with user $D_0$ is given by

$$
q^{ON}_L(\alpha_L) = \mathbb{P}(\Gamma_0 \rightarrow k(t, \alpha_L) \geq \Gamma_{th}, \Gamma_k(t, \alpha_L) \geq \Gamma_{th}) = \mathbb{P}\left(|h_k(t)|^2 \geq \frac{\rho}{\alpha_L^2 - \Gamma_{th}\alpha_L^2}\ell(x_k), |h_k(t)|^2 \geq \frac{\rho}{\alpha_L^2}\ell(x_k)\right) = \mathbb{E}_{x_k}\left[\exp\left(-\varepsilon_1/\ell(x_k)\right)\right],
$$

(10)

where $\varepsilon_1 = \max\left\{\frac{\rho}{\alpha_L^2 - \Gamma_{th}\alpha_L^2}, \frac{\rho}{\alpha_L^2}\right\}$ and $\mathbb{E}_{x_k}$ [$\cdot$] denotes the expectation of user $D_k$’s location $x_k$. Due to the uniform distribution of low-priority users within a circle with radius $r_1$, the probability density function (PDF) of user $D_k$’s location is given by $f(x_k) = 1/(\pi r_1^2)$. Hence, we have

$$
q^{ON}_{L|H}(\alpha_L) = \frac{2}{r_L^2} \int_{r_L}^{r_1} \exp\left(-\varepsilon_1\left(1 + r_k^\beta\right)\right) r_k dr_k = \frac{2}{r_L^2} \varepsilon_1^{-2/\beta} \exp\left(-\varepsilon_1\gamma\left(\frac{2}{\beta}, \varepsilon_1 r_L^\beta\right)\right),
$$

(11)

where $\gamma(u, v) = \int_0^u e^{-z}z^{u-1}dz$ is the lower incomplete Gamma function [17].

When queue $Q_H$ is empty and $D_k \neq D_l$, base station $S$ transmits the first and second packets from queue $Q_L$ using NOMA according to the distances of their intended users. In particular, among these two users, the near and far users are denoted as $D_n$ and $D_l$ with distances $r_n$ and $r_l$, respectively, and $r_n \leq r_l$. Users $D_k$ and $D_l$ have the same probability (i.e.,
Based on (8) and (18), the stable throughput region in dominant system \( \Phi_{1ON} \) can be expressed as

\[
\mathcal{R}_{1ON} = \left\{ (\lambda_H, \lambda_L) : \frac{\lambda_H}{\eta \mu_{ONT}} + \frac{\lambda_L}{\eta} < 1, \text{for } 0 \leq \lambda_H < \mu_{L}^{ON1} \right\},
\]

where \( \eta = (1 - \frac{1}{M}) \left( q_{L}^{ON}(\alpha_{L}) + q_{H}^{ON}(\alpha_{H}) \right) + \frac{1}{M} \mu_{L}^{OMA} \) and

\[
\delta = \exp \left( -\theta \left( 1 + r_{0}^{2} \right) \right) q_{H|H}^{ON}(\alpha_{L}).
\]

According to (19), stable throughput region \( \mathcal{R}_{1ON} \) depends on the values of threshold \( \theta \) and power allocation coefficients \( (\alpha_{H}, \alpha_{L}) \). In dominant system \( \Phi_{1ON} \), some \( \lambda_{L} \) would make queue \( Q_{L} \) always non-empty. As long as queue \( Q_{L} \) always has packets to transmit, the behaviour of dominant system \( \Phi_{1ON} \) is identical to that of the original opportunistic NOMA system \( \Phi_{1OMA} \). Hence, dominant system \( \Phi_{1ON} \) and the original system \( \Phi_{1OMA} \) are indistinguishable at the boundary points of the stable throughput region.

2) Stable throughput region of dominant system \( \Phi_{2ON}^{OMA} \): In dominant system \( \Phi_{2ON}^{OMA} \), if queue \( Q_{H} \) is empty, then queue \( Q_{L} \) contributes a dummy packet, while queue \( Q_{L} \) acts the same as in the original system \( \Phi_{1OMA} \).

The average service rate of queue \( Q_{L} \) in dominant system \( \Phi_{2ON}^{OMA} \), denoted as \( \mu_{L}^{ON2} \), can be expressed as

\[
\mu_{L}^{ON2} = \exp \left( -\theta \left( 1 + r_{0}^{2} \right) \right) q_{H|H}^{ON}(\alpha_{L}),
\]

where \( q_{H|H}^{ON}(\alpha_{L}) \) is given in (11). Hence, queue \( Q_{L} \) is stable if \( \lambda_{L} \neq \mu_{L}^{ON2} \).

The service process of queue \( Q_{H} \) can also be categorized into two cases: a) if queue \( Q_{L} \) is empty, then base station \( S \) transmits one packet to user \( D_{0} \) using OMA; b) if queue \( Q_{L} \) is non-empty, then base station \( S \) transmits one packet to user \( D_{0} \) using NOMA when \( |h_{0}(t)|^{2} \epsilon(x_{0}) > \theta \), and transmits one packet to user \( D_{0} \) using OMA when \( |h_{0}(t)|^{2} \epsilon(x_{0}) < \theta \). As a result, the average service rate of queue \( Q_{H} \) in dominant system \( \Phi_{2ON}^{OMA} \), denoted as \( \mu_{H}^{ON2} \), can be expressed as

\[
\mu_{H}^{ON2} = \exp \left( -\theta \left( 1 + r_{0}^{2} \right) \right) q_{H|H}^{ON}(\alpha_{L}),
\]

where the probability of queue \( Q_{L} \) being empty is \( \mathbb{P}(Q_{L}(t) = 0) = 1 - \lambda_{L}/\lambda_{H}^{ON2} \), and \( q_{H|H}^{ON}(\alpha_{L}) \) and \( q_{H|H}^{OMA}(\alpha_{L}, \theta) \) are given by (6) and (7), respectively.

After deriving the average service rates of queues \( Q_{L} \) and \( Q_{H} \), by Loynes’ theorem, the stable throughput region in dominant system \( \Phi_{2ON}^{OMA} \) can be expressed as

\[
\mathcal{R}_{2ON} = \left\{ (\lambda_{H}, \lambda_{L}) : \frac{\lambda_{H}}{\eta \mu_{ONT}} + \frac{\lambda_{L}}{\eta} < 1, \text{for } 0 \leq \lambda_{L} < \mu_{L}^{ON2} \right\},
\]

where \( \xi = q_{H|H}^{OMA}(\alpha_{L}) - q_{H|H}^{OMA}(\alpha_{L}, \theta) \). Similarly, stable throughput region \( \mathcal{R}_{2ON}^{OMA} \) depends on the values of threshold \( \theta \) and power allocation coefficients \( (\alpha_{H}^{2}, \alpha_{L}^{2}) \), and dominant system \( \Phi_{2ON}^{OMA} \) and the original system \( \Phi_{2OMA} \) are indistinguishable at the boundary points of the stable throughput region.

Based on the above discussions, the stable throughput region of the original opportunistic NOMA system \( \Phi_{1OMA} \) is equal to the union of the stable throughput regions in dominant systems \( \Phi_{1ON}^{OMA} \) and \( \Phi_{2ON}^{OMA} \), i.e., \( \mathcal{R}_{1ON}^{OMA} = \mathcal{R}_{1ON} \cup \mathcal{R}_{2ON}^{OMA} \).

B. Orthogonal Multiple Access

In this subsection, we present a TDMA-based OMA system, \( \Phi_{OMA} \), as a baseline, where base station \( S \) transmits one packet in one time slot. As queues \( Q_{H} \) and \( Q_{L} \) are not interacting when OMA is utilized, the stability condition of
these two queues can be separately analyzed. As user $D_0$ has a high priority to be served, the average service rate of queue $Q_H$ is $\mu_{Q_H}^{OMA} = \exp \left( -\rho \left( 1 + r_0^\beta \right) \right)$.

On the other hand, when queue $Q_H$ is empty, base station $S$ transmits a packet from queue $Q_L$ to the corresponding user. The average service rate of queue $Q_L$ is given by $\mu_{Q_L}^{OMA} = \mathbb{P}(Q_H = 0) \mathbb{P}(\Gamma(t, 1) \geq \Gamma_{th}) = \frac{1 - \lambda_H}{\mu_{Q_L}^{OMA}} q_{Q_L}^{OMA}$, where $q_{Q_L}^{OMA}$ is given in (17). Based on the above discussions, the stable throughput region of the OMA system can be expressed as

$$R_{OMA} = \left\{ \left( \lambda_H, \lambda_L \right) : \frac{\lambda_H}{\mu_{Q_H}^{OMA}} + \frac{\lambda_L}{q_{Q_L}^{OMA}} < 1, \right. \left. \text{for } 0 \leq \lambda_H < \exp \left( -\rho \left( 1 + r_0^\beta \right) \right) \right\}. \quad (22)$$

IV. NUMERICAL RESULTS

In this section, we evaluate the stable throughput regions of the proposed opportunistic NOMA and baseline OMA schemes. The radius of the network coverage area is $r = 1.5$ km. High-priority user $D_0$ is located at $r_0 = 1.2$ km away from base station $S$, and $M = 10$ low-priority users are randomly distributed within a circle with radius $r_L = 1$ km centered at base station $S$. Transmission power $P_S$ and noise power $\sigma^2$ are set to be 1 W and -100 dBm, respectively. We consider Rayleigh fading channels and the path loss exponent $\beta$ is set to be 4. The power allocation coefficients of far and near users of queue $Q_L$, $(\alpha_H^2, \alpha_L^2)$, are set to be $(0.8, 0.2)$. In this section, we evaluate the stable throughput regions of the proposed opportunistic NOMA and baseline OMA schemes. The radius of the network coverage area is $r = 1.5$ km. High-priority user $D_0$ is located at $r_0 = 1.2$ km away from base station $S$, and $M = 10$ low-priority users are randomly distributed within a circle with radius $r_L = 1$ km centered at base station $S$. Transmission power $P_S$ and noise power $\sigma^2$ are set to be 1 W and -100 dBm, respectively. We consider Rayleigh fading channels and the path loss exponent $\beta$ is set to be 4. The power allocation coefficients of far and near users of queue $Q_L$, $(\alpha_H^2, \alpha_L^2)$, are set to be $(0.8, 0.2)$.

Fig. 3 shows the impact of threshold $\theta$ on the stable throughput region of the opportunistic NOMA system when $(\alpha_H^2, \alpha_L^2) = (0.8, 0.2)$ and $\Gamma_{th} = 2$. The stable throughput region of opportunistic NOMA is the union of that of dominant systems $\Phi_1^{ON}$ and $\Phi_2^{ON}$, given by (19) and (21), respectively. When threshold $\theta = 2\rho = 2\Gamma_{th} \sigma^2/P_S$, the achievable $\lambda_L$ in dominant system $\Phi_1^{ON}$ is much larger than that in OMA system $\Phi_1^{OMA}$, but the maximum achievable $\lambda_H$ in dominant system $\Phi_1^{ON}$ is smaller than that in OMA system $\Phi_1^{OMA}$. This is due to the fact that the opportunistic NOMA scheme provides more transmission opportunities to low-priority users, at the cost of reducing the average service rate of high-priority user $D_0$. When $\theta = 2.5\rho$, the maximum achievable $\lambda_H$ in dominant system $\Phi_1^{ON}$ and OMA system $\Phi_1^{OMA}$ are the same, which shows that the opportunistic NOMA scheme can enhance the performance of low-priority users without sacrificing the performance of high-priority user $D_0$ by appropriately selecting the value of threshold $\theta$. By increasing the value of threshold $\theta$, the achievable $\lambda_H$ in dominant systems $\Phi_1^{ON}$ and $\Phi_2^{ON}$ increases, while the opportunity to perform NOMA decreases. The opportunistic NOMA system enhances the stable throughput region when compared with the OMA system, i.e., $R_1^{ON} \supset R_1^{OMA}$.

Fig. 4 illustrates the impact of power allocation coefficients $(\alpha_H^2, \alpha_L^2)$ on the stable throughput region of opportunistic NOMA system $\Phi_1^{ON}$ with parameters $\theta = 2\rho$ and $\Gamma_{th} = 2$. Stable throughput region $R_1^{ON} \cup R_2^{ON}$ changes significantly with power allocation coefficients $(\alpha_H^2, \alpha_L^2)$. When $(\alpha_H^2, \alpha_L^2) = (0.7, 0.3)$, the achievable $\lambda_L$ in dominant systems $\Phi_1^{ON}$ and $\Phi_2^{ON}$ is less than that in OMA system $\Phi_1^{OMA}$ when $\lambda_H > 0.16$, as low-priority user $D_m$ is bottlenecked by successful decoding of the signal intended for high-priority user $D_0$, which is the prerequisite of performing SIC. By increasing $\alpha_H^2$, the maximum achievable $\lambda_H$ and $\lambda_L$ in dominant systems $\Phi_1^{ON}$ and $\Phi_2^{ON}$ increases and decreases, respectively, as more transmission power is allocated to high-priority user $D_0$. By enabling NOMA to serve the packets from queue $Q_L$, the maximum achievable $\lambda_L$ in dominant system $\Phi_1^{ON}$ is much greater than that in OMA system $\Phi_1^{OMA}$.

Fig. 5 plots the impact of reception threshold $\Gamma_{th}$ on the stable throughput region of opportunistic NOMA system $\Phi_1^{ON}$ when $\theta = 2\rho$ and $(\alpha_H^2, \alpha_L^2) = (0.8, 0.2)$. With a decrease of reception threshold $\Gamma_{th}$, the maximum achievable $\lambda_H$ and $\lambda_L$ in both opportunistic NOMA system $\Phi_1^{ON}$ and OMA system $\Phi_1^{OMA}$ increases, as the probability of successful packet reception at each user increases. With a smaller reception threshold, the probability of queue $Q_H$ being empty is higher, which leads to more time slots available for the base station to serve queue $Q_H$ using NOMA. Hence, the performance gap between dominant system $\Phi_1^{ON}$ and OMA system $\Phi_1^{OMA}$

---

**Fig. 3:** Stable throughput region with different values of threshold $\theta$ and parameters $(\alpha_H^2, \alpha_L^2) = (0.8, 0.2)$ and $\Gamma_{th} = 2$.

**Fig. 4:** Stable throughput region with different values of power allocation coefficients $(\alpha_H^2, \alpha_L^2)$ and parameters $\theta = 2\rho$ and $\Gamma_{th} = 2$.\n
---

\[ R_{OMA} = \left\{ \left( \lambda_H, \lambda_L \right) : \frac{\lambda_H}{\mu_{Q_H}^{OMA}} + \frac{\lambda_L}{q_{Q_L}^{OMA}} < 1, \right. \left. \text{for } 0 \leq \lambda_H < \exp \left( -\rho \left( 1 + r_0^\beta \right) \right) \right\}. \quad (22) \]
downlink NOMA transmission with dynamic traffic arrival for when the channel gain is lower. Average service rate of queue $Q_l$ becomes larger when reception threshold $\Gamma_{th}$ is smaller.

Fig. 6 shows the impact of threshold $\theta$ and power allocation coefficients $(\alpha_H^2, \alpha_L^2)$ on the average service rate of queue $Q_L$ when $\lambda_H = 0.3$ and $\Gamma_{th} = 2$. With the variation of threshold $\theta$, there exists an optimal point of the average service rate of queue $Q_L$. The average service rate of queue $Q_L$ can be greater than 1 as NOMA is enabled to simultaneously serve two packets from queue $Q_L$ when queue $Q_H$ is empty. If $(\alpha_H^2, \alpha_L^2) = (0.8, 0.2)$, the average service rate of queue $Q_L$ increases with $\theta$ when $\theta < 0.5 \times 10^{-12}$. By enabling NOMA when the channel gain between base station $S$ and user $D_0$ is larger, less packet retransmissions are required to guarantee the stability of queue $Q_H$, which in turn provide more transmission opportunities to low-priority users. The average service rate of queue $Q_L$ decreases with $\theta$ when $\theta > 0.5 \times 10^{-12}$ and converges to 0.87, as the probability of enabling NOMA becomes smaller. By increasing $\alpha_L^2$ to 0.9, the optimal threshold $\theta$ that can maximize the average service rate of queue $Q_L$ becomes smaller, as allocating more transmission power to user $D_0$ allows NOMA to be enabled when the channel gain is lower.

V. CONCLUSION

In this paper, we studied the stable throughput region of downlink NOMA transmission with dynamic traffic arrival for users with different priorities. To reduce the adverse effect of channel sharing and transmission power splitting due to NOMA on the high-priority user, we proposed an opportunistic NOMA scheme by using limited instantaneous CSI at the base station. By utilizing tools from stochastic geometry and queueing theory, we characterized the stable throughput region of the opportunistic NOMA system. Numerical results showed that the proposed NOMA scheme can significantly increase the transmission opportunities and enhance the stable throughput region. For future work, we will jointly optimize the values of threshold $\theta$ and transmission power allocation coefficients to maximize the stable throughput region of the proposed opportunistic NOMA scheme.

REFERENCES