Delay Violation Probability and Effective Rate of Downlink NOMA over $\alpha$-$\mu$ Fading Channels

Vaibhav Kumar, Student Member, IEEE, Barry Cardiff, Senior Member, IEEE, Shankar Prakriya, Senior Member, IEEE, and Mark F. Flanagan, Senior Member, IEEE

Abstract

Non-orthogonal multiple access (NOMA) is a potential candidate to further enhance the spectrum utilization efficiency in beyond fifth-generation (B5G) standards. However, there has been little attention on the quantification of the delay-limited performance of downlink NOMA systems. In this paper, we analyze the performance of a two-user downlink NOMA system over generalized $\alpha$-$\mu$ fading in terms of delay violation probability (DVP) and effective rate (ER). In particular, we derive an analytical expression for an upper bound on the DVP and we derive the exact sum ER of the downlink NOMA system. We also derive analytical expressions for high and low signal-to-noise ratio (SNR) approximations to the sum ER, as well as a fundamental upper bound on the sum ER which represents the ergodic sum-rate for the downlink NOMA system. We also analyze the sum ER of a corresponding time-division-multiplexed orthogonal multiple access (OMA) system. Our results show that while NOMA consistently outperforms OMA over the practical SNR range, the relative gain becomes smaller in more severe fading conditions, and is also smaller in the presence a more strict delay quality-of-service (QoS) constraint.

I. INTRODUCTION

NOMA has drawn tremendous attention as a potential solution to the spectrum congestion problem in 5G and B5G wireless communication systems. However, the majority of available literature related to NOMA primarily considers the system performance based on Shannon-perspective metrics such as achievable rate and outage probability, as well as performance optimization using these metrics. In contrast to this, a delay-constrained QoS metric termed as effective rate (ER) was introduced in [1], which is defined as the maximum constant arrival rate that a service process can support under a statistical delay constraint. The delay violation probability (DVP) is a related metric, which indicates the probability that an information bit arriving at the source will not be successfully delivered to the destination within a certain delay bound. The ER performance of orthogonal multiple access (OMA) based wireless communication systems over different fading channel models has been well-studied in the past, but in the case of NOMA, the literature is mostly limited to analysis over Rayleigh, Rician and Nakagami-$m$ fading channels.

The performance analysis of OMA systems, including single and multiple-antenna architectures over different fading channels in terms of ER, has been presented in [2]–[11]. An optimal rate and power adaptation policy for the maximization of the ER of a secondary/unicast user in an average-interference-constrained underlay spectrum sharing system over Nakagami-$m$ fading channels was presented in [2]. Using a moment generating function approach, the ER analysis of a single user with independent but not necessarily identically distributed multiple-input single-output (MISO) channel assuming hyper Fox’s $H$ as well as generalized-$K$ fading was presented in [3]. Some other notable contributions also exist on the analysis of the ER for a single-user communication system over more flexible/generalized channel models including (MISO) Weibull [4], (MISO) $\eta$-$\mu$ [5], (MISO) $\alpha$-$\mu$ [6], $\kappa$-$\mu$ shadowed [7], composite $\alpha$-$\eta$-$\mu$/Gamma [8] and (MISO) Fisher-Snedecor $F$ fading [9]. Moreover, the ER analysis of a single-user equal gain combining receiver over $\alpha$-$\kappa$-$\mu$ and $\alpha$-$\eta$-$\mu$ fading channels was presented in [10]. Very
recently, the ER analysis of a single-user $L_\infty$-norm diversity combiner with adaptive transmission schemes over arbitrary generalized fading channels was presented in [11].

The impact of finite blocklength coding and age of information (AoI) on DVP, as well as the DVP analysis for a multiuser MISO OMA system, were presented in [12]–[15]. Using stochastic network calculus (SNC) and the Mellin transform, the impact of finite blocklength coding on the DVP of a point-to-point single-user communication system over the Rayleigh fading channels was analyzed in [12]. The optimal design criteria to minimize the DVP in a point-to-point communication system over the Rayleigh fading channel using finite blocklength coding were presented in [13]. The impact of finite blocklength coding on the DVP in a multiuser MISO downlink OMA system over Rayleigh fading channels with imperfect CSI was presented in [14]. The impact of average age of information on the DVP in a two-user time-slotted multiple access Rayleigh fading channel was presented in [15]. In [16], the DVP analysis for a large wireless network over Rayleigh fading channels was presented, where the transmitter locations are modeled using a Poisson point process.

The delay-constrained performance analysis of NOMA systems over conventional (Rayleigh, Rician and Nakagami-$m$) fading channels was presented in [17]–[27]. The analysis of the sum ER for a two-user downlink NOMA system in Rayleigh fading was first presented in [17], where a suboptimal power allocation policy was proposed for the maximization of the sum ER using a truncated channel inversion power control policy. A cross-layer optimal power allocation scheme for a half-duplex cooperative multi-user downlink NOMA system over Rayleigh fading channels was suggested in [18], where the optimization problem was shown to be quasi-concave and a bisection-based solution was proposed in order to maximize the ER of the user having the minimum ER. Building on [18], a similar max-min criterion based optimal power allocation scheme was suggested for a half-duplex/full-duplex cooperative two-user downlink NOMA system over Rayleigh fading channels in [19]. A comprehensive analysis of the ER for a multi-user downlink NOMA system with a per-user delay QoS requirement over Rayleigh fading was presented in [20] and it was shown that the OMA system achieves a higher sum ER in the low signal-to-noise ratio (SNR) regime, whereas the NOMA system prevails in the high-SNR regime. The performance analysis of a two-user uplink NOMA system over Rayleigh fading channels in terms of DVP was presented in [21], where the authors also presented an optimal transmit power allocation scheme. The ER and DVP analysis for multiuser downlink NOMA over Nakagami-$m$ Rician fading using SNC and the Mellin transform was presented in [22]. In that work, two different optimal power allocation schemes, namely MaxMinEC and MinMaxDVP, were also suggested to maximize the minimum ER and minimize the maximum DVP, respectively. In order to analyze the performance of a two-user downlink NOMA system under statistical QoS constraints in the finite-blocklength regime, and to reduce the complications associated with the ER maximization problem due to the existence of statistical delay-bounded and error-rate-bounded QoS constraints, the concept of “$\epsilon$-effective capacity” was introduced in [23]. For the analysis of the delay-constrained secrecy performance of a multiuser downlink NOMA system over Rayleigh fading channels, the notion of effective secrecy rate was introduced in [24]. The closed-form expression for the sum ER of a two-user uplink NOMA system over Rayleigh fading was presented in [25]. A closed-form expression for the sum ER in a multiuser downlink NOMA system in the finite-blocklength regime over Rayleigh fading channels was presented in [26]. The analysis of DVP for a two-user uplink NOMA system using finite blocklength coding over Rayleigh fading channels, considering imperfect channel state information (CSI), was presented in [27].

To the best of authors’ knowledge, the analysis of DVP and ER for NOMA systems in generalized fading channels have not yet been presented in the literature. In this paper, we analyze the performance of a two-user downlink NOMA system in terms of the DVP and the sum ER over $\alpha$-$\mu$ fading. The $\alpha$-$\mu$ distribution is a mathematically tractable and generalized distribution which provides a remarkable unification of the system’s performance over a variety of fast fading channel models including one-sided Gaussian, Chi-squared, Rayleigh, Nakagami-$m$, Erlang and Weibull fading. Against this background, the main contributions of this paper are listed below:

- We analyze the DVP of a two-user downlink NOMA system over $\alpha$-$\mu$ fading channels. Using SNC and the Mellin transform, we derive a closed-form upper bound on the DVP for each user, and use these expressions to demonstrate the effect of different channel parameters on the DVP.

2 In [22], although the authors considered a multiuser scenario, it was assumed that the users are grouped into multiple NOMA pairs. The analysis and the results were presented for an arbitrary NOMA pair while the problem of optimal user-pairing was not considered.

3 In [26], the authors considered a multiuser system consisting of multiple NOMA pairs, and showed that the NOMA pairs with more distinct channel conditions attain a higher sum ER.
We also derive exact closed-form expressions for the ER of each user in the downlink NOMA system\(^4\) as well as a corresponding downlink OMA system over \(\alpha-\mu\) channels, and show the effect of non-linearity, clustering, and delay parameters on the difference between the sum ER in NOMA and OMA systems. Furthermore, we also derive approximate expressions for the sum ER of downlink NOMA in the high-SNR and low-SNR\(^5\) regimes.

Moreover, using Jensen’s inequality, we derive an upper-bound on the sum ER of downlink NOMA which is independent of the delay-constraint, which represents the ergodic sum-rate of the NOMA system. By evaluating the difference between the ergodic sum-rate and sum ER (which we term as the rate loss) in the downlink NOMA system, we demonstrate that the rate loss increases with an increase in the SNR, the delay exponent and the severity of fading.

**II. System Model**

We consider a half-duplex downlink NOMA scenario, where a source \(S\) simultaneously communicates with two users \(U_s\) (the user with *strong* average channel condition) and \(U_w\) (the user with *weak* average channel condition). The channel fading coefficient of the \(S-U_i\) link \((i \in \{s, w\})\) is denoted by \(h_i\), and the corresponding channel gain is denoted by \(g_i = (\hat{h}_i)^2\) with \(\alpha\)-root-mean value denoted by \(\Omega_i\). It is assumed that both wireless links are \(\alpha-\mu\) distributed with *non-linearity* parameter denoted by \(\alpha\) and *clustering* parameter denoted by \(\mu\) \([28]\).

Assuming \(\alpha, \mu \in \mathbb{Z}_+\) (the set of all positive integers), and using \([29]\) eqns. (1), (2) and \([30]\) eqn. (8.352-1), p. 899, the cumulative distribution function (CDF) and probability density function (PDF) of the channel gain \(g_i\) are, respectively, given by

\[
F_{g_i}(x) = 1 - \exp \left( -\frac{\mu x^{0.5\alpha}}{\Omega_i^{\alpha}} \right) \sum_{j=0}^{\mu-1} \frac{\mu^j x^{0.5j\alpha}}{j!\Omega_i^{j\alpha}},
\]

\[
f_{g_i}(x) = \frac{\alpha \mu^\mu x^{\mu\alpha-1}}{2\Omega_i^{\mu\alpha} \Gamma(\mu)} \exp \left( -\frac{\mu x^{0.5\alpha}}{\Omega_i^{\alpha}} \right),
\]

where \(\Gamma(\cdot)\) denotes the Gamma function. Furthermore, we assume that \(\Omega_{w}^{\alpha} < \Omega_{s}^{\alpha}\). We also assume that (perfect) instantaneous CSI is available at \(U_i\) regarding the \(S-U_i\) link, while only the statistical CSI (comprising the values of \(\Omega_s\) and \(\Omega_w\)) is available at \(S\).

In the case of downlink NOMA, the source \(S\) transmits a superimposed symbol \(\sum_i \sqrt{a_i} \mathcal{E} x_i\) to both users, where \(\mathcal{E}\) denotes the energy budget per (superseded) symbol at the source, \(a_i\) denotes the power-allocation coefficient for user \(U_i\) with \(a_s < a_w\) and \(\sum_i a_i = 1\), and \(x_i\) represents the data-bearing (complex) constellation symbol intended for user \(U_i\) with unit energy. The received signal at user \(U_i\) is given by \(y_i = h_i(\sqrt{a_s} \mathcal{E} x_s + \sqrt{a_w} \mathcal{E} x_w) + z_i\), where \(z_i \sim \mathcal{C}\mathcal{N}(0, \sigma^2)\) denotes complex additive white Gaussian noise (AWGN). Upon reception of the signal, \(U_w\) decodes \(x_w\) treating the interference due to the presence of \(x_s\) as additional noise. On the other hand, \(U_s\) also decodes \(x_w\) first (by treating the interference due to \(x_s\) as additional noise) and then applies successive interference cancellation (SIC) to remove \(x_w\) from \(y_s\), and subsequently decodes \(x_s\). Therefore, the instantaneous received signal-to-interference-plus-noise ratio (SINR) and SNR to decode \(x_w\) and \(x_s\) are, respectively, given by \(\gamma_w = a_w p g_{min}/(a_s p g_{min} + 1)\) and \(\gamma_s = a_s p g_s\), where \(g_{min} \triangleq \min\{g_w, g_s\}\) and \(p \triangleq \mathcal{E}/\sigma^2\) (we will refer to \(p\) as the SNR in the remainder of this paper).

Using (1), (2) and a transformation of random variables, the PDF of \(g_{min}\) can be given by

\[
f_{g_{min}}(x) = \frac{\alpha \exp \left( -\frac{\mu x^{0.5\alpha}}{\Omega_s^{\alpha}} \right)}{2\Omega_s^{\mu\alpha} \Gamma(\mu)} \sum_{m=0}^{\mu-1} \frac{\mu^m x^{0.5m\alpha}}{m!\Omega_s^{m\alpha}} + \frac{\alpha \exp \left( -\frac{\mu x^{0.5\alpha}}{\Omega_w^{\alpha}} \right)}{2\Omega_w^{\mu\alpha} \Gamma(\mu)} \sum_{n=0}^{\mu-1} \frac{\mu^n x^{0.5n\alpha}}{n!\Omega_w^{n\alpha}},
\]

where \(\tilde{\Omega} = 1/\left(1/\Omega_s^{\alpha} + 1/\Omega_w^{\alpha}\right)\).

\(^4\)Note that this is different from \([22]\), where approximate expressions were only derived for the ER of the weak user over Nakagami-\(m\) and Rician fading.

\(^5\)Note that for low-SNR approximation in \([22]\), a first-order Taylor approximation of the form \(\ln(1 + x) \approx x; x \to 0\) was used. In this paper, we use a second-order Taylor approximation, i.e., \(f(x) \approx x f'(0) + 0.5 x^2 f''(0); \ x \to 0\), which helps in quantifying the minimum energy per bit required for reliable transmission of information and the wideband slope.
III. PERFORMANCE ANALYSIS

A. Delay violation probability in downlink NOMA

Assuming that the instantaneous traffic arrival for user $U_i \in \{s, w\}$ at $S$ in the k-th time slot is denoted by $\lambda_i(k)$ and the corresponding traffic departure is denoted by $d_i(k)$, the cumulative arrival and departure processes from time slot $\tau$ to $t-1$ can be represented by $A_i(\tau, t) = \sum_{k=\tau}^{t-1} \lambda_i(k)$ and $D_i(\tau, t) = \sum_{k=\tau}^{t-1} d_i(k)$, respectively. Therefore, the delay experienced in the successful delivery of the information bits (intended for user $U_i$) arriving at $S$ in time slot $t$ is given by

$$\vartheta_i(t) \triangleq \inf \{u \geq 0 : A_i(0, t) \leq D_i(0, t + u)\}.$$ 

Using [22, eqn. (17)], an upper bound on the probability of violation of a target delay $\vartheta$ is given by

$$\Pr(\vartheta_i > \vartheta) \leq \inf_{\mathcal{F} > 0} \left\{ \frac{[\mathcal{M}_{\varphi_s}(1 - \mathcal{F})]^{\vartheta}}{1 - \exp(\lambda \mathcal{F})[\mathcal{M}_{\varphi_s}(1 - \mathcal{F})]} \right\},$$

where $\mathcal{M}_X(\mathcal{F}, \tau, t) = \mathbb{E}[[X(\tau, t)]^{\mathcal{F}-1}]$ represents the Mellin transform of the random process $X(\tau, t)$, $\varphi_s \triangleq (1 + a_s \rho_g s)^N/\ln 2$, $\varphi_w \triangleq [1 + a_w \rho_{g_{\min}}/(1 + a_s \rho_{g_{\min}})]N/\ln 2$ and $N$ denotes the number of symbols transmitted by $S$ in each time slot.

**Theorem 1.** For the case of $U_s$, an analytical expression for $\mathcal{M}_{\varphi_s}(1 - \mathcal{F})$ can be given by

$$\mathcal{M}_{\varphi_s}(1 - \mathcal{F}) = \frac{\alpha \varpi \mu G_{\alpha,2+\alpha}^{\alpha,\alpha} \left[ \frac{\mu^2}{(s, \rho)^{-\frac{1}{2}}} \right]}{\sqrt{2(2\pi)^{\alpha-0.5} \Gamma(\alpha)}} \left[ \Delta_{(0,2)}[\alpha,1-0.5\mu] \right],$$

where $\varpi \triangleq N/\ln 2$, $\Delta(x, y) = \frac{y}{x}, \frac{y+1}{x}, \ldots, \frac{y+x-1}{x}$, and $G[\cdot]$ denotes Meijer’s $G$-function.

**Proof.** See Appendix A.

Substituting the expression for $\mathcal{M}_{\varphi_s}(1 - \mathcal{F})$ into (4), one can obtain an analytical expression for an upper bound on the DVP for $U_s$.

**Theorem 2.** For the case of the weak user, an analytical expression for $\mathcal{M}_{\varphi_w}(1 - \mathcal{F})$ can be given by

$$\mathcal{M}_{\varphi_w}(1 - \mathcal{F}) = \frac{1}{\Gamma(\mu)\Gamma(\varpi)\Gamma(-\varpi)} \left[ \frac{1}{\Omega_w^\mu} \sum_{m=0}^{\mu-1} \frac{\tilde{\Omega}^{\mu+m}}{m!\Omega_m^{\alpha}} \mathcal{H}(\varpi, m) + \frac{1}{\Omega_w^\mu} \sum_{n=0}^{\mu-1} \frac{\tilde{\Omega}^{\mu+n}}{n!\Omega_n^{\alpha}} \mathcal{H}(\varpi, n) \right],$$

where $\mathcal{H}(x, y)$ denotes the extended generalized bivariate Fox’s $H$-function (EGBFH)\(^6\) i.e.,

$$\mathcal{H}(x, y) = \mathcal{H}_b^\alpha \left[ \begin{array}{c} (1-(\mu+y), \frac{2}{\alpha}; \frac{2}{\alpha}); \frac{2}{\alpha}; \frac{2}{\alpha} \\ (1-x, 1) \end{array} \right] \left[ \begin{array}{c} (1+x, 1); \frac{2}{\alpha}; \frac{2}{\alpha} \\ (0, 1) \end{array} \right] \rho \left( \frac{\Omega}{\mu} \right)^\frac{\alpha}{\mu}, a_s \rho \left( \frac{\Omega}{\mu} \right)^\frac{2}{\alpha},$$

$a = 1, 0 : 1, 1 : 1, 1$ and $b = 0, 1 : 1, 1 : 1, 1$.

**Proof.** See Appendix B.

Substituting the expression for $\mathcal{M}_{\varphi_w}(1 - \mathcal{F})$ into (4), one can obtain an analytical expression for an upper bound on the DVP for $U_w$. It is important to note that in [22], an approximate expression for $f_{g_{\min}}(x)$ had been used to obtain the analytical expression for $\mathcal{M}_{\varphi_w}(1 - \mathcal{F})$ (over Nakagami-$m$ and Rician fading); in contrast, we use the exact expression for $f_{g_{\min}}(x)$ to find the closed-form expression for $\mathcal{M}_{\varphi_w}(1 - \mathcal{F})$ over $\alpha$-$\mu$ fading.

\(^6\)Computationally efficient implementations of EGBFH using MATHEMATICA and MATLAB are given in [31] and [32], respectively.
B. Effective rate of downlink NOMA

The ER for $U_i$ can be defined as (c.f. [20])

$$R_{i,\text{NOMA}} = -\frac{1}{\nu} \log_2 \left[ \mathbb{E}_{\gamma_i} \{(1 + \gamma_i)^{-\nu}\} \right],$$

where $\nu \triangleq \theta TB / \ln 2$, $\theta$ represents the delay QoS exponent, $T$ denotes the length of each fading block (which is assumed to be same for both $S-U_s$ and $S-U_w$ links and is an integer multiple of the symbol duration, with a further assumption that the durations of both user symbols are the same), and $B$ denotes the total bandwidth.

Therefore, using (7), the ER for $U_s$ is given by

$$R_{s,\text{NOMA}} = -\frac{1}{\nu} \log_2 \left[ \int_0^\infty (1 + a_s \rho x)^{-\nu} f_{g_s}(x) dx \right].$$

Substituting the expression for $f_{g_s}(x)$ from (2) into the preceding equation and following the arguments in Appendix A, an analytical expression for the ER of $U_s$ is given by

$$R_{s,\text{NOMA}} = -\frac{1}{\nu} \log_2 \left\{ \frac{\alpha^\nu \mu^2}{\sqrt{2(2\pi)^{\alpha-0.5}}} \left[ \frac{\Delta(2,0)}{\Delta(\alpha,1-0.5\alpha\mu)} \right] \right\},$$

where $\chi = \Delta(\alpha,1-0.5\alpha\mu)$ and $\xi = \Delta(\alpha,\nu - 0.5\alpha\mu)$. On the other hand, using (7), the ER of $x_w$ can be given by

$$R_{w,\text{NOMA}} = -\frac{1}{\nu} \log_2 \left[ \int_0^\infty (1 + \rho x)^{-\nu} (1 + a_s \rho x)^\nu f_{g_{min}}(x) dx \right].$$

Substituting the expression for $f_{g_{min}}(x)$ from (5) in the preceding expression and following the arguments in Appendix B, an analytical expression for the ER of $U_w$ is given by

$$R_{w,\text{NOMA}} = -\frac{1}{\nu} \log_2 \left[ \frac{1}{\Gamma(\mu)\Gamma(\nu)} \left\{ \sum_{m=0}^{\mu-1} m! \Omega_{\mu+m} H(\nu,m) + \sum_{n=0}^{\mu-1} n! \Omega_{\mu+n} H(\nu,n) \right\} \right].$$

Therefore, the sum ER of the downlink NOMA system is given by adding (8) and (9), i.e., $R_{\text{sum,NO}M} = R_{s,\text{NOMA}} + R_{w,\text{NOMA}}$. Similar to the case of DVP, an approximate expression for $f_{g_{min}}(x)$ was used to find the analytical expressions for the ER of the weak user over Nakagami-$m$ and Rician fading in [22]. In contrast, we use the exact expression for $f_{g_{min}}(x)$ to derive the closed-form expression for $R_{w,\text{NOMA}}$ over $\alpha-\mu$ fading channels.

C. High-SNR approximation for the ER of downlink NOMA

Although the analytical expressions for the ER derived in the previous subsection are exact, they are computationally expensive. Therefore, we derive a high-SNR ($\rho \gg 1$) approximation for the ER of downlink NOMA in this subsection. For the case when $\rho \gg 1$, the ER of $U_s$ can be approximated as

$$R_{s,\text{NOMA}} \approx -\frac{1}{\nu} \log_2 \left[ \mathbb{E}_{\gamma_s} \{ (a_s \rho g_s)^{-\nu} \} \right] = \log_2(a_s \rho) - \frac{1}{\nu} \log_2 \left[ \int_0^\infty x^{-\nu} f_{g_s}(x) dx \right].$$

Substituting the expression for $f_{g_s}(x)$ from (2) into the preceding expression and solving the integral using [30, eqn. (3.326-2), p. 377], the preceding approximation can be written in closed-form as

$$R_{s,\text{NOMA}} \approx \log_2(a_s \rho) - \frac{1}{\nu} \log_2 \left[ \left( \frac{\mu^{1/\alpha}}{\Omega_s} \right)^{2\nu} \frac{\Gamma \left( \mu - \frac{2\nu}{\alpha} \right)}{\Gamma(\mu)} \right],$$

where the integral holds good only for the case when $\alpha\mu > 2\nu$. On the other hand, it is straightforward to show that for $\rho \gg 1$, the ER of $U_w$ can be approximated as

$$R_{w,\text{NOMA}} \approx \log_2 \left( 1 + \frac{a_w}{a_s} \right).$$

Note that the high-SNR approximation of $R_{w,\text{NOMA}}$ does not depend on the QoS exponent $\theta$ or on the channel parameters ($\alpha$ and $\mu$). A high-SNR approximation for $R_{\text{sum,NO}M}$ can be obtained by adding (10) and (11). It can also be shown that the high-SNR approximation of the ER of $U_w$ is equal to the high-SNR approximation of the ergodic rate of $U_w$, which means that at high SNR, the effective rate of the weak user becomes (almost) equal to the ergodic rate, regardless of the delay QoS exponent.
D. Low-SNR approximation for the ER of downlink NOMA

Using a second-order Taylor expansion around $\rho \to 0$, an approximation for the ER of $U_i$ is given by

$$R_{i,\text{NOMA}} = \rho \dot{R}_{i,\text{NOMA}} + 0.5 \rho^2 \ddot{R}_{i,\text{NOMA}} + O(\rho^2),$$

(12)

where $\dot{R}_{i,\text{NOMA}}$ and $\ddot{R}_{i,\text{NOMA}}$ represent the first-order and second-order derivative of $R_{i,\text{NOMA}}$ (shown in (7)) with respect to $\rho$ and evaluated at $\rho \to 0$, respectively.

Theorem 3. The derivatives for $U_s$ and $U_w$ are given by

$$\begin{align*}
\dot{R}_{s,\text{NOMA}} &= \log_2(e) a_s \mathbb{E}\{g_s\}, \\
\ddot{R}_{s,\text{NOMA}} &= \log_2(e) a_s^2 \nu (\mathbb{E}\{g_s\})^2 - (\nu + 1) \mathbb{E}\{g_s^2\}, \\
\dot{R}_{w,\text{NOMA}} &= \log_2(e) a_w \mathbb{E}\{g_{\min}\}, \\
\ddot{R}_{w,\text{NOMA}} &= \log_2(e) a_w \left[ \nu a_w (\mathbb{E}\{g_{\min}\})^2 - \{ (\nu + 1) a_w + 2a_s \} \mathbb{E}\{g_{\min}^2\} \right].
\end{align*}$$

(13)

Proof. See Appendix C

It can be noted from (13) that $\dot{R}_{i,\text{NOMA}}$ is independent of the delay QoS exponent $\theta$, and is in fact equal to the first derivative of the ergodic rate evaluated at $\rho \to 0$. It can also be noted from (13) that $\ddot{R}_{i,\text{NOMA}}$ decreases as the value of $\theta$ increases\(^7\) (i.e., for a system with more stringent delay constraints). It is also noteworthy that following the arguments in [33], it can be shown that the first derivative $\dot{R}_{i,\text{NOMA}}$ is related to the minimum energy per bit required for reliable transmission of information, denoted by $\left(\mathcal{E}_{b,i}/\sigma^2\right)_{\min}$, as follows:

$$\left(\frac{\mathcal{E}_{b,i}}{\sigma^2}\right)_{\min} = \frac{1}{R_{i,\text{NOMA}}},$$

(14)

whereas the second derivative $\ddot{R}_{i,\text{NOMA}}$ is related to the wideband slope, denoted by $S_{0,i}$, via

$$S_{0,i} = -2 \frac{(\dot{R}_{i,\text{NOMA}})}{R_{i,\text{NOMA}}} \ln 2.$$

(15)

Therefore, (13)-(15) imply that with an increase in the value of the delay QoS exponent, the minimum energy per bit required for reliable transmission does not change, but the wideband slope vanishes. The relevant mean values in (13) can be obtained using (2), (3) and [30] eqn. (3.326-2), p. 337, as follows:

$$\begin{align*}
\mathbb{E}\{g_s\} &= \frac{\Omega_s^2}{\mu^2/\alpha \Gamma(\mu)} \Gamma \left( \mu + \frac{2}{\alpha} \right), \quad \mathbb{E}\{g_s^2\} = \frac{\Omega_s^4}{\mu^4/\alpha \Gamma(\mu)} \Gamma \left( \mu + \frac{4}{\alpha} \right), \\
\mathbb{E}\{g_{\min}\} &= \frac{1}{\Omega_s^2 \Gamma(\mu)} \sum_{m=0}^{\mu-1} \tilde{\Omega}_{w}^{\mu+m+(2/\alpha)} \Gamma \left( \mu + m + \frac{2}{\alpha} \right) + \frac{1}{\Omega_w^2 \Gamma(\mu)} \sum_{n=0}^{\mu-1} \tilde{\Omega}_{s}^{\mu+n+(2/\alpha)} \Gamma \left( \mu + n + \frac{2}{\alpha} \right), \\
\mathbb{E}\{g_{\min}^2\} &= \frac{1}{\Omega_s^4 \Gamma(\mu)} \sum_{m=0}^{\mu-1} \tilde{\Omega}_{w}^{\mu+m+(4/\alpha)} \Gamma \left( \mu + m + \frac{4}{\alpha} \right) + \frac{1}{\Omega_w^4 \Gamma(\mu)} \sum_{n=0}^{\mu-1} \tilde{\Omega}_{s}^{\mu+n+(4/\alpha)} \Gamma \left( \mu + n + \frac{4}{\alpha} \right).
\end{align*}$$

Therefore, using (12), (13) and the preceding expressions, one can find a closed-form low-SNR approximation for $R_{\text{sum,NOMA}}$.

\(^7\)This occurs because $\nu \propto \theta$, and $\text{Var}(g_s) > 0 \Longrightarrow \mathbb{E}\{g_s^2\} > (\mathbb{E}\{g_s\})^2$. The same is also true for $g_{\min}$ also.
E. Upper bound on the ER of downlink NOMA

Using the concavity of the logarithmic function and Jensen’s inequality, it follows from (7) that

\[ R_{i,NOMA} \leq - \frac{1}{\nu} \mathbb{E}_{\gamma_i} \{ \log_2 [(1 + \gamma_i)^{-\nu}] \} = \mathbb{E}_{\gamma_i} \{ \log_2 (1 + \gamma_i) \} \triangleq C_{i,NOMA}. \]

(16)

Note that \( C_{i,NOMA} \) does not depend on the delay exponent \( \theta \) and represents the ergodic rate of \( U_i \), and that equality holds for the case when \( \theta \to 0 \), i.e., a system without any delay QoS constraint. Using (16), the upper bound on the ER of \( U_s \) is given by

\[ C_{s,NOMA} = \int_0^\infty \log_2(1 + a_s \rho x) f_{g_s}(x) dx. \]

Substituting the expression for \( f_{g_s}(x) \) from (2) into the preceding equation and following the arguments in Appendix A, an analytical expression for \( C_{s,NOMA} \) is obtained as

\[ C_{s,NOMA} = \frac{\mu^\mu G^{2+2\alpha,\alpha}}{\sqrt{2(\ln 2)}(2\pi)^{\alpha-0.5} \Gamma(\mu)} \begin{Bmatrix} \zeta, \chi, \psi \end{Bmatrix}_\Delta(2,0), 
\]

(17)

where \( \zeta \triangleq \Delta(\alpha, -0.5\alpha\mu) \). Similarly, using (16), an upper bound on the ER of \( U_w \) is given by

\[ C_{w,NOMA} = \int_0^\infty \log_2(1 + \rho x) f_{g_{\min}}(x) dx - \int_0^\infty \log_2(1 + a_s \rho x) f_{g_{\min}}(x) dx. \]

Substituting the expression for \( f_{g_{\min}}(x) \) and following the arguments in Appendix A, an analytical expression for \( C_{w,NOMA} \) is given by

\[ C_{w,NOMA} = \frac{1}{\sqrt{2(\ln 2)}(2\pi)^{\alpha-0.5} \Gamma(\mu)} \times \left[ \sum_{m=0}^{\mu-1} \frac{\mu^{\mu+m}}{m! \Omega_s^\alpha} \left\{ \mathcal{G}(\rho, m) \frac{\rho^{0.5\alpha(\mu+m)}}{(a_s \rho)^{0.5\alpha(\mu+m)}} \right\} + \sum_{n=0}^{\mu-1} \frac{\mu^{\mu+n}}{n! \Omega_s^\alpha} \left\{ \mathcal{G}(\rho, n) \frac{\rho^{0.5\alpha(\mu+n)}}{(a_s \rho)^{0.5\alpha(\mu+n)}} \right\} \right] \]

(18)

where

\[ \mathcal{G}(x, y) = G^{2+2\alpha,\alpha} \frac{1}{x^\alpha} \left[ \frac{\mu}{2\Omega} \right]^{2} \begin{Bmatrix} \psi_x, \phi_x \end{Bmatrix}_\Delta(2,0), \psi_y, \phi_y \]

\[ \psi_y \triangleq \Delta(\alpha, -0.5\alpha\{\mu + y\}) \text{ and } \phi_y \triangleq \Delta(\alpha, 1 - 0.5\alpha\{\mu + y\}). \]

Therefore using (17) and (18), an analytical upper bound on the sum ER (representing the ergodic sum-rate of the downlink NOMA system) can be obtained as

\[ C_{\text{sum,NOMA}} = C_{s,NOMA} + C_{w,NOMA}. \]

F. Effective rate of downlink OMA

In order to quantify the performance difference between the downlink NOMA system and the corresponding OMA system, we analyze the ER of the downlink OMA system in this subsection. In the case of OMA, the source transmits \( \sqrt{\mathbb{E}_x x_i} \) to \( U_i \) in orthogonal time slots. Therefore, the instantaneous received SNR at \( U_i \) to decode \( x_i \) is given by \( \gamma_i = \rho g_i \). The ER for the user \( U_i \) in the downlink OMA system is therefore given by

\[ R_{s,OMA} = -\frac{1}{\nu} \log_2 [\mathbb{E}_{\gamma_i} \{ (1 + \gamma_i)^{-0.5\nu} \}] = -\frac{1}{\nu} \log_2 \left[ \int_0^\infty (1 + \rho x)^{-0.5\nu} f_{g_s}(x) dx \right]. \]

Substituting the expression for \( f_{g_s}(x) \) from (2) and following the arguments in Appendix A yields

\[ R_{s,OMA} = -\frac{1}{\nu} \log_2 \left[ a_s^{0.5\nu} \mu^\mu G^{2+2\alpha,\alpha} \frac{\mu^2}{4\nu^2 \Omega_s^\alpha} \Delta(2,0), \xi \right] \]

(19)

where \( \xi \triangleq \Delta(\alpha, 0.5\nu - 0.5\alpha\mu) \). An analytical expression for the ER of \( U_w \) in the downlink OMA system, denoted by \( R_{w,OMA} \), can be obtained by replacing \( \Omega_s \) with \( \Omega_w \) in (19). Finally, the sum ER can be obtained as \( R_{\text{sum,OMA}} = R_{s,OMA} + R_{w,OMA} \).
To analyze the performance of the system, we assume that $BT = 1$ and $\Omega_s = 1$. First we analyze the effect of $\Omega_w$ on the ER of NOMA, and also on the difference between the performance of the NOMA and OMA systems. Fig. 1 shows numerically evaluated results for the ER of the strong and weak user, as well as the sum ER of downlink NOMA, for $\Omega_s = 1$ and different values of $\Omega_w$. It can be noted from the figure that as the value of $\Omega_w$ increases, $R_{w,\text{NOMA}}$ also increases in the low-to-mid SNR regime. This happens because an increase in the value of $\Omega_w$ results in a better (average) channel quality between the source $S$ and the weak user $U_w$. Interestingly, for high SNR values, $R_{w,\text{NOMA}}$ becomes (almost) constant regardless of the value of $\Omega_w$, which is in line with (11). As for all the cases of $\Omega_w$ we have considered the value of $\Omega_s$ in the corresponding system to be fixed at 1, the value of $R_{\text{sum,\text{NOMA}}}$ is larger for larger values of $\Omega_w$ in the low-to-mid SNR regime and converges to the same value in the high-SNR regime. For the same system parameters, Fig. 2 shows the numerically computed results on the difference between the sum ER of downlink NOMA and the corresponding downlink OMA system, i.e., $R_{\text{sum,\text{NOMA}}} - R_{\text{sum,\text{OMA}}}$. It can be noted form the figure that this difference is larger for smaller values of $\Omega_w$, i.e., the case when the channel quality of the $S-U_s$ link and $S-U_w$ links are significantly different. This occurs because, for the case when $\Omega_s$ is fixed at 1, although a larger value of $\Omega_w$ in the case of NOMA results in a higher sum ER in the low-to-mid SNR range, the corresponding increase in the ER of the weak user in the case of
OMA is much larger, resulting in a smaller value of $R_{\text{sum,NOMA}} - R_{\text{sum,OMA}}$. Therefore, in the remaining figures in this paper, we will consider $\Omega_w = 0.1$. Assuming the target data rate for both $U_s$ and $U_w$ to be $r_{\text{target}} = 2 \text{ bpcu}$ and following the arguments in [34], a legitimate range for $a_s$ is given by $0 < a_s < 1/2^{r_{\text{target}}}$. Therefore, we use a one-dimensional numerical search over the discrete set $a_s \in \{0.01, 0.02, \ldots, 0.24\}$ in order to maximize the sum ER of downlink NOMA. For $\rho$ in the range from 10 dB to 40 dB, the optimal value of $a_s$ to maximize $R_{\text{sum,NOMA}}$ is found to be 0.24, which is the maximum value considered for numerical evaluation. This occurs because most of the ER in $R_{\text{sum,NOMA}}$ is obtained by $R_{s,\text{NOMA}}$ (as the link between $S$ and $U_s$ is comparatively stronger). Therefore, throughout this section, we use $a_s = 0.24$.

Fig. 7 shows the variation of the ER of downlink NOMA for $U_s$ and $U_w$ for different $\alpha$ and $\mu$ w.r.t. the power-allocation coefficient. It can be noted from the figure that although the ER of $U_w$ decreases linearly with increasing value of $a_s$, the corresponding increase in the ER of $U_s$ is non-linear, resulting in a significant increase in $R_{\text{sum,NOMA}}$. Similar results were shown in [17, Figs. 5 and 6] for the case of Rayleigh fading.

Fig. 4 shows the variation of the exact DVP and the corresponding upper bound w.r.t. the target delay $\vartheta$. It can be noted from the figure that the delay violation probability for $U_s$ is small as compared to that of $U_w$, because of the higher departure rate of the data intended for $U_s$. The figure also shows that with an increase in the data arrival rate (which results in a higher building rate of the queue), the delay violation probability increases for both $U_s$ and $U_w$. It is noteworthy that although the derived upper bounds on the probability of delay violation are not very
tight (bounds of a similar nature were derived in [12], [13], [16], [21], [22] and [27]) the slope of the bounds are almost identical to that of the actual (simulated) curves. Therefore, these bounds can give good insights regarding the rate of decay of the delay violation probability curve.

Figs. 5 and 6 show the upper bound on the delay violation probability for different values of $\alpha$ (with $\mu$ fixed at 1) and for different values of $\mu$ (with $\alpha$ fixed at 2), respectively. It can be noted from these figures that with an increase in the value of $\alpha$ or $\mu$, the probability of delay violation decreases at a higher rate. A large value of $\alpha$ (with fixed $\mu$) or $\mu$ (with fixed $\alpha$) represents a less severe fading, resulting in a higher departure rate of the data at the queue at $S$.

Figs. 7 and 8 show the approximations for the sum ER of the downlink NOMA system for $\rho \gg 1$ (high-SNR) and $\rho \to 0$ (low-SNR), respectively. The figures demonstrate a linear growth in the high-SNR regime and an exponential growth in the low-SNR regime for the sum ER. An excellent match between the exact (numerical)

8Note that for $\mu = 1$ the $\alpha-\mu$ distribution converges to the Weibull distribution with shape parameter $\alpha$, for which the amount of fading is given by $\lambda F = [\Gamma(1 + \frac{1}{\alpha})/\Gamma(1 + \frac{2}{\alpha})]^2 - 1$. Similarly, for $\alpha = 2$ the $\alpha-\mu$ fading model is equivalent to Nakagami-$m$ fading with shape parameter $m = \mu$, for which the amount of fading is given by $\lambda F = 1/\mu$. 
and the approximated (closed-form) results (in the relevant SNR regimes) confirms the correctness of the derived analytical approximations.

Figs. 9 and 10 show the difference between the ergodic sum-rate and the sum ER (i.e., the rate loss $C_{\text{sum,NOMA}} - R_{\text{sum,NOMA}}$) for the downlink NOMA system, with varying delay QoS exponent $\theta$, for different values of $\alpha$ (with $\mu = 1$) and $\mu$ (with $\alpha = 2$), respectively. It can be noted from Fig. 9 that the rate loss decreases with increasing $\alpha$. Therefore, Fig. 9 indicates that a given delay QoS target will result in a more detrimental impact on the effective capacity in more severe fading conditions. Also, for a fixed value of $\alpha$ (with $\mu = 1$), the capacity loss becomes more pronounced as the SNR increases. Fig. 10 shows the rate loss $C_{\text{sum,NOMA}} - R_{\text{sum,NOMA}}$ for different values of $\mu$ with $\alpha = 2$. Effects qualitatively similar to those observed in the previous case can be noticed here also; the loss in capacity increases with the severity of fading and also with the operational SNR.

Figs. 11 and 12 demonstrate the difference between the sum ER of downlink NOMA and the corresponding OMA system (i.e., $R_{\text{sum,NOMA}} - R_{\text{sum,OMA}}$) for different values of $\alpha$ (with $\mu = 1$) and $\mu$ (with $\alpha = 2$), respectively, with varying SNR $\rho$. It is clear from Fig. 11 that for a fixed value of $\alpha$ and $\theta$, the performance gap between downlink NOMA and the corresponding OMA system increases (with NOMA outperforming OMA) in the low-to-mid-SNR regime and then this difference decreases in the high-SNR regime. Also, while NOMA outperforms OMA in most of the practical SNR range, the gain is smaller in more severe fading conditions (smaller $\alpha$ or $\mu$). It can also be seen from both figures that in the presence of a more stringent delay QoS constraint, the performance advantage of NOMA over OMA is considerably lessened. In particular, it can be seen that for the case of the stricter delay constraint ($\theta = 2$) and severe fading ($\alpha, \mu = (2, 1)$), the difference in achievable rate performance between NOMA and OMA is negligible.

V. Conclusion

In this paper, we analyzed the performance of a delay-constrained two-user downlink NOMA system over $\alpha$-$\mu$ fading. We derived an upper bound on the delay violation probability for each user, and an exact analytical expression for the sum effective rate of the downlink NOMA system over the generalized $\alpha$-$\mu$ channel, which as special cases yields the sum ER of the downlink NOMA system over a variety of channel models of practical interest, including Rayleigh, Nakagami-m, Weibull, Erlang, and Chi-squared fading. Analytical expressions for the approximations to the sum ER of downlink NOMA in high-SNR and low-SNR regimes were also derived, which are accurate in the corresponding regions. Simulation results indicate that the delay violation probability for the strong user is less compared to that of the weak user, and this probability of delay violation decreases with a decrease in the arrival data rate at the source as well as with the severity of fading. The results also indicate that for the case of NOMA, with increasing value of delay exponent the difference between ergodic sum-rate and sum ER increases very rapidly, but this difference decreases as the fading becomes less severe. Moreover, for a fixed channel parameters and delay exponent, this rate loss increases at high SNR. Finally, the difference between the sum ER of downlink NOMA and the corresponding OMA decreases for large values of delay exponent $\theta$ as well as under severe fading conditions.

APPENDIX A

PROOF OF THEOREM 11

For the case of $U_s$, we have

$$M_{\varphi_s}(1 - \mathcal{F}) = \mathbb{E}_{g_s}(\{(1 + a_s\rho g_s)^{-\varpi}\} = \int_0^\infty (1 + a_s\rho x)^{-\varpi} f_{g_s}(x)dx,$$

where $\varpi \triangleq N_\mathcal{F}/\ln 2$. Substituting the expression for $f_{g_s}(x)$ from (2) into the preceding equation, we obtain

$$M_{\varphi_s}(1 - \mathcal{F}) = \frac{\alpha \mu}{2\Omega^\varpi} \int_0^\infty x^{0.5\alpha - 1} (1 + a_s\rho x)^{-\varpi} \exp\left(-\frac{\mu x^{0.5\alpha}}{\Omega^\varpi}\right) dx.$$

Using the relations $(1 + x)^y = \frac{1}{(1-y)^y} \sum_{\nu=0}^\infty \frac{\nu! y^\nu}{\nu!(1-y)^\nu}$ and $\exp(-x) = \sum_{\nu=0}^\infty \frac{x^{\nu}}{\nu!}$ from (35) eqn. (10) and (35) eqn. (11), respectively, it follows from the preceding equation that

$$M_{\varphi_s}(1 - \mathcal{F}) = \frac{\alpha \mu}{2\Omega^\varpi} \int_0^\infty x^{0.5\alpha - 1} \sum_{\nu=0}^\infty \frac{\nu! y^\nu}{\nu!(1-y)^\nu} \frac{\nu! y^\nu}{\nu!(1-y)^\nu} \sum_{\nu=0}^\infty \frac{x^{\nu}}{\nu!} dx.$$

Solving the integral given in the equation above using (35) eqn. (21), an analytical expression for $M_{\varphi_s}(1 - \mathcal{F})$ can be given by (5); this concludes the proof.
APPENDIX B  
PROOF OF THEOREM \[2\]  
For the case of the weak user, we have  
\[
\mathcal{M}_{\varphi_w}(1 - \mathcal{S}) = E_{g_{\text{min}}} \left\{ \left( 1 + \frac{a_w \rho g_{\text{min}}}{a_s \rho g_{\text{min}} + 1} \right)^{-\varpi} \right\} = \int_0^\infty (1 + \rho x)^{-\varpi} \left( 1 + a_s \rho x \right)^{\varpi} f_{g_{\text{min}}}(x)dx.
\]
Substituting the expression for \( f_{g_{\text{min}}}(x) \) from (3) into the preceding equation, we obtain  
\[
\mathcal{M}_{\varphi_w}(1 - \mathcal{S}) = \frac{\alpha}{2 \Omega_{s}^{\mu} \Gamma(\mu) \Gamma(\nu) \Gamma\left(-\varpi\right)} \sum_{m=0}^{\mu \mu + m} \frac{\mu^{\mu + m}}{m! \Omega_{s}^{\mu \nu \alpha}} \int_0^\infty (1 + \rho x)^{-\varpi} \left( 1 + a_s \rho x \right)^{\varpi} \exp \left( - \frac{\mu x 0.5 \alpha}{\Omega} \right) x^{0.5 \alpha (\mu + m) - 1} dx \times G_{1,1}^{1,1} \left[ 1 + a_s \rho x \left| \begin{array}{c} \frac{1 + \varpi}{0} \end{array} \right. \right],
\]
Using [35] eqns. (10) and (11), it follows from the preceding equation that  
\[
\mathcal{M}_{\varphi_w}(1 - \mathcal{S}) = \frac{\alpha}{2 \Omega_{s}^{\mu} \Gamma(\mu) \Gamma(\nu) \Gamma\left(-\varpi\right)} \sum_{m=0}^{\mu \mu + m} \frac{\mu^{\mu + m}}{m! \Omega_{s}^{\mu \nu \alpha}} \int_0^\infty x^{0.5 \alpha (\mu + m) - 1} G_{1,1}^{1,1} \left[ 1 + a_s \rho x \left| \begin{array}{c} \frac{1 + \varpi}{0} \end{array} \right. \right].
\]
Using the relation between Meijer’s G-function and Fox’s H-function given in [36] eqn. (6.2.8), the preceding equation yields  
\[
\mathcal{M}_{\varphi_w}(1 - \mathcal{S}) = \frac{\alpha}{2 \Omega_{s}^{\mu} \Gamma(\mu) \Gamma(\nu) \Gamma\left(-\varpi\right)} \sum_{m=0}^{\mu \mu + m} \frac{\mu^{\mu + m}}{m! \Omega_{s}^{\mu \nu \alpha}} \int_0^\infty x^{0.5 \alpha (\mu + n) - 1} H_{1,1}^{1,1} \left[ 1 + a_s \rho x \left| \begin{array}{c} \frac{1 + \varpi}{0} \end{array} \right. \right],
\]
Using the substitution \( \tilde{x} = x^{0.5 \alpha} \) and solving the resulting integrals using [37] eqn. (2.3), the analytical expression for \( \mathcal{M}_{\varphi_w}(1 - \mathcal{S}) \) reduces to (6); this completes the proof.

APPENDIX C  
PROOF OF THEOREM \[3\]  
For the case of the strong user, from (7) it follows that  
\[
\dot{R}_{s,\text{NOMA}} \triangleq \left. \frac{dR_{s,\text{NOMA}}}{d\rho} \right|_{\rho \to 0} = \left. \frac{d}{d\rho} \left( - \frac{1}{\nu} \log_2 \left[ \mathbb{E}_{g_s} \left\{ (1 + a_s \rho g_s)^\nu \right\} \right] \right) \right|_{\rho \to 0} = - \frac{\log_2(\nu)}{\nu} \times \mathbb{E}_{g_s} \left\{ (1 + a_s \rho g_s)^{-\nu} a_s g_s \right\},
\]
and  
\[
\dot{R}_{s,\text{NOMA}} = \left. \frac{d^2 R_{s,\text{NOMA}}}{d\rho^2} \right|_{\rho \to 0} = - \frac{\log_2(\nu)}{\nu} \left( (1 + a_s \rho g_s)^{-\nu} \mathbb{E}_{g_s} \left\{ \nu (\nu + 1)(1 + a_s \rho g_s)^{-\nu - 2} a_s^2 g_s^2 \right\} - (\mathbb{E}_{g_s} \left\{ (1 + a_s \rho g_s)^{-\nu} a_s g_s \right\} )^2 \right) \left. \right|_{\rho \to 0} = - \frac{\log_2(\nu)}{\nu} \left( (1 + a_s \rho g_s)^{-\nu} \mathbb{E}_{g_s} \left\{ \nu (\nu + 1)(1 + a_s \rho g_s)^{-\nu} a_s g_s \right\} \right) = \log_2(\nu) a_s^2 \mathbb{E}_{g_s} \left\{ g_s^2 \right\} - \nu^2 a_s^2 \left( \mathbb{E}_{g_s} \left\{ g_s^2 \right\} \right)^2 = \log_2(\nu) a_s^2 \left( \nu \mathbb{E}_{g_s} \left\{ g_s \right\}^2 - (\nu + 1) \mathbb{E}_{g_s} \left\{ g_s^2 \right\} \right).
Similarly, for the weak user, it follows from (7) that

\[
\hat{R}_{w,NOMA} \triangleq \frac{dR_{w,NOMA}}{d\rho} \bigg|_{\rho \to 0} = \frac{d}{d\rho} \left( -\frac{1}{\nu} \log_2 \left[ E_{g_{\min}} \left\{ \left( \frac{1 + \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu} \right\} \right] \right)_{\rho \to 0}
\]

\[
= - \frac{\log_2(e)}{\nu} \left[ E_{g_{\min}} \left\{ \left( \frac{1 + \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu} \right\} \right]^{-1} \times E_{g_{\min}} \left\{ -\nu \left( \frac{1 + \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu-1} \left( \frac{1 + a_s \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu} \right\}_{\rho \to 0}
\]

\[
= - \frac{\log_2(e)}{\nu} E_{g_{\min}} \{-\nu g_{\min}(1 - a_s)\} = \log_2(e)a_w E \left\{ g_{\min} \right\},
\]

and

\[
\hat{R}_{w,NOMA} \triangleq \frac{d^2R_{w,NOMA}}{d\rho^2} \bigg|_{\rho \to 0} = -\frac{\log_2(e)}{\nu} \left[ E_{g_{\min}} \left\{ \left( \frac{1 + \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu} \right\} \right]^{-2} \times E_{g_{\min}} \left\{ -\nu \left( \frac{1 + \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu-1} \left( \frac{1 + a_s \rho g_{\min}}{1 + a_s \rho g_{\min}} \right)^{-\nu} \right\}_{\rho \to 0}
\]

\[
= - \frac{\log_2(e)}{\nu} \left[ E_{g_{\min}} \left\{ \nu(\nu + 1)(g_{\min} - a_s g_{\min})^2 + 2\nu a_s g_{\min}(g_{\min} - a_s g_{\min}) \right\} - (E_{g_{\min}} \{-\nu (g_{\min} - a_s g_{\min})\})^2 \right]
\]

\[
= \log_2(e) a_w \left[ \nu a_w \left( E \left\{ g_{\min}^2 \right\} \right)^2 - \left\{ (\nu + 1)a_w + 2a_s \right\} E \left\{ g_{\min}^2 \right\} \right].
\]

REFERENCES

[1] Dapeng 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, July 2003.

[2] L. Musavian and S. Aissa, “Effective capacity of delay-constrained cognitive radio in Nakagami fading channels,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1054–1062, 2010.

[3] M. You, H. Sun, J. Jiang, and J. Zhang, “Unified framework for the effective rate analysis of wireless communication systems over MISO fading channels,” *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1775–1785, 2017.

[4] ——, “Effective rate analysis in Weibull fading channels,” *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 340–343, 2016.

[5] J. Zhang, M. Matthaiou, Z. Tan, and H. Wang, “Effective rate analysis of MISO η-μ fading channels,” in *IEEE Int. Conf. Commun. (ICC)*, 2013, pp. 5840–5844.

[6] J. Zhang, L. Dai, Z. Wang, D. W. K. Ng, and W. H. Gerstacker, “Effective rate analysis of MISO systems over α-μ fading channels,” in *IEEE Global Commun. Conf. (GLOBECOM)*, 2015, pp. 1–6.

[7] J. Zhang, L. Dai, W. H. Gerstacker, and Z. Wang, “Effective capacity of communication systems over κ-μ shadowed fading channels,” *Electron. Lett.*, vol. 51, no. 19, pp. 1540–1542, 2015.

[8] H. Al-Hmood and H. S. Al-Raweshidy, “Unified approaches based effective capacity analysis over composite α-η-μ/Gamma fading channels,” *Electron. Lett.*, vol. 54, no. 13, pp. 852–853, 2018.

[9] S. Chen, J. Zhang, G. K. Karagiannidis, and B. Ai, “Effective rate of MISO systems over Fisher-Snedecor F fading channels,” *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2619–2622, December 2018.

[10] K. D. Kanellopoulou, K. P. Peppas, and P. T. Mathiopoulos, “Effective capacity analysis of equal gain diversity combiners over generalized fading channels,” in *IEEE Veh. Technol. Conf. (VTC Spring)*, June 2018, pp. 1–5.

[11] ——, “Effective capacity of Lp-norm diversity receivers over generalized fading channels under adaptive transmission schemes,” *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1240–1253, 2020.

[12] S. Schiessl, J. Gross, and H. Al-Zubaidy, “Delay analysis for wireless fading channels with finite blocklength channel coding,” in *ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2015, pp. 13–22.

[13] 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, 2018.
[14] 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, 2019.

[15] N. Pappas and M. Kountouris, “Delay violation probability and age of information interplay in the two-user multiple access channel,” in IEEE Int. Wksp Signal Proc. Adv. Wireless Commun. (SPAWC), 2019, pp. 1–5.

[16] M. Kountouris, N. Pappas, and A. Avranas, “QoS provisioning in large wireless networks,” in Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), May 2018, pp. 1–6.

[17] 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, April 2017.

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

[19] X. Chen, G. Liu, and Z. Ma, “Statistical QoS provisioning for half/full-duplex cooperative non-orthogonal multiple access,” in IEEE Veh. Technol. Conf. (VTC-Fall), 2017, pp. 1–5.

[20] 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, October 2018.

[21] C. Xiao, J. Zeng, B. Liu, X. Su, and J. Wang, “Cross-layer power control for uplink NOMA in IoT applications with statistical delay constraints,” in IEEE Global Commun. Conf. (GLOBECOM), 2018, pp. 1–7.

[22] C. Xiao, J. Zeng, W. Ni, R. P. Liu, X. Su, and J. Wang, “Delay guarantee and effective capacity of downlink NOMA fading channels,” IEEE J. Sel. Topics Signal Proc., vol. 13, no. 3, pp. 508–523, June 2019.

[23] S. Schiessl, M. Skoglund, and J. Gross, “NOMA in the uplink: Delay analysis with imperfect CSI and finite-length coding,” IEEE Trans. Wireless Commun., to appear.

[28] M. D. Yacoub, “The α-µ distribution: A physical fading model for the Stacy distribution,” IEEE Trans. Veh. Technol., vol. 56, no. 1, pp. 27–34, 2007.

[29] V. Kumar, B. Cardiff, and M. F. Flanagan, “Performance analysis of NOMA-based cooperative relaying in α-µ fading channels,” in 2019 IEEE Int. Conf. Commun. (ICC), May 2019, pp. 1–7.

[30] A. Jeffrey and D. Zwillinger, Table of Integrals, Series, and Products, 7th ed. Elsevier Science, 2007.

[31] H. Lei, I. S. Ansari, G. Pan, B. Alomair, and M. Alouini, “Secrecy capacity analysis over α–µ fading channels,” IEEE Commun. Lett., vol. 21, no. 6, pp. 1445–1448, June 2017.

[32] K. P. Peppas, “A new formula for the average bit error probability of dual-hop amplify-and-forward relaying systems over generalized shadowed fading channels,” IEEE Wireless Commun. Lett., vol. 1, no. 2, pp. 85–88, April 2012.

[33] V. Kumar, B. Cardiff, and M. F. Flanagan, “Performance analysis of NOMA with generalised selection combining receivers,” Electron. Lett., vol. 55, no. 25, pp. 1364–1367, December 2019.

[34] V. Adamchik and O. I. Marichev, “The algorithm for calculating integrals of hypergeometric type functions and its realization in REDUCE system,” in Proceeding of the International Symposium on Symbolic and Algebraic Computation. New York, NY, USA: ACM, 1990, pp. 212–224

[35] M. D. Springer, The Algebra of Random Variables. John Wiley & Sons, 1979.

[36] P. K. Mittal and K. C. Gupta, “An integral involving generalized function of two variables,” Proceedings of the Indian Academy of Sciences - Section A, vol. 75, no. 3, pp. 117–123, Mar 1972.