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-

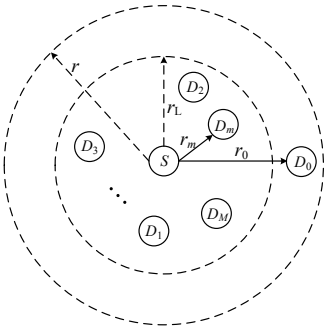

Fig. 1: Illustration of the network topology for downlink NOMA transmission with spatially random users.

gion 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., $\{D_m, m \in \mathcal{M} = \{1, \ldots, M\}\}$) 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 $r_{\mathrm{L}} < r$. On the other hand, the distance between base station S and high-priority user D_0 is fixed and denoted as $r_0 \in (r_{\mathrm{L}}, r]$. 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 λ_m (packets/time slot). Hence, the average arrival rate of queue Q_{L} is $\lambda_{\mathrm{L}} = \sum_{m=1}^{M} \lambda_m$. 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.

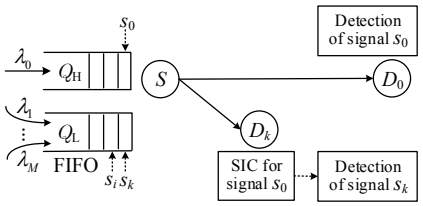

Fig. 2: Illustration of the queueing model for downlink NOMA transmission with dynamic traffic arrival.

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 $Q_{\mathrm{H}}(t)$ and $Q_{\mathrm{L}}(t)$ 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 Loynes' theorem [15], the sufficient condition for the stability of queue Q_{H} is that $\lambda_{\mathrm{H}} < \mu_{\mathrm{H}}$, where $\lambda_{\mathrm{H}} = \lambda_0$ and μ_{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 $\{\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}\}$, 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_k, 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 $\alpha_{\mathrm{H}}\sqrt{P_S}s_0(t) + \alpha_{\mathrm{L}}\sqrt{P_S}s_k(t)$, where P_S denotes the total transmission power of base station S, α_{H} and α_{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 $r_m < r_0$, $\forall\ m \in \mathcal{M}$, according to the design principle of NOMA, we have $\alpha_{\rm H} > \alpha_{\rm L}$ and $\alpha_{\rm H}^2 + \alpha_{\rm L}^2 = 1$. Over the block fading channel, the signal received by user D_m at time slot t is given by

$$y_m(t) = (\alpha_{\rm H} s_0(t) + \alpha_{\rm L} s_k(t))\sqrt{P_S} h_m(t)\sqrt{\ell\left(x_m\right)} + n_m(t), \quad (1)$$

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 σ^2 at time slot t, $\ell(x_m) = \left(1 + r_m^{\beta}\right)^{-1}$ 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 β 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

$$\Gamma_{0|\rm L}\left(t, \alpha_{\rm H}\right) = \frac{\alpha_{\rm H}^2 P_S |h_0(t)|^2 \ell\left(x_0\right)}{\alpha_{\rm L}^2 P_S |h_0(t)|^2 \ell\left(x_0\right) + \sigma^2}, \quad (2)$$

where $\Gamma_{0|\rm L}\left(t, \alpha_{\rm H}\right)$ 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

$$\Gamma_{0\to k}(t, \alpha_{\rm H}) = \frac{\alpha_{\rm H}^2 P_S |h_k(t)|^2 \ell\left(x_k\right)}{\alpha_{\rm L}^2 P_S |h_k(t)|^2 \ell\left(x_k\right) + \sigma^2}, \quad (3)$$

where $\Gamma_{0\to k}(t, \alpha_{\rm H})$ denotes the SINR of signal $s_0(t)$ observed by user D_k at time slot t.

Let $\Gamma_{\rm th}$ denote the threshold for successful packet reception. If user D_k successfully decodes signal $s_0(t)$, i.e., $\Gamma_{0\to k}(t, \alpha_{\rm H}) \geq \Gamma_{\rm th}$, 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

$$\Gamma_k(t, \alpha_{\rm L}) = \frac{\alpha_{\rm L}^2 P_S |h_k(t)|^2 \ell\left(x_k\right)}{\sigma^2}, \quad (4)$$

where $\Gamma_k(t, \alpha_{\rm L})$ 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 $\{\Gamma_{0|\rm L}\left(t, \alpha_{\rm H}\right) \geq \Gamma_{\rm th}\}$ and $\{\Gamma_{0\to k}(t, \alpha_{\rm H}) \geq \Gamma_{\rm th} \cap \Gamma_k(t, \alpha_{\rm L}) \geq \Gamma_{\rm th}\}$ 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 $\{\Gamma_k(t, 1) \geq \Gamma_{\rm th}\}$ 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., $\Gamma_{0|\rm L}\left(t, \alpha_{\rm H}\right) < \Gamma_0(t, 1) = P_S |h_0(t)|^2 \ell(x_0)/\sigma^2$. As a result, to guarantee the stability of queue $Q_{\rm H}$, NOMA cannot always be enabled, especially when the average arrival rate $\lambda_{\rm H}$ 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 $Q_{\rm H}$ 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, $|h_0(t)|^2 \ell\left(x_0\right)$, is not less than a threshold, θ, and feeds back 0 to base station S otherwise. Second, when queue $Q_{\rm H}$ is empty at time slot t, the intended receivers of the first two packets from queue $Q_{\rm L}$ 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 $\Phi^{\rm ON}$, is described as follows. As user D_0 has a high priority to be served, base station S transmits a packet from queue $Q_{\rm H}$ whenever it is non-empty. Without loss of generality, the intended receiver of the second packet from queue $Q_{\rm L}$ at time slot t, when available, is denoted as user D_i. Depending on the status of queues $Q_{\rm H}$ and $Q_{\rm L}$ at time slot t, the packet transmissions in opportunistic NOMA system $\Phi^{\rm ON}$ can be categorized into the following three cases:

Case 1: If $Q_{\rm H}(t) > 0$ and $Q_{\rm L}(t) > 0$, then base station S transmits the first packet from queue $Q_{\rm H}$ and the first packet from queue $Q_{\rm L}$ to users D_0 and D_k, respectively, using NOMA with fixed power allocation coefficients $(\alpha_{\rm H}^2, \alpha_{\rm L}^2)$ when $|h_0(t)|^2 \ell(x_0) \geq \theta$, and transmits the first packet from queue $Q_{\rm H}$ to user D_0 using OMA with power P_S when $|h_0(t)|^2 \ell(x_0) < \theta$.

Case 2: If $Q_{\rm H}(t) > 0$ and $Q_{\rm L}(t) = 0$, then base station S transmits the first packet from queue $Q_{\rm H}$ to user D_0 using OMA with power P_S.

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

Based on the opportunistic NOMA system described above, the average service rate of queue $Q_{\rm H}$ depends on the status of queue $Q_{\rm L}$. In particular, when queue $Q_{\rm L}$ is empty, base station S transmits a packet from queue $Q_{\rm H}$ to user D_0 using OMA. On the other hand, when queue $Q_{\rm L}$ is non-empty, base station S transmits the first packet from queue $Q_{\rm H}$ and the first packet from queue $Q_{\rm L}$ to users D_0 and D_k using NOMA with probability $\mathbb{P}\left(|h_0(t)|^2 \ell(x_0) \geq \theta\right) = \exp\left(-\theta\left(1 + r_0^{\beta}\right)\right)$. 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 $Q_{\rm L}$ also depends on the status of queue $Q_{\rm H}$. As a result, queues $Q_{\rm H}$ and $Q_{\rm L}$ 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* $\Phi_1^{\rm ON}$ and $\Phi_2^{\rm ON}$ based on the original opportunistic NOMA system $\Phi^{\rm ON}$. The dominant systems, as a modification of the original system (i.e., $\Phi^{\rm ON}$), 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 $\Phi_1^{\rm ON}$: In dominant system $\Phi_1^{\rm ON}$, if queue $Q_{\rm L}$ is empty, then queue $Q_{\rm L}$ contributes a dummy packet when user D_0 feeds back 1 to base station S, while queue $Q_{\rm H}$ acts the same as in the original system $\Phi^{\rm ON}$.

In this case, the service process of queue $Q_{\rm H}$ 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 $Q_{\rm H}$ in dominant system $\Phi_1^{\rm ON}$, denoted as $\mu_{\rm H}^{\rm ON1}$, can be expressed as

$$\begin{aligned}\mu_{\rm H}^{\rm ON1} =& \mathbb{P}\left(\Gamma_0(t,1) \geq \Gamma_{\rm th}, |h_0(t)|^2\ell(x_0) < \theta\right) \\ &+ \mathbb{P}\left(\Gamma_{0|{\rm L}}(t,\alpha_{\rm H}) \geq \Gamma_{\rm th}, |h_0(t)|^2\ell(x_0) \geq \theta\right). \quad (5)\end{aligned}$$

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 $q_{\rm H}^{\rm OMA}(\theta)$. For simplicity of notation, we denote $\rho = \Gamma_{\rm th}\sigma^2/P_S$. If $\theta \leq \rho$, we have $q_{\rm H}^{\rm OMA}(\theta) = 0$. Otherwise, we have

$$\begin{aligned}q_{\rm H}^{\rm OMA}(\theta) &= \mathbb{P}\left(\tfrac{\rho}{\ell(x_0)} \leq |h_0(t)|^2 < \tfrac{\theta}{\ell(x_0)}\right) \\ &\overset{(a)}{=} \exp\left(-\rho\left(1+r_0^\beta\right)\right) - \exp\left(-\theta\left(1+r_0^\beta\right)\right), \quad (6)\end{aligned}$$

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 $q_{\rm H|L}^{\rm ON}(\alpha_{\rm H},\theta)$, can be expressed as

$$\begin{aligned}q_{\rm H|L}^{\rm ON}(\alpha_{\rm H},\theta) &= \mathbb{P}\left(|h_0(t)|^2 \geq \max\left\{\frac{\rho}{\alpha_{\rm H}^2 - \Gamma_{\rm th}\alpha_{\rm L}^2}, \theta\right\}\frac{1}{\ell(x_0)}\right) \\ &= \exp\left(-\max\left\{\frac{\rho}{\alpha_{\rm H}^2 - \Gamma_{\rm th}\alpha_{\rm L}^2}, \theta\right\}\left(1+r_0^\beta\right)\right), \quad (7)\end{aligned}$$

where $\alpha_{\rm H}^2 > \Gamma_{\rm th}\alpha_{\rm L}^2$. Otherwise, we have $q_{\rm H|L}^{\rm ON}(\alpha_{\rm H},\theta) = 0$.

After deriving the average service rate of queue $Q_{\rm H}$, by Loynes' theorem, queue $Q_{\rm H}$ is stable if

$$\lambda_{\rm H} < \mu_{\rm H}^{\rm ON1} = \begin{cases} q_{\rm H|L}^{\rm ON}(\alpha_{\rm H},\theta), & \text{if } \theta \leq \rho, \\ q_{\rm H}^{\rm OMA}(\theta) + q_{\rm H|L}^{\rm ON}(\alpha_{\rm H},\theta), & \text{if } \theta > \rho. \end{cases} \quad (8)$$

On the other hand, the service process of queue $Q_{\rm L}$ can also be categorized into two cases: a) if queue $Q_{\rm H}$ is non-empty, base station S transmits one packet to user D_k using NOMA when $|h_0(t)|^2\ell(x_0) \geq \theta$; b) if queue $Q_{\rm H}$ is empty, base station S transmits two packets to users D_k and D_i using NOMA when $D_k \neq D_i$ (which occurs with probability $1 - \frac{1}{M}$), and transmits one packet to user D_k using OMA when $D_k = D_i$ (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., $\lambda_m = \lambda_{\rm L}/M$, $\forall\, m \in \mathcal{M}$, 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 $Q_{\rm L}$ in dominant system $\Phi_1^{\rm ON}$, denoted as $\mu_{\rm L}^{\rm ON1}$, can be expressed as

$$\begin{aligned}\mu_{\rm L}^{\rm ON1} =& \mathbb{P}(Q_{\rm H}(t) > 0)\mathbb{P}(|h_0(t)|^2\ell(x_0) \geq \theta) q_{\rm L|H}^{\rm ON}(\alpha_{\rm L}) \\ &+ \mathbb{P}(Q_{\rm H}(t) = 0)\left(\left(1 - \frac{1}{M}\right)q_{\rm L}^{\rm ON} + \frac{1}{M}q_{\rm L}^{\rm OMA}\right), \quad (9)\end{aligned}$$

where the probability of queue $Q_{\rm H}$ being non-empty is $\mathbb{P}(Q_{\rm H}(t) > 0) = \lambda_{\rm H}/\mu_{\rm H}^{\rm ON1}$, $q_{\rm L|H}^{\rm ON}(\alpha_{\rm L})$ denotes the probability of successful packet reception at user D_k with power allocation coefficient $\alpha_{\rm L}$ when pairing with user D_0, $q_{\rm L}^{\rm ON}$ is the summation of the probabilities of successful packet reception at users D_k and D_i using NOMA, and $q_{\rm L}^{\rm OMA}$ is the probability of successful packet reception at user D_k using OMA.

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

$$\begin{aligned}q_{\rm L|H}^{\rm ON}(\alpha_{\rm L}) &= \mathbb{P}\left(\Gamma_{0\to k}(t,\alpha_{\rm H}) \geq \Gamma_{\rm th}, \Gamma_k(t,\alpha_{\rm L}) \geq \Gamma_{\rm th}\right) \\ &= \mathbb{P}\left(|h_k(t)|^2 \geq \frac{\rho}{\left(\alpha_{\rm H}^2 - \Gamma_{\rm th}\alpha_{\rm L}^2\right)\ell(x_k)}, |h_k(t)|^2 \geq \frac{\rho}{\alpha_{\rm L}^2\ell(x_k)}\right) \\ &= \mathbb{E}_{x_k}\left[\exp\left(-\varepsilon_1/\ell(x_k)\right)\right], \quad (10)\end{aligned}$$

where $\varepsilon_1 = \max\left\{\frac{\rho}{\alpha_{\rm H}^2 - \Gamma_{\rm th}\alpha_{\rm L}^2}, \frac{\rho}{\alpha_{\rm L}^2}\right\}$ 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 $r_{\rm L}$, the probability density function (PDF) of user D_k's location is given by $f(x_k) = 1/\left(\pi r_{\rm L}^2\right)$. Hence, we have

$$\begin{aligned}q_{\rm L|H}^{\rm ON}(\alpha_{\rm L}) &= \frac{2}{r_{\rm L}^2}\int_0^{r_{\rm L}} \exp\left(-\varepsilon_1\left(1+r_k^\beta\right)\right) r_k {\rm d}r_k \\ &= \frac{2}{r_{\rm L}^2\beta}\varepsilon_1^{-2/\beta}\exp\left(-\varepsilon_1\right)\gamma\left(\frac{2}{\beta}, \varepsilon_1 r_{\rm L}^\beta\right), \quad (11)\end{aligned}$$

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

When queue $Q_{\rm H}$ is empty and $D_k \neq D_i$, base station S transmits the first and second packets from queue $Q_{\rm L}$ using NOMA according to the distances of their intended users. In particular, among these two users, the near and far users are denoted as $D_{\rm n}$ and $D_{\rm f}$ with distances $r_{\rm n}$ and $r_{\rm f}$, respectively, and $r_{\rm n} \leq r_{\rm f}$. Users D_k and D_i have the same probability (i.e.,

0.5) to be the near or far user. For instance, if $r_k \leq r_i$, we have $D_{\rm n} = D_k$ and $D_{\rm f} = D_i$, and we have $D_{\rm n} = D_i$ and $D_{\rm f} = D_k$ otherwise. In addition, we set $\alpha_{\rm f} \geq \alpha_{\rm n}$ and $\alpha_{\rm n}^2 + \alpha_{\rm f}^2 = 1$.

Due to the uniform distribution of users $D_{\rm n}$ and $D_{\rm f}$ [18], the PDF of the distance of far user $D_{\rm f}$ is given by

$$f(r_{\rm f}) = 4r_{\rm f}^3/r_{\rm L}^4, \quad 0 \leq r_{\rm f} \leq r_{\rm L}. \tag{12}$$

The probability of successful packet reception at user $D_{\rm f}$ using NOMA, denoted as $q_{\rm f|n}^{\rm ON}(\alpha_{\rm f})$, can be expressed as

$$\begin{aligned} q_{\rm f|n}^{\rm ON}(\alpha_{\rm f}) &= \mathbb{P}\left(\Gamma_{\rm f|n}(t,\alpha_{\rm f}) \geq \Gamma_{\rm th}\right) \\ &= \mathbb{E}_{x_{\rm f}}\left[\exp\left(-\varepsilon_2/\ell(x_{\rm f})\right)\right] \\ &= \frac{4}{r_{\rm L}^4}\int_0^{r_{\rm L}} \exp\left(-\varepsilon_2\left(1+r_{\rm f}^{\beta}\right)\right) r_{\rm f}^3 {\rm d}r_{\rm f} \\ &= \frac{4}{r_{\rm L}^4\beta}\varepsilon_2^{-4/\beta}\exp\left(-\varepsilon_2\right)\gamma\left(\frac{4}{\beta}, \varepsilon_2 r_{\rm L}^{\beta}\right), \end{aligned} \tag{13}$$

where $\varepsilon_2 = \frac{\rho}{\alpha_{\rm f}^2 - \Gamma_{\rm th}\alpha_{\rm n}^2}$ and $\alpha_{\rm f}^2 > \Gamma_{\rm th}\alpha_{\rm n}^2$.

The PDF of the distance of near user $D_{\rm n}$ is given by

$$f(r_{\rm n}) = 4\frac{r_{\rm n}}{r_{\rm L}^2}\left(1 - \frac{r_{\rm n}^2}{r_{\rm L}^2}\right), \quad 0 \leq r_{\rm n} \leq r_{\rm L}. \tag{14}$$

The probability of successful packet reception at user $D_{\rm n}$ using NOMA, denoted as $q_{\rm n|f}^{\rm ON}(\alpha_{\rm n})$, is given by

$$\begin{aligned} q_{\rm n|f}^{\rm ON}(\alpha_{\rm n}) &= \mathbb{P}\left(\Gamma_{{\rm f}\to{\rm n}}(t,\alpha_{\rm f}) \geq \Gamma_{\rm th}, \Gamma_{\rm n}(t,\alpha_{\rm n}) \geq \Gamma_{\rm th}\right) \\ &= \mathbb{E}_{x_{\rm n}}\left[\exp\left(-\varepsilon_3/\ell(x_{\rm n})\right)\right] \\ &= \frac{4}{r_{\rm L}^2}\int_0^{r_{\rm L}} \exp\left(-\varepsilon_3\left(1+r_{\rm n}^{\beta}\right)\right)\left(r_{\rm n} - \frac{r_{\rm n}^3}{r_{\rm L}^2}\right){\rm d}r_{\rm n} \\ &= \frac{4}{r_{\rm L}^2\beta}\varepsilon_3^{-2/\beta}\exp\left(-\varepsilon_3\right)\gamma\left(\frac{2}{\beta}, \varepsilon_3 r_{\rm L}^{\beta}\right) \\ &\quad - \frac{4}{r_{\rm L}^4\beta}\varepsilon_3^{-4/\beta}\exp\left(-\varepsilon_3\right)\gamma\left(\frac{4}{\beta}, \varepsilon_3 r_{\rm L}^{\beta}\right), \end{aligned} \tag{15}$$

where $\varepsilon_3 = \max\left\{\frac{\rho}{\alpha_{\rm f}^2 - \Gamma_{\rm th}\alpha_{\rm n}^2}, \frac{\rho}{\alpha_{\rm n}^2}\right\}$ and $\alpha_{\rm f}^2 > \Gamma_{\rm th}\alpha_{\rm n}^2$.

Based on (13) and (15), we have

$$q_{\rm L}^{\rm ON} = q_{\rm f|n}^{\rm ON}(\alpha_{\rm f}) + q_{\rm n|f}^{\rm ON}(\alpha_{\rm n}). \tag{16}$$

Similarly, the probability of successful packet reception at user D_k using OMA can be expressed as

$$\begin{aligned} q_{\rm L}^{\rm OMA} &= \mathbb{P}\left(\Gamma_k(t,1) \geq \Gamma_{\rm th}\right) \\ &= \frac{2}{r_{\rm L}^2\beta}\rho^{-2/\beta}\exp(-\rho)\gamma\left(\frac{2}{\beta}, \rho r_{\rm L}^{\beta}\right). \end{aligned} \tag{17}$$

By substituting (11), (16), and (17) into (9), the average service rate of queue $Q_{\rm L}$ in dominant system $\Phi_1^{\rm ON}$ can be derived. By Loynes' theorem, queue $Q_{\rm L}$ is stable if

$$\begin{aligned} \lambda_{\rm L} < \mu_{\rm L}^{\rm ON1} &= \frac{\lambda_{\rm H}}{\mu_{\rm H}^{\rm ON1}}\exp\left(-\theta\left(1+r_0^{\beta}\right)\right) q_{\rm L|H}^{\rm ON}(\alpha_{\rm L}) \\ &\quad + \left(1 - \frac{\lambda_{\rm H}}{\mu_{\rm H}^{\rm ON1}}\right)\left(\left(1 - \frac{1}{M}\right) q_{\rm L}^{\rm ON} + \frac{1}{M} q_{\rm L}^{\rm OMA}\right). \end{aligned} \tag{18}$$

Based on (8) and (18), the stable throughput region in dominant system $\Phi_1^{\rm ON}$ is given by

$$\mathcal{R}_1^{\rm ON} = \left\{(\lambda_{\rm H}, \lambda_{\rm L}) : \frac{\lambda_{\rm H}(\eta - \delta)}{\eta\mu_{\rm H}^{\rm ON1}} + \frac{\lambda_{\rm L}}{\eta} < 1, \text{for } 0 \leq \lambda_{\rm H} < \mu_{\rm H}^{\rm ON1}\right\}, \tag{19}$$

where $\eta = \left(1 - \frac{1}{M}\right)\left(q_{\rm f|n}^{\rm ON}(\alpha_{\rm f}) + q_{\rm n|f}^{\rm ON}(\alpha_{\rm n})\right) + \frac{1}{M} q_{\rm L}^{\rm OMA}$ and $\delta = \exp\left(-\theta\left(1+r_0^{\beta}\right)\right) q_{\rm L|H}^{\rm ON}(\alpha_{\rm L})$. According to (19), stable throughput region $\mathcal{R}_1^{\rm ON}$ depends on the values of threshold θ and power allocation coefficients $(\alpha_{\rm H}^2, \alpha_{\rm L}^2)$. In dominant system $\Phi_1^{\rm ON}$, some $\lambda_{\rm L}$ would make queue $Q_{\rm L}$ always non-empty. As long as queue $Q_{\rm L}$ always has packets to transmit, the behaviour of dominant system $\Phi_1^{\rm ON}$ is identical to that of the original opportunistic NOMA system $\Phi^{\rm ON}$. Hence, dominant system $\Phi_1^{\rm ON}$ and the original system $\Phi^{\rm ON}$ are indistinguishable at the boundary points of the stable throughput region.

2) Stable throughput region of dominant system $\Phi_2^{\rm ON}$: In dominant system $\Phi_2^{\rm ON}$, if queue $Q_{\rm H}$ is empty, then queue $Q_{\rm H}$ contributes a dummy packet, while queue $Q_{\rm L}$ acts the same as in the original system $\Phi^{\rm ON}$.

The average service rate of queue $Q_{\rm L}$ in dominant system $\Phi_2^{\rm ON}$, denoted as $\mu_{\rm L}^{\rm ON2}$, can be expressed as $\mu_{\rm L}^{\rm ON2} = \exp\left(-\theta\left(1+r_0^{\beta}\right)\right) q_{\rm L|H}^{\rm ON}(\alpha_{\rm L})$, where $q_{\rm L|H}^{\rm ON}(\alpha_{\rm L})$ is given in (11). Hence, queue $Q_{\rm L}$ is stable if $\lambda_{\rm L} < \mu_{\rm L}^{\rm ON2}$.

The service process of queue $Q_{\rm H}$ can also be categorized into two cases: a) if queue $Q_{\rm L}$ is empty, then base station S transmits one packet to user D_0 using OMA; b) if queue $Q_{\rm L}$ is non-empty, then base station S transmits one packet to user D_0 using NOMA when $|h_0(t)|^2\ell(x_0) \geq \theta$, and transmits one packet to user D_0 using OMA when $|h_0(t)|^2\ell(x_0) < \theta$. As a result, the average service rate of queue $Q_{\rm H}$ in dominant system $\Phi_2^{\rm ON}$, denoted as $\mu_{\rm H}^{\rm ON2}$, can be expressed as

$$\begin{aligned} \mu_{\rm H}^{\rm ON2} &= \mathbb{P}(Q_{\rm L}(t) = 0) q_{\rm H}^{\rm OMA}(\infty) \\ &\quad + \mathbb{P}\left(Q_{\rm L}(t) > 0\right)\left(q_{\rm H}^{\rm OMA}(\theta) + q_{\rm H|L}^{\rm ON}(\alpha_{\rm H}, \theta)\right), \end{aligned} \tag{20}$$

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

After deriving the average service rates of queues $Q_{\rm L}$ and $Q_{\rm H}$, by Loynes' theorem, the stable throughput region in dominant system $\Phi_2^{\rm ON}$ can be expressed as

$$\begin{aligned} \mathcal{R}_2^{\rm ON} = \Big\{ (\lambda_{\rm H}, \lambda_{\rm L}) : &\frac{\lambda_{\rm H}}{q_{\rm H}^{\rm OMA}(\infty)} + \frac{\xi\lambda_{\rm L}}{\mu_{\rm L}^{\rm ON2} q_{\rm H}^{\rm OMA}(\infty)} < 1, \\ &\text{for } 0 \leq \lambda_{\rm L} < \mu_{\rm L}^{\rm ON2} = \exp\left(-\theta\left(1+r_0^{\beta}\right)\right) q_{\rm L|H}^{\rm ON}(\alpha_{\rm L})\Big\}, \end{aligned} \tag{21}$$

where $\xi = q_{\rm H}^{\rm OMA}(\infty) - q_{\rm H}^{\rm OMA}(\theta) - q_{\rm H|L}^{\rm ON}(\alpha_{\rm H}, \theta)$. Similarly, stable throughput region $\mathcal{R}_2^{\rm ON}$ depends on the values of threshold θ and power allocation coefficients $(\alpha_{\rm H}^2, \alpha_{\rm L}^2)$, and dominant system $\Phi_2^{\rm ON}$ and the original system $\Phi^{\rm ON}$ 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 $\Phi^{\rm ON}$ is equal to the union of the stable throughput regions in dominant systems $\Phi_1^{\rm ON}$ and $\Phi_2^{\rm ON}$, i.e., $\mathcal{R}^{\rm ON} = \mathcal{R}_1^{\rm ON} \cup \mathcal{R}_2^{\rm ON}$.

B. Orthogonal Multiple Access

In this subsection, we present a TDMA-based OMA system, $\Phi^{\rm OMA}$, as a baseline, where base station S transmits one packet in one time slot. As queues $Q_{\rm H}$ and $Q_{\rm L}$ are not interacting when OMA is utilized, the stability condition of

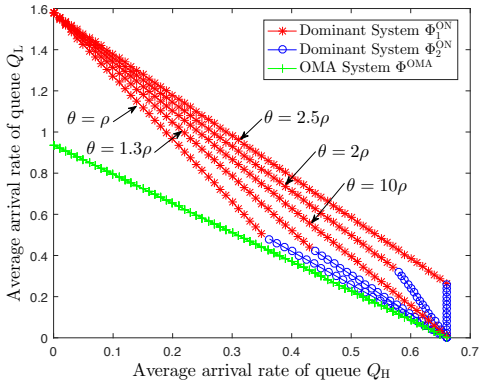

Fig. 3: Stable throughput region with different values of threshold θ and parameters $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$ and $\Gamma_{\mathrm{th}} = 2$.

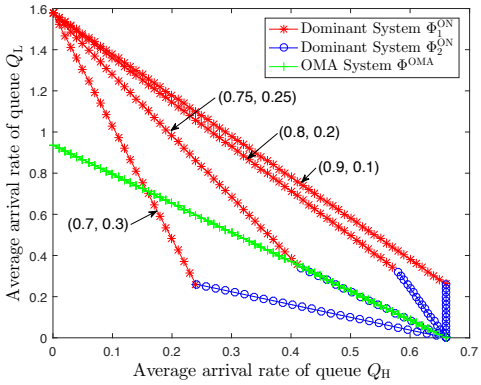

Fig. 4: Stable throughput region with different values of power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$ and parameters $\theta = 2\rho$ and $\Gamma_{\mathrm{th}} = 2$.

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 $\mu_{\mathrm{H}}^{\mathrm{OMA}} = \exp\left(-\rho\left(1 + r_0^{\beta}\right)\right)$.

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 $\mu_{\mathrm{L}}^{\mathrm{OMA}} = \mathbb{P}\left(Q_{\mathrm{H}} = 0\right)\mathbb{P}\left(\Gamma_k(t,1) \geq \Gamma_{\mathrm{th}}\right) = \left(1 - \lambda_{\mathrm{H}}/\mu_{\mathrm{H}}^{\mathrm{OMA}}\right) q_{\mathrm{L}}^{\mathrm{OMA}}$, where $q_{\mathrm{L}}^{\mathrm{OMA}}$ is given in (17). Based on the above discussions, the stable throughput region of the OMA system can be expressed as

$$\mathcal{R}^{\mathrm{OMA}} = \Bigg\{ (\lambda_{\mathrm{H}}, \lambda_{\mathrm{L}}) : \frac{\lambda_{\mathrm{H}}}{\exp\left(-\rho\left(1 + r_0^{\beta}\right)\right)} + \frac{\lambda_{\mathrm{L}}}{q_{\mathrm{L}}^{\mathrm{OMA}}} < 1, \text{ for } 0 \leq \lambda_{\mathrm{H}} < \exp\left(-\rho\left(1 + r_0^{\beta}\right)\right) \Bigg\}. \quad (22)$$

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 $r_{\mathrm{L}} = 1$ km centered at base station S. Transmission power P_S and noise power σ^2 are set to be 1 W and -100 dBm, respectively. We consider Rayleigh fading channels and the path loss exponent β is set to be 4. The power allocation coefficients of far and near users of queue Q_{L}, $(\alpha_{\mathrm{f}}^2, \alpha_{\mathrm{n}}^2)$, are set to be $(0.8, 0.2)$.

Fig. 3 shows the impact of threshold θ on the stable throughput region of the opportunistic NOMA system when $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$ and $\Gamma_{\mathrm{th}} = 2$. The stable throughput region of opportunistic NOMA is the union of that of dominant systems Φ_1^{ON} and Φ_2^{ON}, given by (19) and (21), respectively. When threshold $\theta = 2\rho = 2\Gamma_{\mathrm{th}}\sigma^2/P_S$, the achievable λ_{L} in dominant system Φ_1^{ON} is much larger than that in OMA system Φ^{OMA}, while the maximum achievable λ_{H} in dominant system Φ_1^{ON} is smaller than that in OMA system Φ^{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 λ_{H} in dominant system Φ_1^{ON} and OMA system Φ^{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 θ. By further increasing the value of threshold θ, the achievable λ_{L} in dominant systems Φ_1^{ON} and Φ_2^{ON} decreases, as the opportunity to perform NOMA decreases. The opportunistic NOMA system enhances the stable throughput region when compared with the OMA system, i.e., $\mathcal{R}^{\mathrm{ON}} \supset \mathcal{R}^{\mathrm{OMA}}$.

Fig. 4 illustrates the impact of power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$ on the stable throughput region of opportunistic NOMA system Φ^{ON} with parameters $\theta = 2\rho$ and $\Gamma_{\mathrm{th}} = 2$. Stable throughput region $\mathcal{R}_1^{\mathrm{ON}} \cup \mathcal{R}_2^{\mathrm{ON}}$ changes significantly with power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$. When $\left(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2\right) = (0.7, 0.3)$, the achievable λ_{L} in dominant systems Φ_1^{ON} and Φ_2^{ON} is less than that in OMA system Φ^{OMA} when $\lambda_{\mathrm{H}} > 0.16$, as low-priority user $D_m, \forall\, m \in \mathcal{M}$, is bottlenecked by successful decoding of the signal intended for high-priority user D_0, which is the prerequisite of performing SIC. By increasing α_{H}^2, the maximum achievable λ_{H} and λ_{L} in dominant systems Φ_1^{ON} and Φ_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 λ_{L} in dominant system Φ_1^{ON} is much greater than that in OMA system Φ^{OMA}. By appropriately selecting the power allocation coefficients, the stable throughput region of opportunistic NOMA system Φ^{ON} can always be larger than that of the OMA system Φ^{OMA}.

Fig. 5 plots the impact of reception threshold Γ_{th} on the stable throughput region of opportunistic NOMA system Φ^{ON} when $\theta = 2\rho$ and $\left(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2\right) = (0.8, 0.2)$. With a decrease of reception threshold Γ_{th}, the maximum achievable λ_{H} and λ_{L} in both opportunistic NOMA system Φ^{ON} and OMA system Φ^{OMA} increase, 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 Φ^{ON} and OMA system Φ^{OMA}

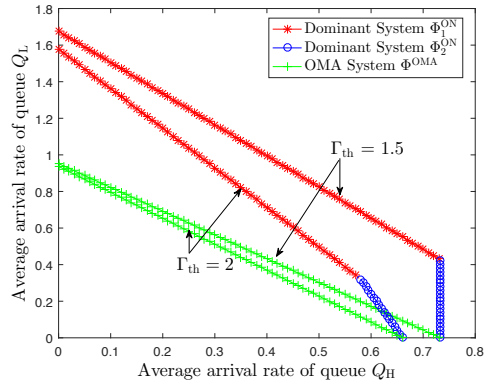

Fig. 5: Stable throughput region with different values of reception threshold Γ_{th} and parameters $\theta = 2\rho$ and $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$.

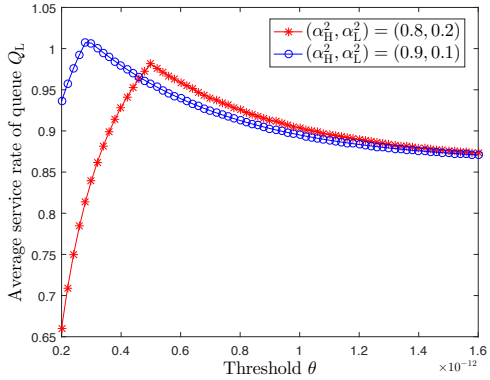

Fig. 6: Average service rate of queue Q_{L} in opportunistic NOMA versus threshold θ and power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$ when $\lambda_{\mathrm{H}} = 0.3$ packets/time slot and $\Gamma_{\mathrm{th}} = 2$.

becomes larger when reception threshold Γ_{th} is smaller.

Fig. 6 shows the impact of threshold θ and power allocation coefficients $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2)$ on the average service rate of queue Q_{L} when $\lambda_{\mathrm{H}} = 0.3$ and $\Gamma_{\mathrm{th}} = 2$. With the variation of threshold θ, 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 $(\alpha_{\mathrm{H}}^2, \alpha_{\mathrm{L}}^2) = (0.8, 0.2)$, the average service rate of queue Q_{L} increases with θ 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 θ when $\theta > 0.5 \times 10^{-12}$ and converges to 0.87, as the probability of enabling NOMA becomes smaller. By increasing α_{H}^2 to 0.9, the optimal threshold θ 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 θ and transmission power allocation coefficients to maximize the stable throughput region of the proposed opportunistic NOMA scheme.

References

[1] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE PIMRC*, London, UK, Sept. 2013.

[2] V. W.S. Wong, R. Schober, D. W. K. Ng, and L. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[3] A. Benjebbovu, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Globecom*, Atlanta, GA, Dec. 2013.

[4] 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.

[5] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[6] 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.

[7] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[8] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393–1405, Apr. 2016.

[9] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Non-orthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 152–10 157, Dec. 2016.

[10] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Stable throughput of cognitive radios with and without relaying capability," *IEEE Trans. Commun.*, vol. 55, no. 12, pp. 2351–2360, Dec. 2007.

[11] S. Kompella, G. D. Nguyen, C. Kam, J. E. Wieselthier, and A. Ephremides, "Cooperation in cognitive underlay networks: Stable throughput tradeoffs," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 1756–1768, Dec. 2014.

[12] P. H. Nardelli, M. Kountouris, P. Cardieri, and M. Latva-Aho, "Throughput optimization in wireless networks under stability and packet loss constraints," *IEEE Trans. Mobile Comput.*, vol. 13, no. 8, pp. 1883–1895, Aug. 2014.

[13] Y. Zhou and W. Zhuang, "Performance analysis of cooperative communication in decentralized wireless networks with unsaturated traffic," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3518–3530, May 2016.

[14] W. Luo and A. Ephremides, "Stability of N interacting queues in random-access systems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1579–1587, Jul. 1999.

[15] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," in *Proc. Camb. Philos. Soc.*, vol. 58, no. 3, pp. 497-520, 1962.

[16] "Study on downlink multiuser superposition transmission (MUST) for LTE," 3GPP TR 36.859, Tech. Rep., Jan. 2016.

[17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. Seventh Edition, Academic Press, 2014.

[18] S. Srinivasa and M. Haenggi, "Distance distributions in finite uniformly random networks: Theory and applications," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 940–949, Feb. 2010.