NOMA in the Uplink: Delay Analysis with Imperfect CSI and Finite-Length Coding

Sebastian Schiessl, *Member, IEEE,* Mikael Skoglund, *Fellow, IEEE,* and James Gross, *Senior Member, IEEE*

***Abstract*—We study whether using non-orthogonal multiple access (NOMA) in the uplink of a mobile network can reduce the queueing delay compared to orthogonal multiple access (OMA) when the system requires communications at very low latency and high reliability. We first consider an ideal system model with perfect channel state information (CSI) at the transmitter and long codewords, where we determine the optimal decoding orders when the decoder uses successive interference cancellation (SIC) and derive closed-form expressions for the optimal rate when joint decoding is used. While joint decoding performs well even under tight delay constraints, NOMA with SIC decoding often performs worse than OMA. For low-latency systems, we must also consider the impact of finite-length channel coding, as well as rate adaptation based imperfect CSI. We derive closed-form approximations for the corresponding outage or error probabilities and find that those effects create a larger performance penalty for NOMA than for OMA. Thus, NOMA with SIC decoding may often be unsuitable for low-latency systems.**

***Index Terms*—Nonorthogonal multiple access (NOMA), stochastic network calculus, effective capacity, quality of service, delay performance, URLLC, imperfect CSI, finite blocklength regime**

I. Introduction

Non-orthogonal multiple access (NOMA) is considered a viable solution for 5G systems due to its increased spectral efficiency over conventional orthogonal multiple access (OMA). In NOMA, multiple users send data simultaneously in the uplink to the base station. The signals create mutual interference, but the base station can employ successive interference cancellation (SIC), i.e., decode one of the signals, and then subtract the corresponding codeword from the received signal, such that the other signal is interference-free. As a result, NOMA can increase the sum ergodic capacity of the system [1].

However, the ergodic capacity is not a meaningful performance metric for applications that require very low latency and high reliablity. For example, industrial control systems often require latencies of at most a few milliseconds. The probability of violating this deadline must be very small, with target values of 10^{-6} and below [2]. In contrast to the ergodic sum capacity, the delay violation probability is affected by the SIC decoding order: the user that is decoded first faces interference by the second user. This interference reduces the data rate achievable by the user that is decoded first. A lower rate can mean that this user's data cannot be transmitted immediately and must be buffered, which leads to a queueing delay. With the reversed decoding order, the user would not suffer from interference, and would therefore experience higher data rate and lower delay. The queueing delay of the two-user NOMA uplink can therefore depend on the choice of decoding orders. Furthermore, if the SNR of both signals is high, then one of the users will always suffer from high interference, so that there is always one user with a very low rate when using SIC decoding. We need to consider a more general joint decoding scheme that can also achieve intermediate rate points where both users transmit at a relatively high rate.

The above discussion on the rate adaptation assumed that the base station has perfect knowledge of the channel state of both users, and that the users can communicate without errors at a rate equal to the capacity of the channel. However, these assumptions become highly inaccurate for low-latency systems. In low-latency systems, the training sequences used for channel estimation are short, so that the base station will only have imperfect channel state information (CSI) for both users. However, the base station must select the data rates based on the CSI. In case of imperfect CSI, the selected rates can be too high, and decoding errors may occur. Furthermore, due to finite blocklength effects, there is always a non-zero probability of decoding errors [3]. Decoding errors are particularly harmful for NOMA systems with SIC decoding due to error propagation: a decoding error of the first user's signal will prevent the removal of that signal's interference and will therefore also cause a decoding error of the second user's signal. This aspect further complicates the selection of rates and decoding orders in low-latency NOMA systems. In order to determine the delay performance of NOMA systems in the presence of decoding errors and error propagation, one must first analyze the decoding error probabilities due to imperfect CSI and finite blocklength channel coding.

A. Related Work

This work combines an analysis of the queueing delay on the link layer of a wireless system with physical layer transmissions based on NOMA. The queueing delay of wireless systems in fading channels can be analyzed using frameworks such as effective capacity [4] or stochastic network calculus [5], [6]. While we consider in this work only stochastic network calculus, the resulting expressions can also be used to derive the effective capacity. We now discuss the state of the art on the delay performance of NOMA systems, as well as on the queueing performance of systems that are subject to imperfect CSI and finite blocklength coding.

The authors are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden (e-mail: schiessl@kth.se, skoglund@kth.se, james.gross@ee.kth.se).

1) NOMA: Several authors have studied the use of NOMA in wireless systems due to its improved spectral efficiency. Some of these works, e.g., [1], [7] study only downlink transmissions, whereas other works [8]–[12] consider uplink transmissions. For both downlink and uplink transmissions, the receivers employ SIC as the decoding principle. Nevertheless, there are some important differences between downlink and uplink transmissions. In the downlink, the decoding order is fixed. The user with the stronger channel must first decode the message intended for the weaker user, subtract the corresponding signal, and can then decode its own message. The weaker user will only be able to decode its own message. Contrary to that, the decoding order in uplink NOMA can be chosen arbitrarily. Many related works on uplink NOMA [8]–[12] suggest that the signal from the stronger user should be decoded first, such that the weaker user's signal is interference-free. Nevertheless, it may be beneficial to deviate from this suggested decoding order to improve the delay performance for one of the users.

With respect to low-latency communications, several authors have considered the use of NOMA. Yang et al. [7] optimize the sum rate subject to constraints on the minimum rates of each user. Similarly, Timotheou et al. [13] consider the max-min fairness for the individual rates, i.e., maximize the minimum rate among all users. However, the minimum rate may not be a meaningful metric for fading channels, where one of the rates may occasionally become very small. In [8] and [14], the outage probability in NOMA systems is analyzed. The outage probability is a meaningful metric for systems where the data rate is kept constant. However, not adapting the data rate to the channel state is generally suboptimal.

More specifically with respect to queueing analysis, several authors have considered NOMA. Choi [15] studied the effective capacity of NOMA systems, assuming that one of the users is always decoded first (i.e., always suffers from interference). Similarly, in our previous work [16], we considered the impact of interference on the queueing performance, which corresponds to the performance of the user that is always decoded first. The queuing performance of downlink NOMA systems was studied by Yu et al. [17] (using effective capacity) and Xiao et al. [18] (using stochastic network calculus). Both [17] and [18] only considered downlink NOMA, where the decoding order is fixed and does not need to be optimized. Close to our work is the work by Qiao et al. [19], where the decoding order of the users in the uplink was varied based on the instantaneous channel states and on the individual QoS parameters. However, the authors assumed CSI to be perfect, and only considered SIC decoding, which may yield suboptimal rate points compared to a more general joint decoder.

2) Imperfect CSI and Finite-Length Coding: Polyanskiy et al. [3] have studied bounds on the decoding error probability of finite-length codes and also presented a closed-form normal approximation. Yang et al. [20] extended these results to quasi-static block-fading channels. Scarlett et al. [21] studied finite blocklength effects in a multiuser scenario. MolavianJazi [22] studied the achievable coding rates with finite-length codewords in the two-user NOMA case. Sun et al. [23] analyzed a downlink NOMA scenario with finite-length codes, but assumed CSI to be perfect and only studied the average throughput, i.e., did not consider queueing.

We are not aware of any results on the queueing performance of NOMA with finite-length codes. For single-user systems with finite-length codes, the queueing delay was studied in [24], [25]. In [26], we have studied the joint impact of imperfect CSI and finite blocklength effects on the delay performance of wireless systems, assuming a single-user single-antenna scenario. Furthermore, we studied the impact of these effects in a more general multiuser multi-antenna downlink scenario in [27], where beamforming was applied to avoid interference.

B. Contributions

In this work, we apply the framework of stochastic network calculus (SNC) [5], [6] to study the queueing delay of the two-user NOMA uplink. Our main contributions are:

- For perfect CSI and SIC decoding, we identify the problem of optimal rate selection (i.e., optimal decoding order) for quantized SNR distributions as a 0-1 knapsack problem that can be solved with a greedy algorithm.
- For perfect CSI and a more general joint decoder, we determine the optimal rate adaptation function in closed form.
- For imperfect CSI and SIC decoding, we derive closed-form approximations for the decoding error probabilities. This allows us to determine optimal rate allocations for this case. Simulation results show that the approximations are sufficiently accurate.
- Using methods from prior work, we present closed-form approximations for the decoding error probabilities under imperfect CSI and finite blocklength coding.
- Our numerical study shows that under ideal assumptions with perfect CSI, NOMA with joint decoding significantly outperforms OMA in terms of queueing delay when there is a large difference between the two users' average channels. NOMA with SIC decoding performs significantly worse, and may be worse than OMA, depending on the parameters.
- Imperfect CSI and finite blocklength effects cause a significant performance loss. Our results indicate that the performance loss is slightly larger for the NOMA than for OMA.

This paper is structured as follows: For the ideal model with perfect CSI and very long codewords, we present the system model in Sec. II and the delay analysis and rate optimization in Sec. III. We then analyze imperfect CSI in Sec. IV, and finite blocklength coding in Sec. V. Numerical results are given in Sec. VI. We present our conclusions in Sec. VII.

II. System Model

We analyze data transmissions in a multiple-access channel (MAC) where two devices send data packets to a central base station in a time-slotted fashion. We consider applications that generate periodic and time-critical data at the device/user side, which should be transmitted to the base station within

a short deadline of w time slots with high reliability. In Sec. II-A, we discuss the data transmission on the physical layer, where we consider only an ideal model with perfect CSI and infinitely long channel codes. Later, in Sec. IV and V, we will consider more realistic physical layer models with imperfect channel estimation and finite blocklength coding. Due to time-varying data rates and transmission errors at the physical layer, the devices must keep their data in a buffer for transmission in subsequent time slots. The resulting queueing delay is described in Sec. II-B. We conclude this section with the problem statement in Sec. II-C.

A. Physical Layer Model

The channel is assumed to be block-fading, i.e., remains constant for the duration of one block or time slot of n channel uses, and changes independently between time slots. We now consider a single time slot. For each channel use, the received signal y is denoted as

$$y = h_1 x_1 + h_2 x_2 + z \tag{1}$$

where z is additive white Gaussian noise. Without loss of generality, we assume that $z \sim \mathcal{CN}(0,1)$ (unit variance) and that $\mathbb{E}\left[|h_k|^2\right] = 1$, such that the average power $\mathbb{E}\left[|x_k|^2\right] = \bar{\gamma}_k$ of the transmitted code symbols corresponds to the *average* SNR of the signal at the receiver in case there is no interference. We assume that the transmit power remains fixed over time. The instantaneous signal-to-noise ratio (SNR) of the received signal of user k is denoted as $\gamma_k = \bar{\gamma}_k |h_k|^2$ and changes along with the fading coefficient h_k from time slot to time slot. We assume Rayleigh-fading with $h_k \sim \mathcal{CN}(0,1)$. For the initial analysis, we assume that the instantaneous SNR values γ_k are perfectly known at transmitting devices and at the base station, so that the entire time slot can be used for transmitting codewords $\mathbf{x}_k$ of length $n_\mathrm{d} = n$ (for imperfect CSI, see Sec. IV). Furthermore, we assume that n_d is sufficiently large so that error-free communication at a rate equal to the capacity is possible, using Gaussian codewords $\mathbf{x}_k$ (we discuss finite-length coding in Sec. V). The base station can try to decode the signals through successive interference cancellation (SIC) or jointly. For SIC decoding, assume that codeword x_1 is decoded first. This is possible if the rate r_1 of the channel code for user 1 satisfies

$$r_1 < c_1^{\mathsf{min}}(\gamma_1, \gamma_2) \triangleq \log_2\left(1 + \frac{\gamma_1}{\gamma_2 + 1}\right). \tag{2}$$

After successfully decoding the signal sent by user 1, the base station reconstructs the codeword x_1 and subtracts $h_1 x_1$ from the received signal y. Then, signal 2 can be decoded if the rate r_2 of the channel code for user 2 satisfies

$$r_2 < c_2^{\mathsf{max}}(\gamma_2) \triangleq \log_2\left(1 + \gamma_2\right). \tag{3}$$

The decoding order can also be reversed, such that x_2 is decoded first and then subtracted.

By using a decoder that decodes x_1 and x_2 jointly, the base station can decode both x_1 and x_2 whenever the rates r_1 and r_2 are inside the capacity region, which is given as [28]:

$$r_1 < c_1^{\mathsf{max}}(\gamma_1) = \log_2(1 + \gamma_1) \tag{4}$$

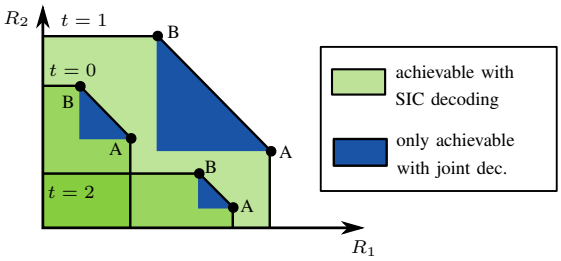


Fig. 1. Schematic illustration of 3 random samples of the capacity region and achievable rate pairs for NOMA.

$$r_2 < c_2^{\mathsf{max}}(\gamma_2) = \log_2(1 + \gamma_2) \tag{5}$$

$$r_1 + r_2 < c_\Sigma(\gamma_1, \gamma_2) \triangleq \log_2(1 + \gamma_1 + \gamma_2). \tag{6}$$

The capacity regions for three time slots, i.e., three random instances of γ_1 and γ_2, are illustrated in Fig. 1. The capacity region always has the shape of a pentagon, with two of the corner points (A and B) corresponding to the maximum rates achieved by the SIC decoder.[1] With joint decoding, one can also achieve the rate pairs on the segment $\overline{\mathrm{AB}}$ between the two corner points. We note here that joint decoding in the information-theoretic sense would mean that the decoder compares the received signal to every combination of the $2^{n_\mathrm{d} r_1}$ and $2^{n_\mathrm{d} r_2}$ possible codewords, which is prohibitively complex for any practical value of n_d. Nevertheless, the points on the segment between the two SIC points can also be achieved through SIC with time sharing between the two decoding orders, or through a rate-splitting approach [29], for which one can employ channel codes like LDPC or turbo codes. SIC requires two decoding steps and is therefore roughly twice as complex as conventional OMA, while rate-splitting requires three decoding steps. Therefore, the rate points on the segment $\overline{\mathrm{AB}}$ can be achieved with practical schemes that have reasonable computational complexity. However, for better clarity, we still apply the term "joint decoding" ("NOMA-joint") to the case where all rate points on $\overline{\mathrm{AB}}$ can be selected, in order to avoid confusion with the case where only the corner points of the rate regions can be selected, which will be exclusively denoted by "SIC" ("NOMA-SIC").

We assume for now that in each time slot, the base station knows the instantaneous SNR values (γ_1, γ_2) perfectly and selects an achievable rate pair, which is denoted as $r_1 = \Phi_1(\gamma_1, \gamma_2)$ and $r_2 = \Phi_2(\gamma_1, \gamma_2)$, or $(r_1, r_2) = \mathbf{\Phi}(\gamma_1, \gamma_2)$ for short. There is no reason to select a rate pair that is below the maximally achievable rates, so the base station will select a pair of rates (r_1, r_2) that either corresponds to one of the corner points A/B (in case of SIC decoding), or to any point on the segment $\overline{\mathrm{AB}}$ (in case of joint decoding). The selected rates are then signaled to the users through a feedback link, which we assume to be instantaneous and error-free.

B. Queueing Model

Due to the time-varying channel conditions, the data arriving at the users must be stored in a buffer until successful

[1]The rates in (2) to (6) can be chosen arbitrarily close to the capacity or capacity region. In order to simplify discussions, we assume that (2) to (6) also hold when the rates are equal to the capacity or on the boundary of the capacity region.

transmission. This leads to a random queueing delay. A queueing system is described by its arrival, service, and departure processes. The arrival process $A_k(t)$ describes the amount of data in bits that arrives and is put into the buffer at user k in time slot t. The service process $S_k(t)$ depends on the instantaneous SNR values $\gamma_1(t)$ and $\gamma_2(t)$ in time slot t, which change from time slot to time slot. The rate is determined by the rate adaptation scheme as $r_k(t) = \Phi_k(\gamma_1(t), \gamma_2(t))$. In case the transmission is successful, the service is $S_k(t) = n_{\mathrm{d}} r_k(t)$. In case of a transmission error (which will become relevant in Sec. IV and Sec. V), the base station will indicate the error event through a feedback bit so that the users will not remove the corresponding data from the queue, which corresponds to $S_k(t) = 0$. Therefore, in case of an error, the data remain in the queue and will be retransmitted in the following time slots. The departure process $D_k(t)$ describes the data that is actually transmitted over the wireless channel, which is the minimum of $S_k(t)$ and the amount of data waiting in the buffer. The virtual delay $W_k(t)$ of the data arriving in time slot t at user k is then defined as [6]

$$W_k(t) \triangleq \inf \left\{ u \geq 0 : \quad \sum_{i=0}^{t-1} A_k(i) \leq \sum_{i=0}^{t+u-1} D_k(i) \right\} \tag{7}$$

The delay $W_k(t)$ is a random variable, and its distribution can be stated in terms of the delay violation probability over all time slots t with respect to a certain deadline w:

$$p_{\mathrm{v},k}(w) \triangleq \sup_{t \geq 0} \mathbb{P}\left\{ W_k(t) > w \right\} \tag{8}$$

The queueing delay $W_k(t)$ only describes the random number of time slots during which the data will remain in the queue. The system will experience additional delays, e.g., due to encoding at the transmitter and signal processing at the receiver. Due to differences in signal processing complexity, those delays might differ between NOMA-joint, NOMA-SIC, and OMA, but the differences depend on numerous factors, e.g., on the chosen hardware, and cannot be quantified in a general fashion. However, the additional delays are often bounded and short compared to $W_k(t)$. For example, signal processing and decoding must always be finished before a new packet is received, so the corresponding delay must be shorter than one time slot. The worst-case decoding delay could then be easily taken into account by adjusting the deadline w by one time slot. To simplify discussions, we ignore those minor or constant delays and restrict our analysis to the random queueing delay $W_k(t)$ and the corresponding delay violation probability $p_{\mathrm{v},k}(w)$.

C. Problem Statement

In this work, we study whether NOMA systems can provide better delay performance, i.e., lower delay violation probabilities, compared to OMA. The delay performance of NOMA depends on the optimal rate allocation scheme $\mathbf{\Phi}$. In order to ensure reliable low-latency communications for both users, we need to determine $\mathbf{\Phi}$ such that the delay violation probabilities $p_{\mathrm{v},k}(w)$ for both users are jointly minimized. We first consider the ideal model with perfect CSI, and find the optimal rate allocation for SIC decoding and for joint decoding. However, in more realistic models with imperfect CSI and finite block-length codes, rate adaptation is more difficult, as we have to find an optimal trade-off between the selected rates and the corresponding decoding error probabilities. We can then address the following questions: can NOMA provide lower delay violation probabilities than OMA also under realistic system models? And how large is the difference between SIC decoding and joint decoding in that case?

III. Analysis – Ideal Case

In this section, we will first present in Sec. III-A the analytical upper bound on the delay violation probability from stochastic network calculus. Then, we show in Sec. III-B how this result can be used to reformulate the problem statement analytically as an optimization problem. We will then solve the optimal rate adaptation problem for SIC decoding and joint decoding in Sec. III-C and Sec. III-D, respectively. In order to introduce the basic optimization methodology, we consider in this section only the ideal system model from Sec. II-A. However, we are ultimately interested in the delay performance for a more realistic system model with imperfect CSI and finite-length coding, which we will introduce in Sec. IV and V.

A. Stochastic Network Calculus

We now give a brief summary of previous results from stochastic network calculus (SNC). Specifically, we show an upper bound from SNC on the delay violation probability $p_{\mathrm{v},k}(w)$ in (8) [5], [6]. This summary closely follows the summary given in our previous works [25]–[27].

We follow [6], where SNC is applied in a transform domain, also referred to as *SNR-domain*. The bit-domain arrival and service processes $A_k(t)$ and $S_k(t)$ defined in Sec. II-B are transformed to the SNR-domain via the exponential function: $\mathcal{A}_k(t) \triangleq e^{A_k(t)}$ and $\mathcal{S}_k(t) \triangleq e^{S_k(t)}$. An upper bound on the delay violation probability $p_{\mathrm{v},k}(w)$ can then be computed in terms of the Mellin transforms of $\mathcal{A}_k(t)$ and $\mathcal{S}_k(t)$, where we can omit the time index t because of i.i.d. arrivals and block-fading. The Mellin transform of a nonnegative random variable $\mathcal{X}$ is defined as [6]

$$\mathcal{M}_{\mathcal{X}}(\theta) \triangleq \mathbb{E}\left[\mathcal{X}^{\theta-1}\right] \tag{9}$$

for a parameter $\theta \in \mathbb{R}$. For the analysis, we always choose $\theta_k > 0$ and first check whether the stability condition $\mathcal{M}_{\mathcal{A}_k}(1+\theta_k)\mathcal{M}_{\mathcal{S}_k}(1-\theta_k) < 1$ holds. If it holds, define the kernel [6], [25]

$$\mathcal{K}_k\left(\theta_k, w\right) \triangleq \frac{\mathcal{M}_{\mathcal{S}_k}(1-\theta_k)^w}{1 - \mathcal{M}_{\mathcal{A}_k}(1+\theta_k)\mathcal{M}_{\mathcal{S}_k}(1-\theta_k)} \,. \tag{10}$$

This kernel is strictly monotonically increasing in both $\mathcal{M}_{\mathcal{A}_k}(1+\theta_k)$ and $\mathcal{M}_{\mathcal{S}_k}(1-\theta_k)$, and provides an upper bound for the delay violation probability, which holds for any time slot t, including the limit $t \to \infty$ (steady-state):

$$p_{\mathrm{v},k}(w) \leq \inf_{\theta_k > 0} \left\{ \mathcal{K}_k\left(\theta_k, w\right) \right\} \,. \tag{11}$$

The above inequality already takes into account that $\mathcal{K}_k(\theta_k, w)$ is an upper bound on $p_{\mathrm{v},k}(w)$ for any parameter $\theta_k > 0$. Thus, one should take the infimum over θ_k in order to find the tightest upper bound on $p_{\mathrm{v},k}(w)$.

B. Rate Allocation Problem

We seek to determine a rate allocation scheme $\mathbf{\Phi}$ which jointly minimizes the delay violation probabilities $p_{\mathrm{v},k}(w)$ in (8) of the two users $k \in \{1,2\}$. Specifically, we iterate over different constraints on $p_{\mathrm{v},1}(w)$ and then minimize $p_{\mathrm{v},2}(w)$ subject to each specific constraint. In order to work with analytical expressions of the system, we use the analytical upper bound (11) on $p_{\mathrm{v},k}(w)$ based on the kernels $\mathcal{K}_k(\theta_k, w)$. Then, the optimization of the rate allocation scheme $\mathbf{\Phi}$ can be formulated as follows: given a specific QoS constraint δ for the first user, how should the base station select the rates such that the bound on $p_{\mathrm{v},k}(w)$ for the second user is minimized? This can be formulated as

$$\begin{aligned} \arg\min_{\mathbf{\Phi}} \quad & \inf_{\theta_2>0} \{\mathcal{K}_2(\theta_2, w)\} \\ \text{s.t.} \quad & \inf_{\theta_1>0} \{\mathcal{K}_1(\theta_1, w)\} \leq \delta \end{aligned} \tag{P.I}$$

The kernels depend on the rate adaptation function $\mathbf{\Phi}$ through the Mellin transform of the SNR-domain service process $\mathcal{M}_{\mathcal{S}_k}(1-\theta_k) = \mathbb{E}\left[e^{-\theta_k n \Phi_k(\gamma_1,\gamma_2)}\right]$. We iterate over all possible combinations of θ_k and solve the problem for a specific choice of the values θ_k. By inspecting (10), we can deduce that the kernel $\mathcal{K}_k(\theta_k, w)$ is monotonically increasing in $\mathcal{M}_{\mathcal{S}_k}(1-\theta_k)$, as long as the stability condition $\mathcal{M}_{\mathcal{A}_k}(1+\theta_k)\mathcal{M}_{\mathcal{S}_k}(1-\theta_k) < 1$ is satisfied. Therefore, the optimization problem can be formulated directly in terms of $\mathcal{M}_{\mathcal{S}_k}(1-\theta_k)$ instead of $\mathcal{K}_k(\theta_k, w)$:

$$\begin{aligned} \arg\min_{\mathbf{\Phi}} \quad & \mathcal{M}_{\mathcal{S}_2}(1-\theta_2) \\ \text{s.t.} \quad & \mathcal{M}_{\mathcal{S}_1}(1-\theta_1) \leq \delta_{\mathrm{M}} \end{aligned} \tag{P.II}$$

with δ_{M} chosen such that $\mathcal{K}_1(\theta_1, w) \leq \delta$.

C. Rate Allocation for SIC

In case the base station employs SIC decoding, the rate allocation can be found as follows:

Result 1. *When using SIC decoding, i.e., when the rate scheduler can only decide between the two rate pairs* $(c_1^{\mathsf{max}}, c_2^{\mathsf{min}})$ *or* $(c_1^{\mathsf{min}}, c_2^{\mathsf{max}})$*, the optimal solution to problem* (P.II) *is given by*

$$\Phi_1(\gamma_1,\gamma_2) = \begin{cases} c_1^{\mathsf{min}}(\gamma_1,\gamma_2) & \text{for } \eta(\gamma_1,\gamma_2) > \lambda \\ c_1^{\mathsf{max}}(\gamma_1) & \text{otherwise} \end{cases} \tag{12}$$

$$\Phi_2(\gamma_1,\gamma_2) = c_\Sigma(\gamma_1,\gamma_2) - \Phi_1(\gamma_1,\gamma_2) \tag{13}$$

with value-to-weight ratio

$$\eta(\gamma_1,\gamma_2) = \frac{e^{-\theta_2 n c_2^{\mathsf{min}}(\gamma_1,\gamma_2)} - e^{-\theta_2 n c_2^{\mathsf{max}}(\gamma_2)}}{e^{-\theta_1 n c_1^{\mathsf{min}}(\gamma_1,\gamma_2)} - e^{-\theta_1 n c_1^{\mathsf{max}}(\gamma_1)}} \tag{14}$$

and $\lambda > 0$ *is the smallest value such that* $\mathcal{M}_{\mathcal{S}_1}(1-\theta_1) \leq \delta_{\mathrm{M}}$ *is still satisfied.*

Proof. We discretize the individual distributions of γ_1 and γ_2 to N_1 and N_2 probability mass points, respectively, such that the joint distribution of γ_1 and γ_2 is described by discrete points $(\gamma_{1,i}, \gamma_{2,i})$, $i = 1 \ldots N$, where $N = N_1 N_2$. The probability mass of each point is denoted as p_i. We then rewrite the optimization problem in terms of $x_i \in \{0,1\}$, where $x_i = 0$ means that decoding order A is selected ($r_{1,i} = c_{1,i}^{\mathsf{max}}$ and $r_{2,i} = c_{2,i}^{\mathsf{min}}$), and $x_i = 1$ means that point B is selected ($r_{1,i} = c_{1,i}^{\mathsf{min}}$ and $r_{2,i} = c_{2,i}^{\mathsf{max}}$):

$$\begin{aligned} \arg\min_{x_i\in\{0,1\}} \quad & \sum_i p_i e^{-\theta_2 n\left(c_{2,i}^{\mathsf{min}} + x_i\left(c_{2,i}^{\mathsf{max}} - c_{2,i}^{\mathsf{min}}\right)\right)} \\ \text{s.t.} \quad & \sum_i p_i e^{-\theta_1 n\left(c_{1,i}^{\mathsf{max}} + x_i\left(c_{1,i}^{\mathsf{min}} - c_{1,i}^{\mathsf{max}}\right)\right)} \leq \delta_{\mathrm{M}} \end{aligned} \tag{P.IIa}$$

When each x_i can only take on one of two possible values $\{0,1\}$, then each term in the objective function can only take on one of two possible values. Let us denote the two possible values as $a_i = p_i e^{-\theta_2 n c_{2,i}^{\mathsf{min}}}$ when $x_i = 0$ and $b_i = p_i e^{-\theta_2 n c_{2,i}^{\mathsf{max}}}$ when $x_i = 1$. The objective function is then equivalent to $\sum_i (a_i + x_i(b_i - a_i))$. The same method is applied to the constraint function. By taking the negative of the objective function, the minimization problem (P.IIa) can be converted into the following equivalent maximization problem:

$$\begin{aligned} \arg\max_{x_i\in\{0,1\}} \quad & z = \sum_i x_i v_i \\ \text{s.t.} \quad & \sum_i x_i w_i \leq \tilde{\delta} \end{aligned} \tag{P.IIb}$$

with

$$v_i = p_i\left(e^{-\theta_2 n c_{2,i}^{\mathsf{min}}} - e^{-\theta_2 n c_{2,i}^{\mathsf{max}}}\right) \tag{15}$$

$$w_i = p_i\left(e^{-\theta_1 n c_{1,i}^{\mathsf{min}}} - e^{-\theta_1 n c_{1,i}^{\mathsf{max}}}\right) \tag{16}$$

$$\tilde{\delta} = \delta_{\mathrm{M}} - \sum_i p_i e^{-\theta_1 n c_{1,i}^{\mathsf{max}}} \tag{17}$$

We identify this optimization problem as the well-known 0-1 knapsack problem of selecting items with value $v_i \geq 0$ and weight $w_i \geq 0$ subject to a weight limit $\tilde{\delta}$. We assume $0 < \tilde{\delta} < \sum_i w_i$, otherwise the problem is either trivial or infeasible. To solve this problem, we first allow x_i in (P.IIb) to vary continuously from 0 to 1, i.e., we relax the problem [30].[2] The continuous problem (CP) can be easily solved using a greedy algorithm, and provides an approximate solution to the discrete problem. Specifically, Dantzig [31] showed that the optimal solution to CP can be found by ordering the items decreasingly according to their value-to-weight ratios $\eta_i = v_i/w_i$ and then selecting the first $j-1$ items with the highest value-to-weight ratios ($x_i = 1$ for $i = 1, \ldots, j-1$, after reordering) such that $\sum_{i=1}^{j-1} x_i w_i \leq \tilde{\delta}$ is still satisfied. Then, for the j-th item, only a fractional value of $0 \leq x_j < 1$ is chosen, such that $\sum_{i=1}^{j} x_i w_i$ becomes equal to δ_{M}. The remaining items are not selected. Clearly, the optimal value z^* of problem (P.IIb) cannot exceed the optimal value z^*_{CP} of

[2]The problem (P.IIb) is only equivalent to (P.IIa) when x_k is integer, and relaxing (P.IIb) is thus *not* equivalent to relaxing the integer constraint in (P.IIa), i.e., not equivalent to allowing rates between the extreme points A and B.

the continuous problem: $z^* \leq z^*_{\mathsf{CP}}$. Furthermore, we obtain a rounded solution by setting $x_j = 0$, with a corresponding value of $z^*_{\lfloor \mathsf{CP} \rfloor}$. The rounded solution is a possible solution to (P.IIb), thus $z^*_{\lfloor \mathsf{CP} \rfloor} \leq z^*$. As $z^*_{\mathsf{CP}} - z^*_{\lfloor \mathsf{CP} \rfloor} = x_j v_j$, with v_j vanishing as the quantization intervals and the corresponding probability masses p_i tend to zero, the rounded/approximate solution converges to the optimal solution. The approximate solution selects $x_i = 1$ for the first $j-1$ items when the items are sorted by descending η_i. This means it will select $x_i = 1$ whenever $\eta_i > \lambda$, where $\lambda > 0$ must be chosen such that the optimization constraint $\sum_i x_i w_i \leq \tilde{\delta}$ is still satisfied. The proof is completed by noting that $x_i = 1$ means decoding order B and by considering continuous values (γ_1, γ_2) instead of quantized values.

□

Remark 1. *The problem above was already addressed in [19], which provided a function that denotes the boundary of the two regions where user 1 or user 2 is decoded first, respectively. However, the proof relies on a specific result from variational calculus, which requires that the end points of the function are fixed and known, which may not be the case.*

Remark 2. *The formulation of the problem as a 0-1 knapsack problem only applies to the two-user scenario. With K users transmitting simultaneously, the base station would need to select one out of $K!$ possible decoding orders, resulting in an integer programming problem that cannot be solved in the same way. However, the system can divide the K users into $\lceil K/2 \rceil$ groups of at most 2 users each and schedule the groups on orthogonal time/frequency resources.*

D. Rate Allocation for Joint Decoding

With SIC decoding, only the rate pairs on the two corner points of the capacity region are achievable. On the other hand, joint decoding allows rate points in between the corner points.

Result 2. *Under joint decoding, i.e., when all rates in the achievable rate region can be selected, the optimal solution to the rate adaptation problem* (P.II) *is given by*

$$\Phi_1(\gamma_1,\gamma_2) = \begin{cases} c_1^{\mathsf{min}}(\gamma_1,\gamma_2) & \text{if } \widehat{\Phi}_1(\gamma_1,\gamma_2) < c_1^{\mathsf{min}}(\gamma_1,\gamma_2) \\ c_1^{\mathsf{max}}(\gamma_1) & \text{if } \widehat{\Phi}_1(\gamma_1,\gamma_2) > c_1^{\mathsf{max}}(\gamma_1) \\ \widehat{\Phi}_1(\gamma_1,\gamma_2) & \text{otherwise} \end{cases} \tag{18}$$

$$\Phi_2(\gamma_1,\gamma_2) = c_\Sigma(\gamma_1,\gamma_2) - \Phi_1(\gamma_1,\gamma_2)\,, \tag{19}$$

where $\widehat{\Phi}_1(\gamma_1,\gamma_2) = \frac{\theta_2}{\theta_1+\theta_2} c_\Sigma(\gamma_1,\gamma_2) + \tilde{\lambda}_1$ and $\tilde{\lambda}_1 \in \mathbb{R}$ is the smallest value such that $\mathcal{M}_{\mathcal{S}_1}(1-\theta_1) \leq \delta_{\mathrm{M}}$ is still satisfied.

Proof. We quantize the joint distribution of γ_1 and γ_2 to points labeled as $i = 1 \ldots N$ with probability mass p_i. Note that if the sum rate constraint (6) does not hold with equality, then one of the rates could be increased without penalty for the other user. An optimal rate allocation will thus always select a rate pair which satisfies the sum rate constraint. We then have $r_{2,i} = c_{\Sigma,i} - r_{1,i}$ and rewrite the optimization problem:

$$\underset{r_{1,i}}{\arg\min} \quad \sum_i p_i e^{-\theta_2 n (c_{\Sigma,i} - r_{1,i})} \tag{20}$$

$$\text{s.t.} \quad \sum_i p_i e^{-\theta_1 n r_{1,i}} \leq \delta_{\mathrm{M}} \tag{21}$$

$$r_{1,i} \leq \log_2(1+\gamma_{1,i}) \qquad i = 1,\ldots,N \tag{22}$$

$$r_{1,i} \geq \log_2\left(1 + \frac{\gamma_{1,i}}{\gamma_{2,i}+1}\right) \quad i = 1,\ldots,N \tag{23}$$

The problem is convex in $r_{1,i}$. We associate the Lagrange multipliers $\lambda_1 \geq 0$, $\mu_i \geq 0$ and $\nu_i \geq 0$ with the contraints (21), (22), and (23), respectively. According to the Karush-Kuhn-Tucker conditions [32], the optimal rates $r_{1,i}$ for all $i = 1,\ldots,N$ must satisfy

$$\theta_2 n e^{-\theta_2 n (c_{\Sigma,i} - r_{1,i})} - \lambda_1 \theta_1 n e^{-\theta_1 n r_{1,i}} + \mu_i - \nu_i = 0\,. \tag{24}$$

When either of the contraints (22) and (23) is satisfied with equality, then the value of $r_{1,i}$ is known. Otherwise, there is slackness in the constraints (22) and (23), and the complementary slackness conditions [32] mandate that μ_i and ν_i must be zero. It follows for those cases:

$$\lambda_1 \frac{\theta_1}{\theta_2} e^{-\theta_1 n r_{1,i}} = e^{-\theta_2 n (c_{\Sigma,i} - r_{1,i})}\,. \tag{25}$$

With the definition $\tilde{\lambda}_1 = \frac{1}{\theta_1 n + \theta_2 n} \log\left(\lambda_1 \frac{\theta_1}{\theta_2}\right)$, we derive

$$r_{1,i} = \frac{\theta_2}{\theta_1+\theta_2} c_{\Sigma,i} + \tilde{\lambda}_1 \triangleq \widehat{\Phi}_1(\gamma_{1,i},\gamma_{2,i}) \tag{26}$$

For a given $\tilde{\lambda}_1$, the rates can be found by first computing $\widehat{\Phi}_1(\gamma_{1,i},\gamma_{2,i})$ according to (26). For those SNR points i where (26) would violate the constraints (22) or (23), the respective constraint must hold with equality. □

IV. Imperfect CSI

A. System Model and Approximations for Imperfect CSI

In a realistic system, the users do not have perfect knowledge of the channel states. Instead, the channels must be estimated first, and the rates must be chosen based on the imperfect channel estimate. We assume that the two users send known training sequences of length $n_{\mathrm{t},1}$ and $n_{\mathrm{t},2}$ at the beginning of each time slot. The two users send their training sequences orthogonally (i.e., without interference) to the receiver, which obtains the MMSE channel estimates $\hat{h}_k$, $k \in \{1,2\}$. The actual channel coefficients H_k are unknown, i.e., random. Given the MMSE estimates $\hat{h}_k$, the actual channel coefficients are given as [26], [33]

$$H_k = \hat{h}_k + \tilde{H}_k\,, \tag{27}$$

with $\tilde{H}_k \sim \mathcal{CN}(0, \sigma^2_{Z,k})$,

$$\sigma^2_{Z,k} = \frac{1}{1 + \bar{\gamma}_{\mathrm{t},k} n_{\mathrm{t},k}}\,, \tag{28}$$

and $\bar{\gamma}_{\mathrm{t},k}$ denoting the SNR during the training phase of user k. The actual SNR $\Gamma_k = \bar{\gamma}_k |H_k|^2$ of signal k is then given as [26]:

$$\Gamma_k = \bar{\gamma}_k |\hat{h}_k|^2 + 2\bar{\gamma}_k |\hat{h}_k| \Re\left\{e^{-j\angle(\hat{h}_k)} \tilde{H}_k\right\} + \bar{\gamma}_k |\tilde{H}_k|^2\,, \tag{29}$$

where $\Re\{\cdot\}$ and $\angle(\cdot)$ denote the real part and the phase of a complex value, respectively. With sufficient training, the estimation error is relatively small, i.e., $|\tilde{H}_k| \ll |\hat{h}_k|$, and

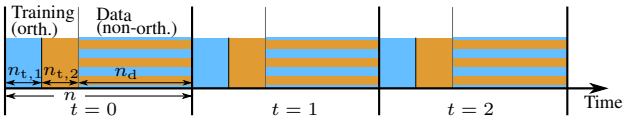


Fig. 2. Illustration of orthogonal training and non-orthogonal data transmission phases in each time slot.

the last term becomes negligible. The term $\Re\{e^{-j\angle(\hat{h}_k)}\tilde{H}_k\}$ is Gaussian distributed with variance $\sigma^2_{Z,k}/2$, so we can approximate the distribution of the SNR Γ_k of signal k as [26]

$$\Gamma_k \sim \mathcal{N}(\hat{\gamma}_k, \sigma_k^2) \tag{30}$$

with $\hat{\gamma}_k = \bar{\gamma}_k|\hat{h}_k|^2$, and

$$\sigma_k^2 = 2\bar{\gamma}_k^2|\hat{h}_k|^2\sigma^2_{Z,k} = 2\bar{\gamma}_k\hat{\gamma}_k\sigma^2_{Z,k}\,. \tag{31}$$

After the channel estimation, the base station must select the pair of rates (r_1, r_2) at which the users should encode their data, based on the channel estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$. The users are informed about the chosen rates through an error-free feedback channel with zero delay.[3] While the rates are chosen based on the estimated SNR, i.e., $(r_1, r_2) = \mathbf{\Phi}(\hat{\gamma}_1, \hat{\gamma}_2)$, the actual SNR of the channels, Γ_1 and Γ_2, is unknown and may be too low, such that decoding of the data sent by the users will fail. The corresponding error probabilities depend on the chosen rates, as well as on the type of decoder used at the base station. In Sec. IV-C, we will derive and approximate the error probabilities for a simple SIC decoder, and in Sec. IV-D, we consider errors under joint decoding. We assume for now that the blocklength n_{d} is very large. In Sec. V, we will also consider the effects of finite blocklength channel coding, i.e., the case where n_{d} is small. It is assumed throughout the paper that rate adaptation is based on imperfect CSI, but that the receiver has perfect CSI when decoding the signal.[4] We note that when $n_{\mathrm{t},1}$ plus $n_{\mathrm{t},2}$ symbols are required for channel estimation in each time slot, only $n_{\mathrm{d}} = n - n_{\mathrm{t},1} - n_{\mathrm{t},2}$ symbols remain for the data transmission. Fig. 2 illustrates the durations of training and data transmission.

B. Rate Optimization

As discussed in Sec. III-B, the optimal rate allocation problem can be formulated as minimizing the Mellin transform $\mathcal{M}_{\mathcal{S}_2}(1-\theta_2)$ of the SNR-domain service process for the second user, subject to a constraint on $\mathcal{M}_{\mathcal{S}_1}(1-\theta_1)$ for the first user. When decoding errors occur with probability ε_k, the Mellin transform of the service process is given as [26]

$$\mathcal{M}_{\mathcal{S}_k}(1-\theta_k) = \mathbb{E}\left[\varepsilon_k + (1-\varepsilon_k)e^{-\theta_k n_{\mathrm{d}} r_k}\right] \tag{32}$$

[3]For the single user case, we found that quantizing the rate to 6 bits is sufficient [26]. It is reasonable to assume that the base station, which can operate at high transmit power, can communicate this small amount of data to the users with very low error probabilities and delays that are negligible compared to those in the uplink transmissions.

[4]Additional pilot symbols sent during the data transmission phase will make the CSI at the receiver almost perfect compared to the imperfect estimate during rate selection. The receiver may also employ joint estimation and decoding [34]. In [26], we computed an achievability bound for finite-length codes, which showed that after imperfect CSI during rate selection is taken into account, the additional performance impact due to imperfect CSI at the receiver is small.

For the original rate adaptation problem under perfect CSI, it is clear that $\mathcal{M}_{\mathcal{S}_k}(1-\theta_k)$ is convex in the chosen rates. We conjecture that the rate adaptation problem remains convex under imperfect CSI, which is also motivated by the findings for the single-user scenario in [26]. In this paper, we will solve the problem for different values of δ_{M} in order to show the full range of possible trade-offs between user 1 and user 2. We obtain the same solutions if we consider the Lagrangian dual problem

$$\arg\min_{\Phi} \quad \mathcal{M}_{\mathcal{S}_2}(1-\theta_2) + \tilde{\lambda}\mathcal{M}_{\mathcal{S}_1}(1-\theta_1) \tag{33}$$

and iterate over $\tilde{\lambda}$ instead of $\tilde{c}$. When quantizing the joint distribution of Γ_1 and Γ_2 to N points, each point i having probability p_i, the problem becomes

$$\begin{aligned}\arg\min_{r_{1,i},r_{2,i}} \quad \sum_{i=1}^{N} p_i \cdot \Big(&\varepsilon_{2,i} + (1-\varepsilon_{2,i})e^{-\theta_2 n_{\mathrm{d}} r_{2,i}} + \\ &+ \tilde{\lambda}\left(\varepsilon_{1,i} + (1-\varepsilon_{1,i})e^{-\theta_1 n_{\mathrm{d}} r_{1,i}}\right)\Big)\end{aligned} \tag{34}$$

where $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ depend on both $r_{1,i}$ and $r_{2,i}$. The optimal $\mathbf{\Phi}^*$ can be determined by finding the optimal rate pairs $(r_{1,i}, r_{2,i})$ individually for each quantized point i of the SNR distribution. However, to solve the problem, the error probabilities $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ must be computed for each quantized point i and each considered rate pair. Assume that the SNR distribution is quantized on a grid with $N = N_1 \times N_2$ points, and that for each of those points, we consider $M_1 \times M_2$ different rate pairs. If one would need to perform a numerical, 2-dimensional integration to determine the error probabilities for each of the $N_1 N_2 M_1 M_2$ rate/SNR combinations, then the rate adaptation problem would quickly become computationally infeasible, even with a coarse quantization of the distribution and the rate points. Thus, in the following sections, we will present closed-form approximations for $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$, which can be efficiently computed.

C. Outage Probability under SIC Decoding

We first focus on the case where SIC decoding is used. In the ideal model, the scheduler would select a rate pair (r_1, r_2) that corresponds to one of the two corner points of the capacity region. However, in case of imperfect CSI, the transmitter only knows the estimated SNR $\hat{\gamma}_1$ and $\hat{\gamma}_2$. In the event that the selected rate pair (r_1, r_2) lies outside the actual capacity region (defined by Γ_1 and Γ_2, whose exact values are unknown), a decoding error occurs. In order to ensure that this event occurs only with small probability, the scheduler should not select the rates exactly at the corner points of the estimated capacity region but should select smaller rates. Assume for now that the rates are close to the corner point where user 1 has priority in terms of rate, i.e., the base station will try to decode user 2 first such that user 1 experiences no interference.[5] The receiver can decode signal 2 directly if $r_2 \leq \log_2(1 + \Gamma_2/(\Gamma_1 + 1))$.

[5]Due to imperfect CSI, there is now a chance that the channel conditions are reversed, such that user 1 could be decoded first. We still obtain an approximate upper bound on the error probability by ignoring these highly unlikely events.

Afterwards, signal 1 can be decoded if $r_1 \leq \log_2(1+\Gamma_1)$, but only if signal 2 has been decoded without errors (otherwise, the interference from signal 2 cannot be removed, and the error propagates). We define $\beta_k = 2^{r_k} - 1$. The probability that user 1 cannot be decoded is then

$$\varepsilon_1 = \mathbb{P}\left\{\{\Gamma_1 < \beta_1\} \quad \cup \quad \left\{\frac{\Gamma_2}{\Gamma_1+1} < \beta_2\right\}\right\}. \tag{35}$$

Result 3. *When the scheduler has selected a rate pair (r_1, r_2) while assuming that user 2 is decoded first, the resulting decoding error probabilities are approximated as*

$$\varepsilon_1 \approx Q\left(\frac{\gamma_{1,\mathrm{turn}} - \hat{\gamma}_1}{\sigma_1}\right) + Q\left(\frac{\hat{\gamma}_1 - \beta_1}{\sigma_1}\right) + \\ + \frac{\sigma_\mathrm{n}}{2\sigma_1} e^{-\kappa}\left(Q\left(\frac{\beta_1 - \mu_\mathrm{n}}{\sigma_\mathrm{n}}\right) - Q\left(\frac{\gamma_{1,\mathrm{turn}} - \mu_\mathrm{n}}{\sigma_\mathrm{n}}\right)\right) \tag{36}$$

$$\varepsilon_2 \approx Q\left(\frac{\gamma_{1,\mathrm{turn}} - \hat{\gamma}_1}{\sigma_1}\right) + \\ + \frac{\sigma_\mathrm{n}}{2\sigma_1} e^{-\kappa}\left(Q\left(\frac{-\mu_\mathrm{n}}{\sigma_\mathrm{n}}\right) - Q\left(\frac{\gamma_{1,\mathrm{turn}} - \mu_\mathrm{n}}{\sigma_\mathrm{n}}\right)\right) \tag{37}$$

with $\gamma_{1,\mathrm{turn}} = \hat{\gamma}_2/\beta_2 - 1$ and

$$\sigma_\mathrm{n}^2 = \sigma_1^2\left(1 + \frac{\sigma_1^2\beta_2^2}{\sigma_2^2}\right)^{-1} \tag{38}$$

$$\mu_\mathrm{n} = \left(\hat{\gamma}_1 + \frac{\sigma_1^2}{\sigma_2^2}\beta_2(\hat{\gamma}_2 - \beta_2)\right)\left(1 + \frac{\sigma_1^2\beta_2^2}{\sigma_2^2}\right)^{-1} \tag{39}$$

$$\kappa = -\frac{1}{2\cdot\sigma_n^2}\mu_n^2 + \frac{1}{2\cdot\sigma_1^2}\left(\hat{\gamma}_1^2 + \frac{\sigma_1^2}{\sigma_2^2}\left(\hat{\gamma}_2 - \beta_2\right)^2\right) \tag{40}$$

In the other case, when user 1 is decoded first, the resulting error probabilities can be obtained from the same expressions, after switching 1 and 2 in all expressions.

Proof. The probability ε_1 that user 1 cannot be decoded is given as $\varepsilon_1 = \varepsilon_{1,\mathrm{SIC}} + \varepsilon_{2\setminus 1}$, where $\varepsilon_{1,\mathrm{SIC}}$ denotes the probability that signal 1 is in outage after applying SIC:

$$\varepsilon_{1,\mathrm{SIC}} = \mathbb{P}\left\{\Gamma_1 < \beta_1\right\} \approx Q\left(\frac{\hat{\gamma}_1 - \beta_1}{\sigma_1}\right), \tag{41}$$

and $\varepsilon_{2\setminus 1}$ is the probability that signal 1 is not in outage ($\Gamma_1 > \beta_1$), but signal 2 is in outage:

$$\varepsilon_{2\setminus 1} = \int_{\gamma_1=\beta_1}^{\infty} \mathbb{P}\left\{\frac{\Gamma_2}{\gamma_1+1} < \beta_2\right\} f_{\Gamma_1}(\gamma_1)d\gamma_1 \tag{42}$$

$$\approx \int_{\gamma_1=\beta_1}^{\infty} Q\left(\frac{\hat{\gamma}_2 - \beta_2(\gamma_1+1)}{\sigma_2}\right) f_{\Gamma_1}(\gamma_1)d\gamma_1 \tag{43}$$

In order to obtain a closed-form approximation to this integral, we note that the argument of the Q-function is positive for $\gamma_1 < \gamma_{1,\mathrm{turn}}$. We then use the Chernoff bound $Q(x) \leq \frac{1}{2}e^{-\frac{x^2}{2}}$ for $x \geq 0$ [35], along with $Q(x) \leq 1$ for $x < 0$, to obtain $\varepsilon_{2\setminus 1} \approx \varepsilon_{2\setminus 1,\mathrm{a}} + \varepsilon_{2\setminus 1,\mathrm{b}}$ with

$$\varepsilon_{2\setminus 1,\mathrm{a}} = \int_{\gamma_1=\beta_1}^{\gamma_{1,\mathrm{turn}}} \frac{1}{2} e^{-\frac{(\hat{\gamma}_2 - \beta_2(\gamma_1+1))^2}{2\cdot\sigma_2^2}} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(\gamma_1 - \hat{\gamma}_1)^2}{2\cdot\sigma_1^2}} d\gamma_1 \tag{44}$$

$$= \frac{1}{2\sqrt{2\pi\sigma_1^2}} \int_{\gamma_1=\beta_1}^{\gamma_{1,\mathrm{turn}}} e^{-\frac{(\gamma_1 - \mu_n)^2}{2\cdot\sigma_n^2} - \kappa} d\gamma_1 \tag{45}$$

$$= \frac{\sigma_n}{2\sigma_1} e^{-\kappa}\left(Q\left(\frac{\beta_1 - \mu_n}{\sigma_n}\right) - Q\left(\frac{\gamma_{1,\mathrm{turn}} - \mu_n}{\sigma_n}\right)\right), \tag{46}$$

where (45) follows after tedious algebra, applying (38), (39), and (40). We also find

$$\varepsilon_{2\setminus 1,\mathrm{b}} = \int_{\gamma_{1,\mathrm{turn}}}^{\infty} f_{\Gamma_1}(\gamma_1)d\gamma_1 \approx Q\left(\frac{\gamma_{1,\mathrm{turn}} - \hat{\gamma}_1}{\sigma_1}\right). \tag{47}$$

The error probability ε_2 for user 2 is computed in the same way as $\varepsilon_{2\setminus 1}$, but the integral in (42) must start from zero. □

D. Outage Probability under Joint Decoding

In the previous section, we considered a SIC receiver. When using joint decoding, the receiver can decode both codewords when the rates are within the capacity region (4)–(6). If any of the conditions are violated, a decoding error occurs. We assume that neither of the codewords can be decoded in that case. The selected rates violate the capacity constraints with probabilities

$$\varepsilon_\mathrm{I} = \mathbb{P}\left\{r_1 > \log_2(1+\Gamma_1)\right\} \tag{48}$$

$$\varepsilon_\mathrm{II} = \mathbb{P}\left\{r_2 > \log_2(1+\Gamma_2)\right\} \tag{49}$$

$$\varepsilon_\mathrm{III} = \mathbb{P}\left\{r_1 + r_2 > \log_2(1+\Gamma_1+\Gamma_2)\right\} \tag{50}$$

We follow [26] and apply the Gaussian approximation (30) to each term. We then obtain a closed-form bound on ε by applying the union bound $\varepsilon \leq \varepsilon_\mathrm{I} + \varepsilon_\mathrm{II} + \varepsilon_\mathrm{III}$:

$$\varepsilon \leq Q\left(\frac{\hat{\gamma}_1 - \beta_1}{\sigma_1}\right) + Q\left(\frac{\hat{\gamma}_2 - \beta_2}{\sigma_2}\right) + Q\left(\frac{\hat{\gamma}_1 + \hat{\gamma}_2 - 2^{r_1+r_2} + 1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right). \tag{51}$$

V. Imperfect CSI and Finite-Length Coding

In the previous section, we used a simple outage model that assumes decoding errors occur if and only if the selected rates are above the channel capacity. However, when the blocklength n_d of the channel code is finite, that model no longer holds. In fact, when the blocklength is finite, errors occur with probability $\varepsilon > 0$ even when the channel is not in outage. In this section we analyze the joint impact of imperfect CSI and finite-length coding on the decoding error probabilities based on methods we developed in [26].

A well-known result for finite-length codes in AWGN channels with SNR γ is given by Polyanskiy et al. [3], who showed in order to achieve an error probability ε, the transmitter should select a rate

$$r_\mathrm{AWGN}(\gamma, n_\mathrm{d}, \varepsilon) \approx \log_2(1+\gamma) - \sqrt{\frac{\mathcal{V}_\mathrm{AWGN}(\gamma)}{n_\mathrm{d}}} Q^{-1}(\varepsilon), \tag{52}$$

where

$$\mathcal{V}_\mathrm{AWGN} = \log_2^2(e)\left(1 - \frac{1}{(1+\gamma)^2}\right) \tag{53}$$

is the channel dispersion. However, the codewords that achieve (52) are not Gaussian distributed, which means that the result cannot be directly applied to the multiuser scenario, where non-Gaussian codewords would create non-Gaussian

interference, which is hard to analyze. When the codewords must be i.i.d. Gaussian, the achievable transmission rate is approximated as [21]

$$r_{\mathrm{iid}}(\gamma, n_{\mathrm{d}}, \varepsilon) \approx \log_2(1+\gamma) - \sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\gamma)}{n_{\mathrm{d}}}} Q^{-1}(\varepsilon)\,, \tag{54}$$

which has the same form as (52), but with a different dispersion term

$$\mathcal{V}_{\mathrm{iid}} = \log_2^2(e) \frac{2\gamma}{1+\gamma}\,. \tag{55}$$

We now analyze the finite blocklength effects separately for the case of a decoder using successive interference cancellation, and for a joint decoder. The analysis in this section closely follows our previous works [26], [27]. To simplify discussions, we assume that (54) holds with equality.

A. Error Probability for SIC Decoding

Before analyzing the joint impact of imperfect CSI and finite blocklength coding, we first assume that the SNR values γ_1 and γ_2 are perfectly known. We again consider the case where the decoder uses successive interference cancellation and decodes user 2 first. In this case, the SINR of the signal is given as $\mathsf{SINR}_2 = \frac{\gamma_2}{\gamma_1+1}$. Due to the assumption of i.i.d. Gaussian codewords for user 1, the signal is equivalent to that of an AWGN channel with SNR SINR_2. Thus, the achievable rate for user 2 can be determined by (54). We can solve (54) for ε to obtain the decoding error probability given the selected rate r_2:

$$\varepsilon_2 = Q\left(\frac{\log_2(1+\mathsf{SINR}_2) - r_2}{\sqrt{\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR}_2)/n_{\mathrm{d}}}} \right). \tag{56}$$

According to the authors in [36], such an expression for the error probability can be interpreted in the following way: the communication channel corresponds to a "bit pipe" of random size. Errors occur when the rate r_2 of the channel code is above the size of the bit pipe, which is Gaussian distributed with mean $\log_2(1+\mathsf{SINR}_2)$ and variance $\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR}_2)/n_{\mathrm{d}}$. We follow our previous work [26] and denote this random size as the *blocklength-equivalent capacity*

$$C_{\mathrm{FBL},2} = \log_2(1+\mathsf{SINR}_2) + \sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR}_2)}{n_{\mathrm{d}}}} U_2 \tag{57}$$

with $U_2 \sim \mathcal{N}(0,1)$ and observe for a given SINR_2 that ε_2 is (by definition) the outage probability of a channel with random capacity $C_{\mathrm{FBL},2}$: $\varepsilon_2 = \mathbb{P}\{C_{\mathrm{FBL},2} < r_2\}$. Similarly, we define $U_1 \sim \mathcal{N}(0,1)$ and find that the error probability $\varepsilon_{1,\mathrm{SIC}}$ of user 1, after SIC was applied, is given as $\varepsilon_{1,\mathrm{SIC}} = \mathbb{P}\{C_{\mathrm{FBL},1} < r_1\}$ with

$$C_{\mathrm{FBL},1} = \log_2(1+\gamma_1) + \sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\gamma_1)}{n_{\mathrm{d}}}} U_1\,. \tag{58}$$

All of the above statements remain true when the instantaneous SNR γ_k is not perfectly known at the transmitters, i.e., we can replace γ_k with Γ_k, where all random variables U_k and Γ_k are mutually independent. At the receiver, we still assume perfect CSI, as motivated in Sec. IV. For user 1, we can directly follow our previous work [26] to obtain a bound on $\varepsilon_{1,\mathrm{SIC}}$. Using a first-order Taylor approximation, we obtain the bound

$$\varepsilon_{1,\mathrm{SIC}} \le \mathbb{P}\left\{\log_2\left(1+\Gamma_1+\sigma_{\mathrm{FBL},1}(\Gamma_1)U_1\right) < r_1\right\} \tag{59}$$

with

$$\sigma_{\mathrm{FBL},1}(\Gamma_1) = \frac{1+\Gamma_1}{\log_2(e)} \sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\Gamma_1)}{n_{\mathrm{d}}}}\,. \tag{60}$$

Using the Gaussian approximation (30) for Γ_1, and replacing $\sigma_{\mathrm{FBL},1}(\Gamma_1)$ with its estimated value $\sigma_{\mathrm{FBL},1}(\hat{\gamma}_1)$, we obtain [26]

$$\varepsilon_{1,\mathrm{SIC}} \approx Q\left(\frac{\hat{\gamma}_1 - \beta_1}{\sigma_{\mathrm{IC,F},1}(\hat{\gamma}_1)} \right), \tag{61}$$

with the variance of the sum of the independent Gaussian variables Γ_1 and U_1 given as

$$\sigma_{\mathrm{IC,F},1}(\hat{\gamma}_1) = \sqrt{\sigma_1^2 + \sigma_{\mathrm{FBL},1}^2(\hat{\gamma}_1)}\,, \tag{62}$$

where σ_1^2 is given by (31). User 2 is decoded directly, with error probability

$$\varepsilon_2 = \mathbb{P}\{C_{\mathrm{FBL},2} < r_2\} \tag{63}$$

$$= \mathbb{P}\left\{ \log_2\left(1+\frac{\Gamma_2}{\Gamma_1+1}\right) + \sqrt{\frac{\mathcal{V}_{\mathrm{iid}}\left(\frac{\Gamma_2}{\Gamma_1+1}\right)}{n_{\mathrm{d}}}} U_2 < r_2 \right\} \tag{64}$$

$$\le \mathbb{P}\left\{ \log_2\left(1+\frac{\Gamma_2}{\Gamma_1+1} + \sigma_{\mathrm{FBL},2}\left(\frac{\Gamma_2}{\Gamma_1+1}\right) U_2\right) < r_2 \right\} \tag{65}$$

$$= \mathbb{P}\left\{ \Gamma_2 + (1+\Gamma_1)\sigma_{\mathrm{FBL},2}\left(\frac{\Gamma_2}{\Gamma_1+1}\right) U_2 < \beta_2(1+\Gamma_1) \right\} \tag{66}$$

$$\approx \mathbb{P}\left\{ \Gamma_2 + (1+\hat{\gamma}_1)\sigma_{\mathrm{FBL},2}\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1+1}\right) U_2 < \beta_2(1+\Gamma_1) \right\}, \tag{67}$$

where (65) follows again from a Taylor approximation, and in (67) we replaced the random variables in and before $\sigma_{\mathrm{FBL},2}$ with their estimated values, similar to [26]. Thus:

$$\varepsilon_2 \approx \int_{\gamma_1=0}^{\infty} \mathbb{P}\left\{ \frac{G_{\mathrm{IC,F,int},2}}{\gamma_1+1} < \beta_2 \right\} f_{\Gamma_1}(\gamma_1) d\gamma_1 \tag{68}$$

where $G_{\mathrm{IC,F,int},2} = \Gamma_2 + (1+\hat{\gamma}_1)\sigma_{\mathrm{FBL},2}\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1+1}\right) U_2$ describes the uncertainty of signal 2 due to imperfect CSI and finite blocklength. $G_{\mathrm{IC,F,int},2}$ is Gaussian with variance

$$\sigma_{\mathrm{IC,F,int},2}^2(\hat{\gamma}_1, \hat{\gamma}_2) = \sigma_2 + (1+\hat{\gamma}_1)^2 \sigma_{\mathrm{FBL},2}^2\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1+1}\right). \tag{69}$$

Following the same steps as in Sec. IV-C, we obtain

$$\varepsilon_2 \approx Q\left(\frac{\gamma_{1,\mathrm{turn}} - \hat{\gamma}_1}{\sigma_1}\right) + \\ + \frac{\sigma_{\mathrm{n}}}{2\sigma_1} e^{-\kappa} \cdot \left(Q\left(\frac{-\mu_{\mathrm{n}}}{\sigma_{\mathrm{n}}}\right) - Q\left(\frac{\gamma_{1,\mathrm{turn}} - \mu_{\mathrm{n}}}{\sigma_{\mathrm{n}}}\right)\right) \tag{70}$$

with σ_{n}, μ_{n}, and κ still given by (38), (39), and (40), but with σ_2 replaced by $\sigma_{\mathrm{IC,F,int},2}$.

Finally, user 1 cannot be decoded when it cannot be decoded after interference cancellation or when user 2 cannot be decoded. Contrary to the analysis with infinite blocklength in Sec. IV, those events may not be independent, so we have to apply the union bound:

$$\varepsilon_1 \leq \varepsilon_{1,\mathrm{SIC}} + \varepsilon_2 \,. \tag{71}$$

B. Error Probability for Joint Decoding

The achievable rate region for a multiple access channel with finite blocklength coding has been studied by Molavian-Jazi [22]. We first assume that the SNR values γ_1, γ_2 are perfectly known, such that these values correspond to the power constraints on the codewords. It was shown in [22, Thm. 7] that given a maximum error probability ε, a second-order approximation for the achievable rate region can be found by splitting the error probability into three arbitrarily large parts with $\varepsilon = \varepsilon_{\mathrm{I}} + \varepsilon_{\mathrm{II}} + \varepsilon_{\mathrm{III}}$. The rates r_1, r_2 are achievable with error probability ε if

$$r_1 \leq \log_2(1+\gamma_1) - \sqrt{\frac{\mathcal{V}_{\mathrm{AWGN}}(\gamma_1)}{n_{\mathrm{d}}}} Q^{-1}(\varepsilon_{\mathrm{I}}) + \mathcal{O}\left(\frac{1}{n_{\mathrm{d}}}\right) \tag{72}$$

$$r_2 \leq \log_2(1+\gamma_2) - \sqrt{\frac{\mathcal{V}_{\mathrm{AWGN}}(\gamma_2)}{n_{\mathrm{d}}}} Q^{-1}(\varepsilon_{\mathrm{II}}) + \mathcal{O}\left(\frac{1}{n_{\mathrm{d}}}\right) \tag{73}$$

$$r_1 + r_2 \leq \log_2(1+\gamma_1+\gamma_2) - \sqrt{\frac{\mathcal{V}_{\mathrm{III}}(\gamma_1,\gamma_2)}{n_{\mathrm{d}}}} Q^{-1}(\varepsilon_{\mathrm{III}}) + \mathcal{O}\left(\frac{1}{n_{\mathrm{d}}}\right) \tag{74}$$

with $\mathcal{V}_{\mathrm{AWGN}}$ given in (53) and

$$\mathcal{V}_{\mathrm{III}}(\gamma_1,\gamma_2) = \mathcal{V}_{\mathrm{AWGN}}(\gamma_1+\gamma_2) + 2\log_2^2(e)\frac{\gamma_1\gamma_2}{(1+\gamma_1+\gamma_2)^2} \,. \tag{75}$$

Like in the previous section, we assume that the second-order approximations are exact, i.e., we ignore the terms $\mathcal{O}\left(1/n_{\mathrm{d}}\right)$. Then, we define random blocklength-equivalent capacities

$$C_{\mathrm{FBL,I}} = \log_2(1+\gamma_1) + \sqrt{\mathcal{V}_{\mathrm{AWGN}}(\gamma_1)/n_{\mathrm{d}}} \cdot U_{\mathrm{I}} \tag{76}$$

$$C_{\mathrm{FBL,II}} = \log_2(1+\gamma_2) + \sqrt{\mathcal{V}_{\mathrm{AWGN}}(\gamma_2)/n_{\mathrm{d}}} \cdot U_{\mathrm{II}} \tag{77}$$

$$C_{\mathrm{FBL,III}} = \log_2(1+\gamma_1+\gamma_2) + \sqrt{\mathcal{V}_{\mathrm{III}}(\gamma_1,\gamma_2)/n_{\mathrm{d}}} \cdot U_{\mathrm{III}} \tag{78}$$

with $U_{\mathrm{I}}, U_{\mathrm{II}}, U_{\mathrm{III}} \sim \mathcal{N}(0,1)$. If we redefine ε_{I}, $\varepsilon_{\mathrm{II}}$, and $\varepsilon_{\mathrm{III}}$ as the probabilities that the rates exceed $C_{\mathrm{FBL,I}}$, $C_{\mathrm{FBL,II}}$, and $C_{\mathrm{FBL,III}}$, respectively, then the overall probability that the rates (r_1, r_2) are outside the random blocklength-equivalent capacity region is bounded by $\varepsilon = \varepsilon_{\mathrm{I}} + \varepsilon_{\mathrm{II}} + \varepsilon_{\mathrm{III}}$. The codewords that achieve (72) to (74) are chosen independently of each other, and are uniformly distributed on the "power shells" [22]. Therefore, the choice of codewords depends only on the selected rates, but not on the fading state, and we can thus extend the above results directly to fading channels. For ε_{I} and $\varepsilon_{\mathrm{II}}$, we can then directly use the approximation (61), but using a dispersion term $\mathcal{V}_{\mathrm{AWGN}}$. Using similar steps, we obtain $\varepsilon_{\mathrm{III}} \approx Q\left(\frac{\hat{\gamma}_1+\hat{\gamma}_2-\beta_{12}}{\sigma_{\mathrm{III}}}\right)$ with

$$\sigma_{\mathrm{III}} = \sqrt{\sigma_1^2 + \sigma_2^2 + \frac{(1+\hat{\gamma}_1+\hat{\gamma}_2)^2}{\log_2^2(e)} \frac{\mathcal{V}_{\mathrm{III}}(\hat{\gamma}_1,\hat{\gamma}_2)}{n_{\mathrm{d}}}} \,. \tag{79}$$

VI. Numerical Results

In this section, we evaluate the queueing delay of NOMA and OMA schemes. First, we present in Sec. VI-A the general methodology used in the evaluation. The queueing performance of NOMA depends on the rate adaptation, and we thus show in Sec. VI-B how different rate adaptation schemes chosen by the base station result in different delay violation probabilities for the two users, assuming an ideal system model. In Sec. VI-C, we consider the effects of imperfect CSI and validate the outage probability approximations from Sec. IV-C. In Sec. VI-D, we study the performance impact due to imperfect CSI, and in Sec. VI-E, we also take finite-length coding into account.

A. Metrics and Methodology

For comparing different schemes, we generally use the upper bound (11) on the delay violation probability $p_{\mathrm{v},k}(w)$. The delay performance depends on the number of bits α_k that arrive in each user's queue per time slot. To compare the delay performance of different schemes over a wide range of parameters, we will often assume fixed target delay parameters (e.g., $p_{\mathrm{v},k}(5) < 10^{-8}$) and show the maximum arrival rates α_1, α_2 such that these constraints can still be met.

1) Optimization Method: We quantize the distribution of the estimated SNR values $\hat{\gamma}_1, \hat{\gamma}_2$ until finer quantization yields no more significant improvement, usually around 300 points each. For the ideal model with perfect CSI ($\gamma_k = \hat{\gamma}_k$), we can determine the optimal rates according to the results in Sec. III. For imperfect CSI and finite blocklength, we consider several rate pairs $(r_{1,i}, r_{2,i})$ for each combination of quantized $(\hat{\gamma}_{1,i}, \hat{\gamma}_{2,i})$, each resulting in different approximate error probabilities $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$. Then, for fixed values θ_1, θ_2, and $\tilde{\lambda}$, we can determine the optimal rates according to (34). Instead of an exhaustive search over all combinations of θ_1, θ_2, and $\tilde{\lambda}$, we propose a suboptimal approach where we start with a coarsely quantized grid for θ_k and $\tilde{\lambda}$. We then iterate between optimizing the rates and choosing θ_k and $\tilde{\lambda}$ such that the kernels $\mathcal{K}_k\left(\theta_K, w\right)$ are minimized. This method seems to offer robust and reasonably fast convergence.

2) Comparison to OMA and Power Allocation: We assume throughout this section that the signals are subject to a constraint on the sum instantaneous transmit power. We denote the average SNR when one device uses the entire transmit power as $\bar{\gamma}_1^{\mathrm{o}}$ and $\bar{\gamma}_2^{\mathrm{o}}$, respectively. These are also the average SNR values of the signals when using OMA. For NOMA, both devices are active, so the transmit powers must be scaled by factors ρ_1 and ρ_2, with $\rho_1 + \rho_2 = 1$, so that the average SNR values are given as $\bar{\gamma}_1 = \rho_1\bar{\gamma}_2^{\mathrm{o}}$ and $\bar{\gamma}_2 = \rho_2\bar{\gamma}_2^{\mathrm{o}}$. We always consider scenarios where $\bar{\gamma}_1^{\mathrm{o}} > \bar{\gamma}_2^{\mathrm{o}}$, and assign $\rho_1 = 0.2$ and $\rho_2 = 0.8$. During the channel training phase, channel access is orthogonal, so the SNR during training is $\bar{\gamma}_1^{\mathrm{o}}$ and $\bar{\gamma}_2^{\mathrm{o}}$, respectively.

B. Results for Ideal Model

First, we consider SIC decoding and assume an ideal system model where the CSI at the transmitter is perfect and finite blocklength effects are ignored. In Fig. 3, we study the effect of different decoding orders, i.e., different rate adaptation functions $(r_1, r_2) = \mathbf{\Phi}(\gamma_1, \gamma_2)$ on the delay performance of both users under SIC. In the upper-most plot of Fig. 3a, we consider the traditional decoding order, suggested in [8]–[12], where user 1 is always decoded first when $\gamma_1 > \gamma_2$. The plot shows the resulting rates $r_1 = \Phi_1(\gamma_1, \gamma_2)$ for user 1. In the middle and bottom plots of Fig. 3a, we show the optimal r_1 when the delay violation probability $p_{\mathrm{v}}(w)$ of the second user is minimized subject to constraints $p_{\mathrm{v}}(10) < 10^{-8}$ (middle plot) and $p_{\mathrm{v}}(5) < 10^{-8}$ (bottom plot) for the first user. In the area below the dotted red line, user 1 is decoded first and its signal suffers from interference (for reference, the dashed red line still marks $\gamma_1 = \gamma_2$). Fig. 3b shows the corresponding delay violation probabilities $p_{\mathrm{v}}(w)$ for both users. When optimizing the decoding orders with a constraint $p_{\mathrm{v}}(10) < 10^{-8}$ for user 1 (middle plot), the resulting rates and delay violation probabilities are similar to those for the stronger-user-first decoding order (upper plot). However, user 1 will often experience fairly long delays, which is problematic in cases where the data from user 1 is time-critical and should be delivered e.g. with at most $w = 5$ time slot delay. The bottom plots of Fig. 3 show the decoding orders and delay violation probabilities under a constraint $p_{\mathrm{v}}(5) < 10^{-8}$ for user 1. We note that the decoding order is now quite different from the upper and middle plots: user 1 is decoded first only when γ_1 is about 3 to 6 dB above γ_2. The delay violation probabilities of user 1 now meet the constraint $p_{\mathrm{v}}(5) < 10^{-8}$. Although user 2 now experiences higher delay violation probabilities, this may still be tolerable for the application served by that user. In conclusion, we find that the traditional stronger-user-first decoding order performs fairly well in some specific cases, but the resulting delay performance may not always meet the delay constraints for both users. In contrast, we derived optimized decoding orders that provide an optimal trade-off between the two users, depending on the performance constraints of the individual applications.

We also determined the actual delay violation probabilities $p_{\mathrm{v}}(w)$ empirically through extensive Monte-Carlo simulations of the queueing system over 10^{10} time slots. For the simulations, one must generate random values of the SNR $(\gamma_1(t), \gamma_2(t))$ for each time slot t and determine the rates of both users as $(r_1(t), r_2(t)) = \mathbf{\Phi}(\gamma_1(t), \gamma_2(t))$ in each time slot. For the queue at each of the users, one can then determine the departure process $D_k(t)$ from the arrivals $A_k(t) = \alpha_k$ and the service $S_k(t) = nr_k(t)$, determine the virtual delay $W(t)$ in (7), and thus determine $p_{\mathrm{v}}(w)$. As expected, the simulated $p_{\mathrm{v}}(w)$ is below the analytical upper bounds from SNC. Because the upper bounds from SNC are derived based on union bounds and moment bounds, there is a gap of one or two orders of magnitude between the simulations and the upper bounds, which was observed also in previous works on SNC [6], [25], [26], [37]. Regardless of the gap, the slopes of the simulation curves match the slopes of the analytical

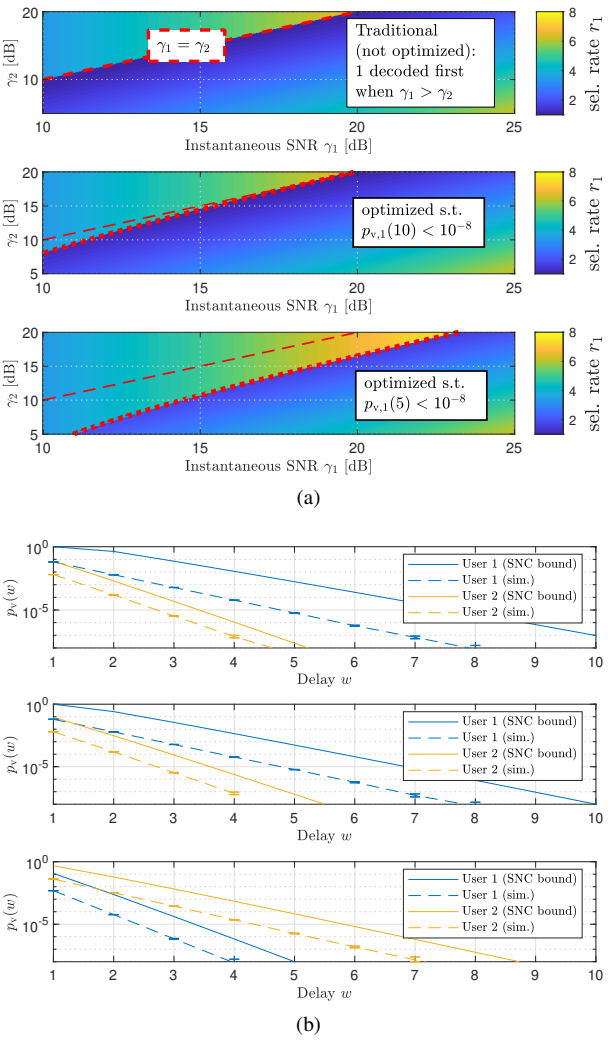


Fig. 3. Average SNR: $\bar{\gamma}_1^{\mathrm{o}} = 30$ dB, $\bar{\gamma}_2^{\mathrm{o}} = 15$ dB. $n = 250$, $n_{\mathrm{d}} = 200$, $\alpha_1 = 560$, $\alpha_2 = 320$ bits per time slot. Top: stronger-user-first. Middle: optimized s.t. $p_{\mathrm{v},1}(10) < 10^{-8}$. Bottom: optimized s.t. $p_{\mathrm{v},1}(5) < 10^{-8}$. a) Optimal rate $r_1 = \Phi_1(\gamma_1, \gamma_2)$. b) Delay violation probability $p_{\mathrm{v}}(w)$ (From SNC bounds and from simulations, 95% confidence intervals shown).

curves, and the SNC bounds accurately predict how different rate adaptations will affect the delay performance of both users without the need for extensive simulations. Thus, the proposed analysis using SNC is sufficiently accurate to determine the optimal trade-offs between the two users.

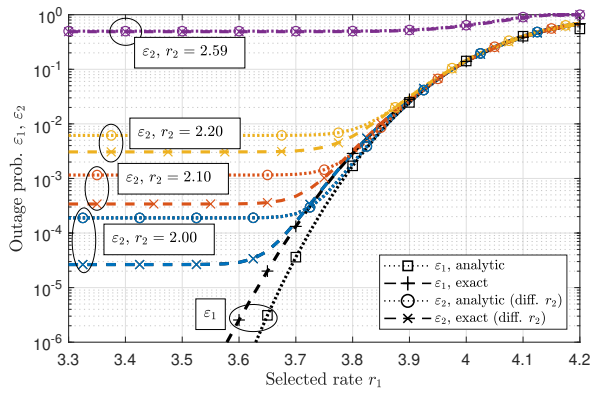
C. Validating the Approximations for Imperfect CSI

When considering that the rate adaptation mechanism only has access to an imperfect estimate of the channel state, outages may occur. For such a scenario, we perform the delay analysis and rate optimization for NOMA with SIC decoding based on the analytical approximations for the outage probabilities ε_1 and ε_2 that were derived in Sec. IV-C. First, we verify that these analytical approximations are sufficiently accurate for the rate adaptation $(r_1, r_2) = \mathbf{\Phi}(\hat{\gamma}_1, \hat{\gamma}_2)$. In Fig. 4a, we plot both ε_1 and ε_2 when the base station estimates the instantaneous SNRs as $\hat{\gamma}_1 = 20$ dB and $\hat{\gamma}_2 = 7$dB and tries to adapt the coding rates (r_1, r_2) according to these estimates. User 1 is decoded first. When the message from user 1 cannot be decoded (which depends on r_1), the signal cannot be removed,

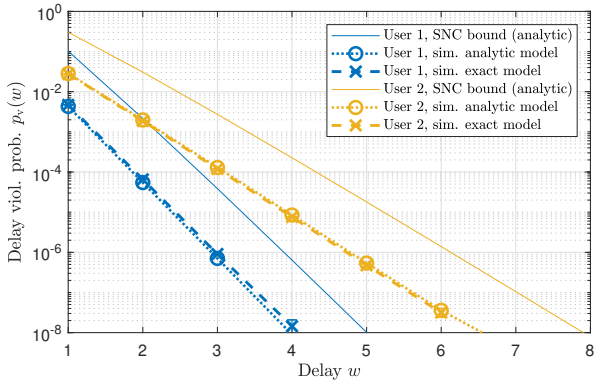
and the SIC decoding fails. Otherwise, the base station will try to decode signal 2, which can again fail, depending on r_2. We note that the analytical approximations for ε_1 (solid curves) are close to the actual error probabilities (dashed curves), obtained from Monte Carlo simulations with 10^8 trials. This is important for rate adaptation: if the rate adaptation scheme would ignore the imperfections in the channel estimates, and select a rate of $r_1 = \log_2(1+\hat{\gamma}_1/(1+\hat{\gamma}_2)) \approx 4.14$ based on the estimates, then the system would experience error probabilities ε_1 of around 50%. If the rate selection wants to keep ε_1 below 10^{-3}, one would actually need to choose $r_1 \approx 3.76$ according to the Monte-Carlo simulations. When using the analytical approximation, one would select approximately the same rate ($r_1 \approx 3.78$). Thus, one can perform accurate rate adaptation without the need for extensive simulations (or numerical integrations). For ε_2, we notice a fairly large gap between the analytical approximations of ε_2 and the actual ε_2 at $r_2 = 2.0$. Nevertheless, this gap may not be harmful for the rate adaptation. When considering systems with a deadline of e.g. $w = 5$ time slots, then individual decoding error probabilities in the range 10^{-3} to 10^{-2} are sufficient to transmit the packets within the deadline. The figure shows that for error probabilities above 10^{-3}, the approximations are fairly accurate. Most importantly, the approximations are either very close to the actual ε_1 or they overestimate ε_1. When the base station selects the rates based on an overestimation of the error probability, it will make a conservative choice and select a slightly smaller rate than necessary, leading to a smaller error probability than expected.

In Fig. 4b, we analyze the queueing performance when the rate is adapted to imperfect CSI. The rate adaptation scheme was optimized using SNC and the analytical approximations for ε_1 and ε_2. We first consider the analytical model based on our derived approximations. The solid curves show the SNC bounds on $p_{\mathrm{v}}(w)$, whereas the dotted curves show the actual $p_{\mathrm{v}}(w)$ from Monte-Carlo simulations for the same model. Similar to the case of perfect CSI in Fig. 3b, we observe a gap between the SNC bounds and the simulations. We note here that the simulation results (dotted curves) and the SNC bounds (solid curves) are based on the same analytical approximations for the error probabilities. Therefore, the gap cannot be caused by inaccuracy of the derived approximations, but is due to the conservativeness of the SNC upper bounds, which was observed also at perfect CSI and in many previous works on SNC.

In addition, Fig. 4b shows that the derived approximations in Result 3 are sufficiently accurate in terms of the resulting queueing performance. Specifically, we compare the Monte-Carlo simulations for the approximate model with simulations for the actual (exact) system model. In both cases, we first generate random values of the estimated channel coefficients $\hat{h}_k$ and select the rates according to the optimized rate adaptation scheme. The dotted curves show $p_{\mathrm{v}}(w)$ for a hypothetical system where decoding error events are Bernoulli random variables, whose probabilities are given by the analytical approximations. The dashed curves show the results from simulating decoding errors according to the actual/exact system model, where we generate random instances of $\tilde{H}_k$



(a)



(b)

Fig. 4. Average SNR: $\bar{\gamma}_1^{\mathrm{o}} = 30$ dB, $\bar{\gamma}_2^{\mathrm{o}} = 15$ dB. $n = 250$ $n_{\mathrm{t},1} = n_{\mathrm{t},2} = 25$. a) Estimated SNR: $\hat{\gamma}_1 = 20$ dB, $\hat{\gamma}_2 = 7$ dB. Decoding error probabilities ε_1, ε_2 according to analytical results and simulations, vs. r_1 for different r_2. b) Delay viol. prob. $p_{\mathrm{v}}(w)$ for a system where the analytical ε_1, ε_2 were used for optimal rate adaptation. Arrivals $\alpha_1 = 560$, $\alpha_2 = 160$ bits/slot.

and declare an error event whenever the rates are above the corresponding actual channel capacities (based on the actual values of the SNR $\Gamma_k = \bar{\gamma}_k|h_k + \tilde{H}_k|^2$). We observe that there is no significant difference between the simulations for the the approximate model (dotted curves) and the actual/exact system model (dashed curves). This shows that the actual error probabilities ε_1 and ε_2 and their analytical approximations in Result 3 are sufficiently close for the parameters considered in the queueing analysis. In other words, whether we consider the actual system model or the approximations, the queueing performance remains very similar.

We conducted further experiments which show that the analytical approximations work very well when the average SNR $\bar{\gamma}_k$ for each user's signal is above 10 dB and for training sequence length $n_{\mathrm{t},k}$ is above 25. We also performed simulations to validate the approximations for finite-length coding in Sec. V, where we found that $p_{\mathrm{v}}(w)$ for the actual system model was slightly below the $p_{\mathrm{v}}(w)$ for the hypothetical/approximate model. This means that although the approximations can be slightly inaccurate, they are upper bounds on the error probability, similar to the approximation in [26]. The actual system performance is then even better than predicted.

D. Effects of Imperfect CSI

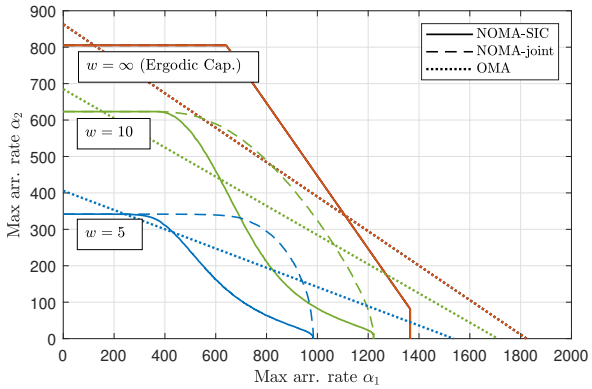
In Fig. 5, we investigate the performance for the ideal model with perfect CSI, and compare the resulting performance also

to a semi-realistic model with imperfect CSI. In both cases, we still assume that errors occur only when the rates are above the Shannon capacity (i.e., infinitely long codewords). First, Fig. 5a shows results for the ideal model with perfect CSI. The uppermost curves show the ergodic capacity per time slot, which corresponds to the maximum supported arrival rates α_1 and α_2 when there is no delay constraint (the delay may be infinite). Then, we investigate the maximum arrival rates α_1 and α_2 such that, with an optimized rate adaptation scheme, the system can still meet delay constraints $p_{\mathrm{v}}(w) < 10^{-8}$ for both users, for different target delays w. We find that imposing tight delay requirements ($w = 5$) degrades the performance, the maximum arrival rate reduces drastically compared to the ergodic case. For a maximum delay $w = 5$, NOMA with SIC decoding now performs significantly worse than NOMA with joint decoding, because the optimal rate points for joint decoding often lie between the corner points of the capacity region. With SIC decoding, the rate adaptation scheme can only select the suboptimal corner points. Interestingly, for NOMA with joint decoding, the performance of the second user remains constant over a wide range of arrival rates α_1 for the first user, i.e., both users can simultaneously achieve a large fraction of the maximum performance. For both system models, NOMA-joint significantly outperforms the orthogonal scheme (OMA), except for the regions where either α_1 or α_2 are very small. We found through further experiments that the performance of NOMA in those regions could be improved by using different power allocations ρ_1,ρ_2, from which we conclude that NOMA-joint always outperforms OMA for the considered scenario with $w = 5$. On the other hand, NOMA with SIC decoding can only provide a small performance improvement over OMA for $w = 5$. For different power allocations, we found that NOMA-SIC may not even achieve same performance as OMA.

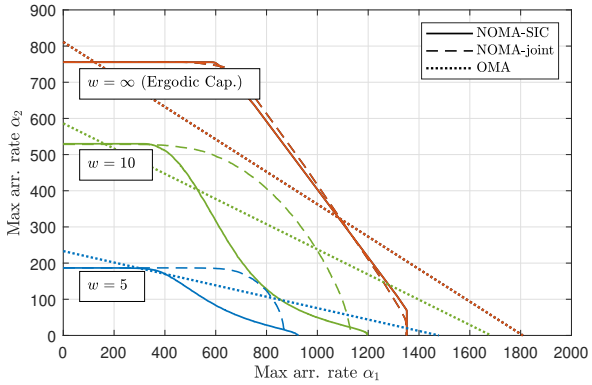
In Fig. 5b, we show results for the same parameters as in Fig. 5a, but considering imperfect CSI. We note first of all that imperfect CSI barely affects the performance under loose delay constraints ($w = \infty$ and $w = 10$), but reduces the maximum achievable α_2 under tighter delay constraints ($w = 5$) by more than 40%. For both models, we observe that NOMA-joint significantly outperforms OMA. However, when considering imperfect CSI, NOMA-SIC can no longer outperform OMA for the considered parameters. We conclude that imperfect CSI creates a slightly larger performance penalty for NOMA-SIC than for OMA.

E. Different Channel Symmetries and Finite-Length Coding

While we have previously considered a setup with $\bar{\gamma}_1^{\mathrm{o}} = 30$ dB, $\bar{\gamma}_2^{\mathrm{o}} = 15$ dB, we investigate in Fig. 6 the cases $\bar{\gamma}_1^{\mathrm{o}} \in \{20, 30, 40\}$ dB and $\bar{\gamma}_2^{\mathrm{o}} = 15$ dB, i.e., we investigate different ratios between the users' average SNR values. Furthermore, we will now consider also finite blocklength effects. In Fig. 6a, we show again results for the semi-realistic model with imperfect CSI, but still assuming that errors occur only when the rates are above the Shannon capacity (i.e., infinite blocklength), whereas Fig. 6b shows results for the realistic model where finite blocklength effects are also considered.



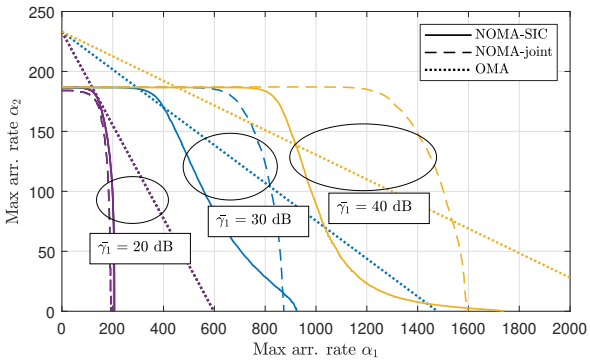
(a)



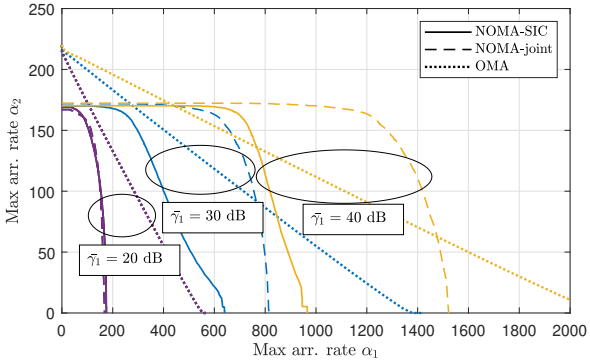
(b)

Fig. 5. Max arrival rates α_1, α_2 in bits per time slot s.t. $p_{\mathrm{v}}(w) < 10^{-8}$ for different w. Average SNR: $\bar{\gamma}_1^{\mathrm{o}} = 30$ dB, $\bar{\gamma}_2^{\mathrm{o}} = 15$ dB. $n = 250$, $n_{\mathrm{t},1} = n_{\mathrm{t},2} = 25$, $n_{\mathrm{d}} = 200$. Shannon capacity model (infinite blocklength). a) Perfect CSI b) Imperfect CSI.

For $\bar{\gamma}_1^{\mathrm{o}} = 30$ dB, the results in Fig. 6a were already shown in Fig. 5b. When comparing them to the new results in Fig. 6b, we find that NOMA-joint still outperforms OMA, but NOMA-SIC performs worse than OMA when finite-length coding is taken into account. For $\bar{\gamma}_1^{\mathrm{o}} = 40$ dB, NOMA-joint outperforms both NOMA-SIC and OMA by a large margin. In this scenario where the difference between the two users' average SNR is large, NOMA-SIC can still outperform OMA, but only by a fairly small margin. For $\bar{\gamma}_1^{\mathrm{o}} = 20$ dB, the chosen power allocation $\rho_1 = 0.2$, $\rho_2 = 0.8$ results in an almost symmetric scenario in terms of average SNR values (NOMA usually performs best in asymmetric scenarios, but we confirmed for this case that NOMA still exceeds OMA in the sum ergodic capacity). We observe that OMA outperforms both NOMA schemes, and that there is a significant margin between OMA and the NOMA schemes once finite blocklength effects are considered. This indicates that the NOMA schemes suffer more from finite blocklength coding than the OMA scheme. It must be noted that the results for finite-length coding from [21] and [22] are approximations, and we are not aware of information-theoretic bounds that can be used to verify their accuracy. Furthermore, the codewords for OMA are below 200 symbols long, so that (52) starts to become inaccurate. However, it is noteworthy that finite blocklength effects seem to create a larger penalty for the NOMA schemes than for OMA, despite the fact that OMA operates with two codewords



(a)



(b)

Fig. 6. Max arrivals α_1, α_2 in bits/slot s.t. $p_{\mathrm{v}}(w) < 10^{-8}$, $w = 5$. Average SNR: $\bar{\gamma}_1^{\mathrm{o}} \in \{20, 30, 40\}$ dB, $\bar{\gamma}_2^{\mathrm{o}} = 15$ dB. $n = 250$, $n_{\mathrm{t},1} = n_{\mathrm{t},2} = 25$, $n_{\mathrm{d}} = 200$. Imperfect CSI. a) Shannon capacity model (infinite blocklength) b) Finite blocklength model

of shorter blocklength. This demonstrates that the blocklength itself is not the only factor that determines the performance impact of finite blocklength channel coding.

VII. Conclusions

In this work, we analyzed the queueing performance of NOMA in the uplink. We found that even under realistic assumptions, NOMA may be suitable for low-latency communications, but only when joint decoding is used and only when there is a large difference between the two users' average SNR values. However, joint decoding may be difficult to implement in practice, especially for low-latency systems. With SIC decoding, NOMA often performs worse than OMA when considering low-latency communications with more realistic system effects.

Aside from the interference-cancellation techniques investigated in this paper, simultaneous uplink from several users can also be enabled through multi-antenna technology, where the interference from different users can be mitigated through receive beamforming. Combining these two approaches would yield an interesting extension to our results.

References

[1] 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 Sig. Proc. Letters*, vol. 21, no. 12, pp. 1501–1505, Dec 2014.

[2] 3GPP, "Study on communication for automation in vertical domains," 3GPP, Tech. Rep. 22.804, 2018. [Online]. Available: http://www.3gpp.org/DynaReport/22804.htm

[3] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[4] 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.

[5] M. Fidler, "A network calculus approach to probabilistic quality of service analysis of fading channels," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2006, pp. 1–6.

[6] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[7] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, "On the optimality of power allocation for NOMA downlinks with individual QoS constraints," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, 2017.

[8] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov 2016.

[9] 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.

[10] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *IEEE Veh. Technol. Conf. (VTC)*, May 2014, pp. 1–5.

[11] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," in *Int. Symp. Wireless Commun. Systems (ISWCS)*, Aug. 2012, pp. 261–265.

[12] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Commun. Letters*, vol. 20, no. 3, pp. 458–461, Mar. 2016.

[13] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, 2015.

[14] Q. Yang, H.-M. Wang, D. W. K. Ng, and M. H. Lee, "NOMA in downlink SDMA with limited feedback: Performance analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2281–2294, 2017.

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

[16] S. Schiessl, F. Naghibi, H. Al-Zubaidy, M. Fidler, and J. Gross, "On the delay performance of interference channels," in *IFIP Networking Conf.*, May 2016, pp. 216–224.

[17] 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.

[18] C. Xiao, J. Zeng, W. Ni, X. Su, R. P. Liu, T. Lv, and J. Wang, "Downlink MIMO-NOMA for ultra-reliable low-latency communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, Apr. 2019.

[19] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Transmission strategies in multiple-access fading channels with statistical QoS constraints," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1578–1593, 2012.

[20] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4243, Jul. 2014.

[21] J. Scarlett, V. Y. F. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-gaussian channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 81–92, Jan. 2017.

[22] E. MolavianJazi, "A unified approach to Gaussian channels with finite blocklength," Ph.D. dissertation, University of Notre Dame, Notre Dame, IN, 2014.

[23] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550–4564, Jul. 2018.

[24] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP J. Wireless Commun. and Networking*, vol. 2013, no. 1, paper 290, pp. 1–13, Dec. 2013.

[25] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, Nov. 2015, pp. 13–22.

[26] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Delay performance of wireless communications with imperfect CSI and finite-length coding," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6527–6541, Dec. 2018.
[27] S. Schiessl, J. Gross, M. Skoglund, and G. Caire, "Delay performance of the multiuser MISO downlink under imperfect CSI and finite-length coding," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 765–779, Apr. 2019.
[28] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
[29] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.
[30] A. Fréville, "The multidimensional 0–1 knapsack problem: An overview," *European Journal of Operational Research*, vol. 155, no. 1, pp. 1–21, 2004.
[31] G. B. Dantzig, "Discrete-variable extremum problems," *Operations research*, vol. 5, no. 2, pp. 266–288, 1957.
[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[33] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.
[34] M. Skoglund, J. Giese, and S. Parkvall, "Code design for combined channel estimation and error protection," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1162–1171, May 2002.
[35] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 840–845, 2003.
[36] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.
[37] N. Petreska, H. Al-Zubaidy, R. Knorr, and J. Gross, "On the recursive nature of end-to-end delay bound for heterogenous networks," in *IEEE Int. Conf. Commun. (ICC)*, Jun. 2015.

Sebastian Schiessl received his Ph.D. degree from KTH Royal Institute of Technology, Stockholm, Sweden in June 2019. He remained at the KTH Division of Information Science and Engineering to work on a project on edge computing applications, focusing on SDR implementations of both IEEE 802.11 and 4G/LTE systems. In January 2020, he joined u-blox AG to advance the standardization of next-generation vehicle-to-everything (V2X) communication systems within the IEEE 802.11bd task group.

Sebastian also holds a Dipl.-Ing degree in Electrical Engineering from Technical University of Munich, Germany. He took part in a student exchange at the University of Illinois at Urbana-Champaign and was a guest researcher at the Technical University of Berlin, Germany. His research interests are in the area of ultra-reliable low-latency wireless communication systems, where he combines information theory with queueing theory to study and optimize the overall system performance.

Mikael Skoglund (S'93-M'97-SM'04-F'19) received the Ph.D. degree in 1997 from Chalmers University of Technology, Sweden. In 1997, he joined the Royal Institute of Technology (KTH), Stockholm, Sweden, where he was appointed to the Chair in Communication Theory in 2003. At KTH, he heads the Division of Information Science and Engineering.

Dr. Skoglund has worked on problems in source-channel coding, coding and transmission for wireless communications, Shannon theory, information and control, and statistical signal processing. He has authored and co-authored some 150 journal and 350 conference papers.

Dr. Skoglund is a Fellow of the IEEE. During 2003–08 he was an associate editor for the IEEE Transactions on Communications and during 2008–12 he was on the editorial board for the IEEE Transactions on Information Theory. He has served on numerous technical program committees for IEEE sponsored conferences, he was general co-chair for IEEE ITW 2019, and he will serve as TPC co-chair for IEEE ISIT 2022.

James Gross received his Ph.D. degree from TU Berlin in 2006. From 2008-2012 he was assistant professor at RWTH Aachen University, and also research associate of the DFG-funded UMIC Research Centre of RWTH. Since November 2012, he has been with the Electrical Engineering and Computer Science School, KTH Royal Institute of Technology, Stockholm, where he is professor for machine-to-machine communications. He served as director for the ACCESS Linnaeus Centre from 2016 to 2019, while he is currently a member of the board of KTHs Innovative Centre for Embedded Systems. His research interests are in the area of mobile systems and networks, with a focus on critical machine-to-machine communications, cellular networks, resource allocation, as well as performance evaluation methods. He has authored over 150 (peer-reviewed) papers in international journals and conferences. His work has been awarded multiple times, including the best paper awards at ACM MSWiM 2015, IEEE WoWMoM 2009, and European Wireless 2009. In 2007, he was the recipient of the ITG/KuVS dissertation award for his Ph.D. thesis. He is also co-founder of R3 Communications GmbH, a Berlin-based young company in the area of ultra-reliable low-latency wireless networking for industrial automation.