QoS Provision for Industrial IoT Networking: Multiantenna NOMA Based on Partial CSIT

Li Bing, *Member, IEEE*, Yating Gu, *Graduate Student Member, IEEE*, Lanke Hu, Tor Aulin, *Life Fellow, IEEE*, Yue Yin, and Jue Wang

***Abstract*—The acquisition of small-scale fading is essential but challenging in industrial Internet of Things (IIoT). This article concerns with the quality of service (QoS) provision for IIoT uplink networking, in terms of reliability, latency, and connectivity without resorting to small-scale fading. To this end, a scheme based on multiantenna nonorthogonal multiple access (mNOMA) technique is proposed. Different from existing setups, the proposed system is designed to work in overloaded regime where the number of devices K significantly outnumbers N_r, the antennas equipped by the access point, i.e., $K \gg N_r$, such that concurrent massive connectivity is enabled using off-the-shelf equipment. The design starts with the development of asymptotic performance analysis in terms of signal to interference-plus-noise ratio, where a closed-form expression is obtained taking massive connectivity, finite blocklength, and code rate into consideration jointly. The analysis implies the possibility of QoS provision without estimating small-scale fading at transmitter side. Then the insight is leveraged to develop mNOMA with special focus on minimum shift keying type waveform, which is widely adopted in low power wide area IIoT standards including IEEE 802.15 series. Power allocation and code rate adaption are employed to offer QoS guaranteed performance as the blocklength is only 512 bits, whereas dozens of devices simultaneously transmit at block error rate down to 10^{-5} or even lower, as required by industrial 5.0 and beyond.**

***Index Terms*—Continuous phase modulation, industrial Internet of Things (IIoT), multiantenna, noncoherent transmission, nonorthogonal multiple access (NOMA), quality of service (QoS).**

Manuscript received 8 May 2023; revised 21 September 2023 and 11 December 2023; accepted 5 February 2024. Date of publication 12 March 2024; date of current version 5 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61601346, Grant 61972318, and Grant 62377039, in part by Shaanxi Natural Science Basic Research Plan under Grant 2018JQ6044, and in part by Shaanxi Provincial Science and Technology Project under Grant 2023-GHZD-47. Paper no. TII-23-1631. *(Corresponding author: Li Bing.)*

Li Bing, Yating Gu, Lanke Hu, and Yue Yin are with Northwestern Polytechnical University, Xi'an 710072, China (e-mail: libingprc@gmail.com).

Tor Aulin, retired, was with the Chalmers University of Technology, SE-41296 Gothenburg, Sweden.

Jue Wang is with the 20th Research Institute of China Electronic Technology Group Corporation, Xi'an 710068, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2024.3366222.

Digital Object Identifier 10.1109/TII.2024.3366222

I. Introduction

THE next generation industrial network features automated manufacturing, for which massive devices (meters, sensors, and trackers) are deployed to render real-time sensing, accurate positioning, reliable control, etc. [1], [2]. Such scenario is termed as massive machine type communication (mMTC), which 5G and beyond have attempted to address.

One of the challenges of mMTC is to enable massive access. Existing industrial Internet of Things (IIoT) relies on orthogonal multiple access (OMA) mechanism, such as time division multiple access that permits only one access each resource block (RB) [1]. Hence, enabling mMTC in OMA is equivalent to deploying massive RB, which is barely acceptable in industry 5.0 and beyond. Fortunately, nonorthogonal multiple access (NOMA) enables simultaneous access of multiple devices by sharing RB among them. Obviously, the key is to mitigate severe interference. Among many techniques, forward error corrections (FECs) and power allocation are prominent methods [3], [4], which have all proved successful in supporting mMTC. However, IIoT networking requires not only simultaneous multiple access, but also improved spectral efficiency (SE), high reliability, and low delay especially as highlighted in [1] and [2].

A. Previous Works

The success of NOMA has triggered extensive interest in IIoT recently concerning diverse quality of service (QoS) provision, as addressed in e.g., [5], [6], [7], [8], and [9]. NOMA is firstly employed to improve SE by accommodating multiple user over a single RB. In [5], coded tandem spreading multiple access for IIoT was proposed to support multiple simultaneous access. In [6], interleave-grid based NOMA method and its testbed was put forward. The scheme proved that of low complexity yet energy-efficient.

Ultrareliable low latency communications (URLLC) in IIoT is made possible with recent development of NOMA in finite blocklength regime. In [7], URLLC provision in IIoT using multicarrier NOMA is addressed, where performance of NOMA using finite blocklength regime theory is presented. In [8], autoconfigurable NOMA is proposed to enable massive and ultrareliable random access in IIoT. However, as user load increases, design of NOMA becomes challenging. Hence, numerical optimization method is proposed in [9], where the resource allocation problem is converted to convex optimization and efficiently solved.

Recently, multiantenna enabled QoS provision in IIoT has been extensively discussed. The benefits brought by multi-antenna NOMA (mNOMA) include significant diversity gain and massive connectivity mainly, as discussed in, e.g, [10], [11], and [12]. In [10], nonlinear multiantenna transmission for IIoT was proposed. Power consumption is significantly reduced due to diversity gain. In [11], feasibility of massive multiantenna technique in IoT is addressed. In this setup, the access point (AP) equips with N_r antennas and serves $K \ll N_r$ devices, such that devices are allocated to mutually orthogonal beams due to channel hardening property. This property is further explored in [12], and massive antenna technique is combined with NOMA to boost connectivity further. Though massive antenna has proved viable in outdoor IIoT networking [13], the feasibility is perhaps questionable in indoor scenario due to the very large size of antenna array when operating in industrial, scientific, and medical band. Hence, overloaded mNOMA with $N_r \ll K$ is desirable for possible indoor IIoT networking.

B. Motivations and Contributions

As pointed out in [13], multiantenna technique is perhaps the most promising breakthrough in IIoT networking. However, it is also revealed that the success presumes perfect acquisition of channel state information (CSI), which is one of the most challenging task in IIoT. In general, there are two methods available to fulfill this purpose. The most convenient method is to use pilots to estimate CSI in uplink transmission at AP and, then, use dedicated overhead to broadcast CSI to devices to avoid computationally-demanding CSI estimation in IIoT devices. Obviously, the pilots/overhead inevitably degrade SE. To overcome this drawback, nonpilot aided method was alternatively suggested in, e.g., [14], [15], and [16], where deep-learning, Gaussian mixture model, and joint data-channel estimation were used, respectively. Disregarding the near optimal estimation, the method is still too computational-demanding to be implemented in IIoT, where most devices are sensors. Hence, the acquisition of CSI at transmitter (CSIT) is solvable but not feasible.

Even perfect CSIT is accessible, concern still arises perceiving that existing mNOMA designs barely address the binary nonlinear modulations, i.e., minimum shift keying (MSK) and Gaussian MSK (GMSK). MSK family is favored in IIoT standards, such as IEEE 802.15.1 (a.k.a Bluetooth) and IEEE 802.15.4 (a.k.a Zigbee) to take the advantage of constant envelope property in mitigating nonlinear distortion due to hardware impairment [1]. Though MSK/GMSK-based NOMA is demonstrated to outperform linear modulation-based NOMA in terms of SEs [17], [18], [19], the design and performance analysis of multiantenna configured MSK/GMSK-based NOMA have gained little attention.

This article attempts to design MSK/GMSK-based mNOMA to provide QoS in terms of improved SE, strengthened reliability, and reduced latency for IIoT, given only partial CSIT, i.e., only the path loss of CSIT is needed. The main contributions of the proposed mNOMA are summarized below.

1) *Asymptotic performance analysis in finite blocklength regime:* The asymptotic performance of overloaded system ($K \gg N_r$) over fast Rayleigh fading channels is analyzed in terms of signal to interference-plus-noise ratio (SINR), taking finite blocklength into consideration. The analysis implies that mNOMA is deterministic, as $\kappa = K/N_r \gg 1$. A favorable property similar to channel hardening in massive antenna setups, where $\kappa \ll 1$. The result is then used to derive sum SE and delay violation probability in closed-form expressions given partial CSIT. Accurate expressions and upper bounds are obtained, and are shown to coincide with experimental results well.
2) *Joint power allocation and rate adaption method:* The analysis is then leveraged to design practical mNOMA based on MSK/GMSK, where FECs of different code rates R_{FEC} are considered as a means for interference mitigation. Power allocation scheme adapts to R_{FEC} is derived to aid the interference mitigation and to offer provable guaranteed performance.
3) *Low complexity implementation of practical mNOMA:* Two IIoT FECs, namely, convolutional code (CC) and repetition code (REP) [1, Ch. 13], are used to validate the proposed method. The complexity at the receiver side is reduced by combing linear minimum mean square error (LMMSE) and successive interference cancellation (SIC). Simulated results show that the gap of the proposed mNOMA to the theoretical limit is very small, even the latter can access perfect CSIT.

In summary, this article offers both analysis and construction of mNOMA scheme based on MSK/GMSK, which can be considered a candidate for IIoT with massive connectivity and URLLC enabled. In contrast to existing power domain multiantenna NOMA schemes, e.g., [10], [11], and [12], one perceived merit is that mNOMA does not acquire small-scale fading. This subtle difference would reduce the computational burden significantly due to CSIT estimation, user-clustering, etc., where iterative optimization is required frequently. The proposed mNOMA also offers an alternative methodology for constructing coded modulation based mNOMA, where density evolution method is favored to offer accurate performance prediction in infinite-long blocklength regime [20], [21]. Irrespective of the success, such method can hardly tackle short blocklength of a few hundreds bits, hence, is unlikely an ideal candidate for IIoT.

The rest of this article is organized as follows. Section II presents the system model. Section III presents the asymptotic performance analysis, joint power allocation, and rate adaption. Section IV presents the simulated results. Finally, Section V concludes this article.

II. System Model and Problem Formulation

A. System Model

This article considers the uplink transmission of mNOMA system, where an N_r-antenna AP serves K single-antenna devices. Since all devices are driven by the same standard,

the transceivers are identical. It is further assumed that the receiver, i.e., AP has access to perfect CSI, using advanced methods presented in, e.g., [14], [15], and [16]. Take device k for example, the workflow is sketched as follows. The information sequence $\boldsymbol{d}_k$ is first encoded by the FEC (of rate R_{FEC}) to generate codeword $\boldsymbol{c}'_k$, which is then fed to user-specific interleaver Π_k to produce a permutated codeword $\boldsymbol{c}_k$. The codeword $\boldsymbol{c}_k$ is the input of MSK/GMSK modulator and the resultant signal is expressed as

$$s(t,\boldsymbol{c}_k)=\sqrt{p_k}\cdot\exp[j\phi(t,\boldsymbol{c}_k)+j\theta_k], t\in[(n-1)T,nT] \tag{1}$$

where $j=\sqrt{-1}$, T is the symbol period, p_k is the transmit power allocated to device-k, θ_k is a user-specific phase offset [8], and the information-bearing phase $\phi(t,\boldsymbol{c}_k)$ reads [22]

$$\phi(t,\boldsymbol{c}_k)=2\pi h\sum_{l=1}^{n}c_k(l)q(t-lT) \tag{2}$$

where $h=1/2$, $c_k(l)$ is the lth code bits of $\boldsymbol{c}_k$, and $q(t)$ is the phase response function. The choice of $q(t)$ leads to different signaling and normalized bandwidth. In the considered scheme, the normalized bandwidth is $B=1$ (or $B=1.2$) as GMSK (or MSK) applies, where the achievable SE is R_{FEC} (or $R_{\text{FEC}}/1.2$).

At the receiver side, i.e., AP, the received signal reads

$$\boldsymbol{r}=\boldsymbol{G}\boldsymbol{s}+\boldsymbol{v} \tag{3}$$

where $\boldsymbol{G}$, $\boldsymbol{s}$, and $\boldsymbol{v}$ are channel matrix between devices and AP, signal vector, and additive white Gaussian noise (AWGN), respectively. In detail, $\boldsymbol{G}=[\sqrt{g_k}\boldsymbol{h}_k]_{N_r\times K}$ [7], [11], where g_k and $\boldsymbol{h}_k$ account for the path-loss and small-scale fading coefficients between device-k and AP, respectively. The former g_k is assumed a constant and known in prior for both ends, and the elements of the latter $\boldsymbol{h}_k$ are complex-Gaussian distributed, i.e., $[\boldsymbol{h}_k]_{N_r\times 1}\sim\mathcal{CN}(0,\boldsymbol{I})$, which is known at the receiver side but not transmitter side. The acquisition of accurate CSI at AP for overloaded mNOMA can employ the methods presented in [23] and [24]; $\boldsymbol{s}$ collects all signals from devices, i.e., $\boldsymbol{s}=[s_1,\ldots,s_K]^{\intercal}$, where s_k is the discrete representation of $s(t,\boldsymbol{c}_k)$ using, e.g., principal component analysis [19]; AWGN $\boldsymbol{v}\sim\mathcal{CN}(0,\boldsymbol{I})$.

The receiver is mounted with LMMSE-SIC to suppress the severe interference. The LMMSE filter/beamformer $\boldsymbol{W}=[w_{ij}]_{K\times N_r}$ is written as

$$\boldsymbol{W}=\text{diag}[g_1\text{Var}(s_1),\ldots,g_k\text{Var}(s_K)]\cdot\boldsymbol{H}^{\text{H}}\boldsymbol{R}^{-1} \tag{4}$$

where $\boldsymbol{H}=[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_K]$, H denotes the Hermitian transpose, and diag stands for diagonalization operations, respectively.

$$\boldsymbol{R}=\sum_{k=1}^{K}g_k\text{Var}(s_k)\cdot\boldsymbol{h}_k\boldsymbol{h}_k^{\text{H}}+\boldsymbol{I}. \tag{5}$$

Hence, the LMMSE-processed signal of user-k is [8]

$$x_k=\boldsymbol{w}_k(\boldsymbol{r}-\boldsymbol{H}\boldsymbol{\mu}_s)+\boldsymbol{\mu}_s \tag{6}$$

where $\boldsymbol{\mu}_s=[\text{E}(s_1),\ldots,\text{E}(s_K)]^{\intercal}$ and $\boldsymbol{w}_k$ is the kth row of $\boldsymbol{W}$, expressed as

$$\boldsymbol{w}_k=g_k\text{Var}(s_k)\boldsymbol{h}_k\left(\sum_{k=1}^{K}g_k\text{Var}(s_k)\cdot\boldsymbol{h}_k\boldsymbol{h}_k^{\text{H}}+\boldsymbol{I}\right)^{-1}. \tag{7}$$

To implement SIC at the receiver side, the detection order is first sorted according to $\rho_k:=g_kp_k\forall k\in[1,K]$ in descending order, i.e., $\rho_1\geq\rho_2\geq\cdots\geq\rho_K$ without loss of generality. Assuming perfect *a priori* information regarding $s_m\forall m\in[1,k-1]$ while no *a priori* information regarding $s_n\forall n\in[k,K]$, there exists [4, Sec.II.B]

$$\begin{cases}[\text{E}((s_m)),\text{Var}(s_m)]=[s_m,0] & \forall m\in[1,k-1]\\ [\text{E}((s_n)),\text{Var}(s_n)]=[0,p_n] & \forall n\in[k,K]\end{cases} \tag{8}$$

and after some manipulations the SINR of x_k reads

$$\begin{aligned}\gamma_k&=\frac{\rho_k\boldsymbol{h}_k^{\text{H}}\left(\sum_{m\geq k}\rho_m\boldsymbol{h}_m\boldsymbol{h}_m^{\text{H}}+\boldsymbol{I}\right)^{-1}\boldsymbol{h}_k}{1-\rho_k\boldsymbol{h}_k^{\text{H}}\left(\sum_{m\geq k}\rho_m\boldsymbol{h}_m\boldsymbol{h}_m^{\text{H}}+\boldsymbol{I}\right)^{-1}\boldsymbol{h}_k}\\&=\rho_k\boldsymbol{h}_k^{\text{H}}\left(\sum_{m>k}\rho_m\boldsymbol{h}_m\boldsymbol{h}_m^{\text{H}}+\boldsymbol{I}\right)^{-1}\boldsymbol{h}_k\end{aligned} \tag{9}$$

where the last line is obtained using matrix inversion lemma $\boldsymbol{x}^{\text{H}}(\boldsymbol{A}+\alpha\boldsymbol{x}\boldsymbol{x}^{\text{H}})^{-1}\boldsymbol{x}=\boldsymbol{x}^{\text{H}}\boldsymbol{A}^{-1}\boldsymbol{x}/1+\alpha\boldsymbol{x}^{\text{H}}\boldsymbol{A}^{-1}\boldsymbol{x}$. The above procedure repeats for all devices, and the transmitted information $\boldsymbol{d}_k\forall k\in[1,K]$ is retrieved successfully, as long as γ_k surpasses a predetermined threshold γ. However finding γ is challenging in general, since γ_k is a random variable due to the independently varying $\boldsymbol{h}_k$ from symbol to symbol. Though recent results demonstrated that γ_k is Gamma-distributed [25], the parameterized-distribution has not been available for overloaded mNOMA so far.

B. Figures of Merit

The QoS key performance indicators (KPIs) considered are connectivity K, target block error rate (BLER) ϵ, and latency w, which are jointly depicted by ergodic SE $\bar{R}_k$ and delay violation probability $p_d(w)$ explicitly. Without loss of generality, each slot transmits one block, whose length is equivalent to N coded bits.

1) Ergodic SE: Given SINR γ_k and taking the parameters ϵ and N jointly into consideration, the attainable *instantaneous* SE per device of mNOMA over time-varying fading channels reads [26, Eq.(4)], [11, Eq.(18)]

$$R_k\approx\mathcal{R}(N,\gamma_k,\epsilon)=C-\sqrt{\frac{V}{N}}Q^{-1}(\epsilon) \tag{10}$$

where $C=1/2\log_2(1+\gamma_k)$, $V=\log_2^2 e/2[1-1/(1+\gamma_k)^2]$, and $Q(x)=\int_x^{\infty}1/\sqrt{2\pi}e^{-t^2/2}dt$ is the Gaussian Q-function. And the ergodic SE per device is expressed as

$$\bar{R}_k\approx\text{E}[\mathcal{R}(N,\gamma_k,\epsilon)]. \tag{11}$$

2) Delay Violation Probability $p_d(w)$: In brief, given the signal to noise ratio (SNR) domain data arrival process $\mathcal{A}(u)$, service process $\mathcal{S}(u)$, and departure process $\mathcal{D}(u)$, the accumulative arrival, service, and departure processes in a concerned time

window $[s,t]$ are, respectively, calculated as [27]

$$\mathcal{A}(s,t)=\prod_{u=s}^{t-1}\mathcal{A}(u),\mathcal{S}(s,t)=\prod_{u=s}^{t-1}\mathcal{S}(u),\mathcal{D}(s,t)=\prod_{u=s}^{t-1}\mathcal{D}(u). \tag{12}$$

Two companion operations $(\min,\times)$-convolution operation $\otimes$ and $(\min,\times)$-deconvolution operation $\oslash$ are, respectively, defined as

$$\mathcal{X}\otimes\mathcal{Y}(s,t):=\inf_{s\le u\le t}\mathcal{X}(s,u)\cdot\mathcal{Y}(u,t) \tag{13}$$

and

$$\mathcal{X}\oslash\mathcal{Y}(s,t):=\sup_{0\le u\le s}\left\{\frac{\mathcal{X}(u,t)}{\mathcal{Y}(u,s)}\right\}. \tag{14}$$

As a result, the departure process can be reexpressed as $\mathcal{D}(0,t)\ge\mathcal{A}\otimes\mathcal{S}(0,t)$, and the latency $W(t)$ after some simplifications is written as

$$\begin{aligned}W(t)&=\inf\left\{u>0:\frac{\mathcal{A}(0,t)}{\mathcal{D}(0,t+u)}\le 1\right\}\\&\le\inf\left\{u>0:\mathcal{A}\oslash\mathcal{S}(t+u,t)\le 1\right\}.\end{aligned} \tag{15}$$

Therefore, the probability that the delay exceeds a tolerable value w is bounded as

$$\mathrm{P}\left\{W(t)>w\right\}\le\mathrm{P}\left\{\mathcal{A}\oslash\mathcal{S}(t+u,t)\le 1\right\}. \tag{16}$$

Resorting to Mellin transform $\mathcal{M}_{\mathcal{X}}(\theta)=\mathrm{E}[\mathcal{X}^{\theta-1}]$ with $\theta>0$ and Chernoff bounding technique, there exists

$$\mathrm{P}\{W(t)>w\}\le\sum_{v=0}^{t}\mathcal{M}_{\mathcal{A}}(1+\theta)^{t-v}\mathcal{M}_{\mathcal{S}}(1-\theta)^{t+w-v}. \tag{17}$$

As long as $\mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}_{\mathcal{S}}(1-\theta)<1$ (Gartner–Ellis condition), an upper bound of the delay violation probability $p_d(w):=\sup_{t\ge0}\{\mathrm{P}(W(t)>w)\}$ can be established as

$$p_d(w)=\frac{\mathcal{M}_{\mathcal{S}}(1-\theta)^w}{1-\mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}_{\mathcal{S}}(1-\theta)} \tag{18}$$

which is obtained from (17) as $t\to\infty$. Therefore, the calculation boils down to evaluating $\mathcal{M}_{\mathcal{A}}(\theta)$ and $\mathcal{M}_{\mathcal{S}}(\theta)$ eventually.

For mMTC, the arrival process is modeled as a Poisson process of parameter η, to indicate the massive and sporadic connectivity. As a result [11], [27]

$$\mathcal{M}_{\mathcal{A}}(\theta)=\mathrm{E}\left[\mathcal{A}^{\theta-1}\right]=e^{\eta(e^{\theta-1}-1)} \tag{19}$$

where η is the arrival rate in bits to depict Poisson arrival process, and

$$\mathcal{M}_{\mathcal{S}}(\theta)=\mathrm{E}\left[\mathcal{S}^{\theta-1}\right]=\mathrm{E}\left[e^{(\theta-1)N\mathcal{R}(N,\gamma_k,\epsilon)}\right] \tag{20}$$

While (19) is easily obtained, (20) admits no closed-form expression in general.

C. Problem Formulation

The major concern of this article is to provide KPI-guaranteed massive connectivity for IIoT, while maximizing the minimum achievable sum rate R_{sum} of the entire network, i.e.,

$$\begin{aligned}\max_{\boldsymbol{p}}\min\quad & R_{\text{sum}}=\sum_{k=1}^{K}\bar{R}_k\\ \text{subject to}\quad & \sum_{k=1}^{K}p_k=p_{\mathsf{T}}\end{aligned} \tag{21}$$

where $\boldsymbol{p}=[p_1,\ldots,p_K]$ is a power allocation scheme given total available power p_{T}. The challenge arises due to the intractable quantify $\bar{R}_k$ and finding the feasible power allocation $\boldsymbol{p}=[p_1,\ldots,p_K]$.

The key for solving this problem is γ_k, which bridges $\boldsymbol{p}$ and KPIs, as already mentioned in (11) and (20) for example. In the sequel, the statistical behavior of γ_k is presented first using large dimensional analysis technique. Then, the derived result enlightens the construction of feasible solution $\boldsymbol{p}$.

III. Design and Performance Analysis: A SINR-Centric Approach

A. Statistical Behavior of γ_k Over Fast Fading Channels

The first step toward solving the problem is to find the distribution of γ_k, which is detailed in Lemma 1. In the sequel, the *effective* user load

$$K_e=K-(k-1) \tag{22}$$

is defined to indicate the residual K_e data streams to be processed at the receiver side if taking perfect SIC into consideration. Accordingly, the overloading factor $\kappa:=K_e/N_r$, which equals K/N_r initially and reduces progressively as LMMSE-SIC applies device by device as pointed out in (8).

Lemma 1: The SINR γ_k is Gamma-distributed, i.e., $\gamma_k\sim\text{Gamma}(\alpha,\beta)$, where the probability density function $f(\gamma_k)$ is expressed as

$$f(\gamma_k)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}{\gamma_k}^{\alpha-1}e^{-\beta\gamma_k} \tag{23}$$

where $\alpha=N_r$, $\beta=(K-N_r)\delta$, and $\Gamma(z)=\int_0^{\infty}t^{z-1}e^{-t}\mathrm{d}t$.

Proof: The proof starts with finding the distribution of γ_k. According to (9)

$$\gamma_k=\rho_k\boldsymbol{h}_k^{\mathrm{H}}\left(\sum_{m>k}\rho_m\boldsymbol{h}_m\boldsymbol{h}_m^{\mathrm{H}}+\boldsymbol{I}\right)^{-1}\boldsymbol{h}_k=\rho_k\boldsymbol{h}_k^{\mathrm{H}}(\boldsymbol{R}_k+\boldsymbol{I})^{-1}\boldsymbol{h}_k \tag{24}$$

where $\boldsymbol{R}_k:=\sum_{m>k}\rho_m\boldsymbol{h}_m\boldsymbol{h}_m^{\mathrm{H}}$ and admits eigenvalue decomposition $\boldsymbol{R}_k=\boldsymbol{Q}_k\boldsymbol{\Lambda}_k\boldsymbol{Q}_k^{\mathrm{H}}$, where $\boldsymbol{Q}_k$ collects the eigenvectors of $\boldsymbol{R}_k$, and $\boldsymbol{\Lambda}_k$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\boldsymbol{R}_k$. By letting $\hat{\boldsymbol{h}}_k=\boldsymbol{Q}_k^{\mathrm{H}}\boldsymbol{h}_k$, SINR γ_k can be rewritten as [25, Eq. (20)], [28, 3.20.14]

$$\gamma_k=\rho_k\hat{\boldsymbol{h}}_k^{\mathrm{H}}(\boldsymbol{\Lambda}_k+\boldsymbol{I})^{-1}\hat{\boldsymbol{h}}_k=\sum_{j=1}^{N_r}\rho_k(\lambda_{kj}+1)^{-1}\|\hat{h}_{kj}\|_{\mathrm{F}}^2 \tag{25}$$

where λ_{kj} is the jth diagonal element of Λ_k and $\hat{h}_{kj}$ is the jth element of $\hat{\boldsymbol{h}}_k$. It is proved in [25] that γ_k, expressed

as the weighted summation of $\|\hat{h}_{kj}\|_{\mathrm{F}}^2$, is Gamma-distributed, i.e., $\gamma_k \sim \Gamma(\alpha, \beta)$. However the shape parameter α and rate parameter β are not readily available. Alternatively, there exists $\alpha = \mu_\gamma^2/\sigma_\gamma^2$ and $\beta = \mu_\gamma/\sigma_\gamma^2$, where μ_γ and σ_γ^2 are the mean and variance of γ_k, respectively.

In this regard, the mean $\mu_\gamma := \mathrm{E}(\gamma_k)$ and variance $\sigma_\gamma^2 := \mathrm{Var}(\gamma_k)$ of γ_k are calcuated assuming $\kappa = K_e/N_r \geq 1$. First the mean μ_γ can be shown to converge to a deterministic value almost surely (a.s.)

$$\mu_\gamma = \mathrm{E}_{\boldsymbol{H}}\left[N_r\rho_k \boldsymbol{h}'^{\mathrm{H}}_k \left(N_r \sum_{m>k} \rho_m \boldsymbol{h}'_m \boldsymbol{h}'^{\mathrm{H}}_m + \boldsymbol{I}\right)^{-1} \boldsymbol{h}'_k\right] \tag{26}$$

where $\boldsymbol{h}'_k = \boldsymbol{h}_k/N_r$, $\boldsymbol{h}'_m = \boldsymbol{h}_m/N_r$, and $(K_e-1)\rho_a = \sum_{m>k}\rho_m$. Using [29, Eq. (46)], there exists

$$\mu_\gamma \approx \quad \mathrm{E}_{\boldsymbol{H}}\left[N_r\rho_k \boldsymbol{h}'^{\mathrm{H}}_k \left(N_r\rho_a \sum_{m>k} \boldsymbol{h}'_m \boldsymbol{h}'^{\mathrm{H}}_m + \boldsymbol{I}\right)^{-1} \boldsymbol{h}'_k\right] \tag{27}$$

the right-hand side (RHS) converges a.s. to the following deterministic value:

$$\mu_\gamma \xrightarrow{\text{a.s.}} \quad \frac{\rho_k}{\rho_a} \cdot \underbrace{N_r\rho_a\left(1 - \frac{\mathcal{F}(N_r\rho_a, \kappa)}{4N_r\rho_a}\right)}_{:=\mu} = \frac{\rho_k}{\rho_a} \cdot \frac{1}{\kappa - 1} \tag{28}$$

where $\mathcal{F}(x, z)$ is the η-transform of the empirical distribution of eigenvalues of $\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}$ and is expressed as $\mathcal{F}(x,y) = (\sqrt{x(1+\sqrt{y})^2+1} - \sqrt{x(1-\sqrt{y})^2+1})^2$ and the final result holds as long as $\kappa > 1$ [30].

A useful conclusion is perhaps that small-scale fading $\boldsymbol{h}_k \forall k$ is simply averaged out and only large-scale fading g_k and power p_k are concerned as revealed in (28). This observation renders the possibility of developing mNOMA transceiver using only partial CSIT g_k, i.e., largely dependent on the placement of devices, which can be obtained a priori to transmission. The details are presented in Section III-C.

Further, the a.s. convergence behavior of γ_k also implies that σ_γ^2 approaches 0, which is now derived. To verify, σ_γ^2 is expressed as

$$\begin{aligned}\sigma_\gamma^2 &= \mathrm{Var}_{\boldsymbol{H}}\left[N_r\rho_k \boldsymbol{h}'^{\mathrm{H}}_k \left(\sum_{m>k} N_r\rho_m \boldsymbol{h}'_m \boldsymbol{h}'^{\mathrm{H}}_m + \boldsymbol{I}\right)^{-1} \boldsymbol{h}'_k\right] \\ &= \left(\frac{\rho_k}{\rho_a}\right)^2 \cdot \mathcal{G}(N_r\rho_a, \kappa)\end{aligned} \tag{29}$$

where $\mathcal{G}(N_r p_a, \kappa)$ reads [31, Sec.VI]

$$\mathcal{G}(N_r\rho_a, \kappa) = \frac{1}{N_r}\left[\frac{2\mu(1+\mu)^2 N_r\rho_a}{(1+\mu)^2 + N_r\rho_a \cdot \kappa} - \mu^2\right] \tag{30}$$

and, hence

$$\sigma_\gamma^2 = \left(\frac{\rho_k}{\sqrt{N_r}\rho_a}\right)^2 \left[\frac{2\mu(1+\mu)^2 N_r\rho_a}{(1+\mu)^2 + N_r\rho_a\kappa} - \mu^2\right]. \tag{31}$$

An interesting observation is that the variance can be effectively reduced letting $N_r \to \infty$, such that the SINR is increasingly deterministic, just as extensively studied in the context of massive antenna setups, see, e.g., [11] and the references therein.

More interestingly, due to the presence of κ in the denominator of (31), the same effect is implied by increasing $\kappa = K_e/N_r$ while keep N_r small. To see this, σ_γ^2 is further calculated as

$$\begin{aligned}\sigma_\gamma^2 &= \left(\frac{\rho_k}{\sqrt{N_r}\rho_a}\right)^2 \left[\frac{\mu}{\kappa} \cdot \frac{2(1+\mu)^2 N_r\rho_a\kappa}{(1+\mu)^2 + N_r\rho_a\kappa} - \mu^2\right] \\ &\approx \left(\frac{\rho_k}{\sqrt{N_r}\rho_a}\right)^2 \left[\mu^2 \frac{2N_r\rho_a\kappa}{1 + N_r\rho_a\kappa} - \mu^2\right] \\ &= \left(\frac{\rho_k}{\sqrt{N_r}\rho_a}\right)^2 \mu^2 \to 0\end{aligned} \tag{32}$$

where $\mu = 1/\kappa - 1 \approx 0$ and $N_r\rho_a\kappa > 1$ are used. As expected. the result $\sigma_\gamma^2 \to 0$ suggests that SINR becomes deterministic and converges to μ_γ irrespective of $\boldsymbol{H}$ with increasing κ, a favorable property in massive antenna system due to channel hardening given $\kappa \ll 1$. Therefore, the target SINR γ_k equals μ_γ asymptotically, which is explored in the sequel.

Then the parameters α and β of (23) are now obtained as

$$\begin{cases} \alpha = \dfrac{\mu_\gamma^2}{\sigma_\gamma^2} = \dfrac{\left(\frac{\rho_k}{\rho_a}\frac{1}{\kappa-1}\right)^2}{\left(\frac{\rho_k}{\sqrt{N_r}\rho_a}\right)^2\mu^2} = N_r & (33) \\ \beta = \dfrac{\mu_\gamma}{\sigma_\gamma^2} = \dfrac{N_r}{\mu_\gamma}\dfrac{\rho_a}{\rho_k} = (K_e - N_r)\delta & (34) \end{cases}$$

where $\delta := \rho_a/\rho_k = 1/\gamma_k(\kappa-1)$, which is calculated once the target SINR γ_k and system configuration κ are given without specifying ρ_a and ρ_k yet.

It is worth noticing that α and β eventually have nothing to do with the small-scale fading $\boldsymbol{h}_k$, which implies the possibility of designing mNOMA given path-loss g_k alone. ■

B. Accurate and Upper-Bounded $\bar{R}_k$ and $p_d(w)$

Given the statistics of γ_k, the ergodic SE $\bar{R}_k$ and the delay violation probability $p_d(w)$ are presented in terms of accurate results and upper bounds. To avoid numerical instability, i.e., $R_k < 0$ if $\gamma_k < \gamma_{\mathrm{th}}$, the actual SE per device is defined as [11]

$$R'_k = \max(R_k, 0). \tag{35}$$

1) Ergodic SE $\bar{R}_k$: The ergodic SE is calculated as $\bar{R}_k = \mathrm{E}(R'_k)$, i.e.,

$$\begin{aligned}\bar{R}_k = &\underbrace{-\int_0^{\gamma_{\mathrm{th}}} R_k f(\gamma_k)\mathrm{d}\gamma_k}_{\bar{R}_3} + \underbrace{\int_0^\infty \frac{1}{2}\log_2(1+\gamma_k) f(\gamma_k)\mathrm{d}\gamma_k}_{\bar{R}_1} \\ &\underbrace{-\int_0^\infty Q^{-1}(\epsilon)\sqrt{\frac{\log_2^2 e}{2\,N}\left[1 - \frac{1}{(1+\gamma_k)^2}\right]} f(\gamma_k)\mathrm{d}\gamma_k}_{\bar{R}_2}\end{aligned} \tag{36}$$

where $\bar{R}_3$ can be obtained numerically. According to Lemma 1, $\bar{R}_1$ is expressed as

$$\bar{R}_1 = \frac{\beta^\alpha \log_2 e}{2\Gamma(\alpha)} \int_0^\infty \ln(1+\gamma_k)\gamma_k{}^{\alpha-1} e^{-\beta\gamma_k} \mathrm{d}\gamma_k \tag{37}$$

where the integral can be solved by applying the formula [32]

$$\int_0^\infty \ln(1+x)x^{n-1}e^{-bx}\mathrm{d}x = \Gamma(n)e^b \sum_{k=1}^{n} \frac{\Gamma(k-n,b)}{b^k} \tag{38}$$

which leads to the following result:

$$\bar{R}_1 = \frac{\beta^\alpha e^\beta \log_2 e}{2} \sum_{k=1}^{\alpha} \frac{\Gamma(k-\alpha,\beta)}{\beta^k} \tag{39}$$

where $\Gamma(a,b) = \int_b^\infty t^{a-1}e^{-t}\mathrm{d}t$ stands for the upper incomplete Gamma function.

Regarding $\bar{R}_2$, there exists a tight approximation [11], [12]

$$\bar{R}_2 \approx Q^{-1}(\epsilon)\log_2 e/\sqrt{2\,N} \tag{40}$$

and an accurate expression is presented in Lemma 2 below.

Lemma 2: Given the mNOMA configuration (K, N_r, δ), the ergodic SE is expressed as

$$\bar{R}_k = \frac{\beta^\alpha e^\beta \log_2 e}{2} \sum_{k=1}^{\alpha} \frac{\Gamma(k-\alpha,\beta)}{\beta^k} - \frac{Q^{-1}(\epsilon)\log_2 e}{\sqrt{2\,N}} + \bar{R}_3 \tag{41}$$

Proof: The proof is a summary of the above discussion. ■

2) Delay Violation Probability $p_d(w)$: The key for assessing $p_d(w)$ is obtaining $\mathcal{M}_{\mathcal{S}}(1-\theta)$. According to the definition (20) and (35), $\mathcal{M}_{\mathcal{S}}(1-\theta)$ is divided into two parts as

$$\mathcal{M}_{\mathcal{S}}(1-\theta) = \underbrace{\int_0^\infty \left(\frac{1+\gamma_k}{e^{\sqrt{v}2\lambda}}\right)^{N_s} f(\gamma_k)\mathrm{d}\gamma_k}_{\mathcal{M}_1} + \underbrace{\int_0^{\gamma_{\mathrm{th}}} \left(1-\left(\frac{1+\gamma_k}{e^{\sqrt{v}2\lambda}}\right)^{N_s}\right) f(\gamma_k)\mathrm{d}\gamma_k}_{\mathcal{M}_2} \tag{42}$$

where $N_s = -1/2\theta N\log_2 e$, $v = 1-(1+\gamma_k)^{-2}$, and $\lambda = Q^{-1}(\epsilon)/\sqrt{2\,N}$, and $\mathcal{M}_1$ is expressed as

$$\mathcal{M}_1 = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty (1+\gamma_k)^{N_s}\gamma_k{}^{\alpha-1}e^{-\beta\gamma_k}e^{-2\lambda N_s\sqrt{v}}\mathrm{d}\gamma_k \tag{43}$$

To solve the RHS integral, the following two series expansion identities are used

$$\begin{cases} e^{-bx} = \sum_{n=0}^\infty \dfrac{(-1)^n b^n x^n}{n!} \\ [1-(1+\gamma_k)^{-2}]^{n/2} = \sum_{l=0}^{n/2} \binom{n/2}{l}(-1)^l(1+\gamma_k)^{-2l} \end{cases} \tag{44}$$

and, hence, there exists

$$e^{-2\lambda N_s\sqrt{v}} = \sum_{l=0}^{n/2} c_l \cdot (1+\gamma_k)^{-2l} \tag{45}$$

where c_l collects the coefficient of $(1+\gamma_k)^{-2l}$,[1] and is expressed as

$$c_l = \sum_{n\geq 2l} \frac{(-1)^{n+l}(2\lambda N_s)^n \binom{n/2}{l}}{n!} \tag{46}$$

and after some algebraic manipulations, $\mathcal{M}_1$ is expressed as

$$\mathcal{M}_1 = \frac{e^\beta\beta^\alpha}{\Gamma(\alpha)} \sum_{l=0}^{\infty} c_l \int_1^\infty x^{N_s-2l}(x-1)^{\alpha-1}e^{-\beta x}\mathrm{d}x \tag{47}$$

where $x = \gamma_k + 1$ and the RHS integral of $\mathcal{M}_1$ is solved using the following formula [32]:

$$\begin{aligned} &\int_u^\infty x^{a-1}(x-u)^{b-1}e^{-cx}\,\mathrm{d}x, \quad [\mathrm{Re}\,b>0, \mathrm{Re}\,cu>0] \\ &= \beta^{-\frac{b+a}{2}} u^{\frac{b+a-2}{2}} \Gamma(b) e^{-\frac{cu}{2}} W_{\frac{a-b}{2},\frac{1-b-a}{2}}(cu) \end{aligned} \tag{48}$$

where $W_{a,b}(z)$ stands for the Whittaker function that reads

$$W_{a,b}(z) = \frac{z^{b+\frac{1}{2}}e^{-z/2}}{\Gamma\left(b-a+\frac{1}{2}\right)} \int_0^\infty e^{-zt}t^{b-a-\frac{1}{2}}(1+t)^{b+a-\frac{1}{2}}\mathrm{d}t. \tag{49}$$

As a result, $\mathcal{M}_{\mathcal{S}}(1-\theta)$ is compactly expressed as

$$\mathcal{M}_{\mathcal{S}}(1-\theta) = e^{\beta/2} \sum_{l=0}^{\infty} c_l \beta^{\frac{b-a}{2}} W_{\frac{a-b}{2},\frac{1-b-a}{2}}(\beta) + \mathcal{M}_2 \tag{50}$$

where $a = N_s - 2l + 1$ and $b = \alpha$. $\mathcal{M}_2$ does not admit closed-form expression in general and is, hence, calculated numerically. Once $\mathcal{M}_1$ and $\mathcal{M}_2$ are obtained $p_d(w)$ is available readily.

Fortunately, experiments reveal that γ_{th} is extremely small and, hence, $\bar{R}_3$ in (41) and $\mathcal{M}_2$ in (50) are both negligible in the considered systems, and closed-form expressions are obtained as

$$\begin{cases} \bar{R}_k = \dfrac{\beta^\alpha e^\beta \log_2 e}{2} \sum_{k=1}^{\alpha} \frac{\Gamma(k-\alpha,\beta)}{\beta^k} - \dfrac{Q^{-1}(\epsilon)\log_2 e}{\sqrt{2\,N}} & (51) \\ \mathcal{M}_{\mathcal{S}}(1-\theta) = e^{\beta/2} \sum_{l=0}^{\infty} c_l \beta^{\frac{b-a}{2}} W_{\frac{a-b}{2},\frac{1-b-a}{2}}(\beta) & (52) \end{cases}$$

This claim is validated in Section IV.

3) Upper-Bounding $\bar{R}_k$ and $p_d(w)$: Though both (51) and (52) are obtained, it is perhaps desirable to derive more tractable results for rapid evaluation.

Lemma 3: $\bar{R}_k$ and $p_d(w)$ are upper-bounded, respectively, as

$$\begin{cases} \bar{R}_k \leq \mathcal{R}(N,\gamma_k,\epsilon)\big|_{\gamma_k=\mu_\gamma} & (53) \\ p_d(w) \leq \dfrac{\mathcal{M}'_{\mathcal{S}}(1-\theta)^w}{1-\mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}'_{\mathcal{S}}(1-\theta)} & (54) \end{cases}$$

where $\mathcal{M}'_{\mathcal{S}}(1-\theta)$ is the truncated result of (50) using the first $n_M < \infty$ terms in the summation, i.e.,

$$\mathcal{M}'_{\mathcal{S}}(1-\theta) = e^{\beta/2} \sum_{l=0}^{n_M} c_l \beta^{\frac{b-a}{2}} W_{\frac{a-b}{2},\frac{1-b-a}{2}}(\beta) + \mathcal{M}_2. \tag{55}$$

Proof: The proof of (53) is to leverage the fact that $\mathrm{E}[f(x)] \leq f[\mathrm{E}(x)]$ when $f(x)$ is concave. In the considered scenario,

[1]Precise calculations of c_l and the Gamma related functions may require specialized software, such as ADVNPIX.

$\mathcal{R}(N,\gamma_k,\epsilon)$ is approximately concave in high SNR regime [33] and, hence, there exists

$$\bar{R}_1 = \mathrm{E}[\mathcal{R}(N,\gamma_k,\epsilon)] \lessapprox \mathcal{R}(N,\gamma_k,\epsilon)\big|_{\gamma_k=\mu_\gamma}. \tag{56}$$

Apart from rapid evaluation, (56) provides a tractable expression for joint power allocation and rate adaption as detailed in Section III-C. Similarly, (54) is obtained since $\mathcal{M}_{\mathcal{S}}(1-\theta)$ is upper-bounded by $\mathcal{M}'_{\mathcal{S}}(1-\theta)$ according to (43) and $p_d(w)$ is a monotonically increasing function of $\mathcal{M}_{\mathcal{S}}(1-\theta)$ seeing that

$$\frac{\mathrm{d}}{\mathrm{d}\mathcal{M}_{\mathcal{S}}(1-\theta)}\left(\frac{\mathcal{M}_{\mathcal{S}}(1-\theta)^w}{1-\mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}_{\mathcal{S}}(1-\theta)}\right) > 0 \tag{57}$$

as long as $\mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}_{\mathcal{S}}(1-\theta) < 1$, which is the Gartner–Ellis condition. ■

C. Joint Power Allocation and Rate Adaption Based on Partial CSIT

The above discussion reveals that both SE and latency are entirely determined by γ_k, which is written as $\gamma_k = \mathcal{R}^{-1}(N, R_{\mathrm{FEC}}/B, \epsilon)$ taking code rate R_{FEC} and normalized bandwidth B into consideration.

The feasibility of designing mNOMA given partial CSIT was indicated in Lemma 1. The method is now detailed, where the main idea is to design a power allocation scheme taking code rate R_{FEC} into consideration, such that rate adaption is permitted to support diverse QoS. Given successful SIC, the SINR of the signal fed to LMMSE $\boldsymbol{w}_k$ is [4, Eq. (4)]

$$\gamma'_k = \frac{\|\boldsymbol{h}_k\|_{\mathrm{F}}^2\rho_k}{\sum_{m>k}\|\boldsymbol{h}_m\|_{\mathrm{F}}^2\rho_m + 1} \tag{58}$$

and the improvement of LMMSE filter is N_r as detailed in Lemma 4.

Lemma 4: Given $\kappa \gg 1$, the SINR improvement of LMMSE is N_r, i.e.,

$$\gamma_k/\gamma'_k = N_r. \tag{59}$$

Proof:

$$\gamma_k/\gamma'_k = \gamma_k\left(\sum_{m>k}\|\boldsymbol{h}_m\|_{\mathrm{F}}^2\rho_m + 1\right)/\|\boldsymbol{h}_k\|_{\mathrm{F}}^2\rho_k \tag{60}$$

seeing $\|\boldsymbol{h}_k\|_{\mathrm{F}}^2 \overset{\text{a.s.}}{\to} N_r$ and $\|\boldsymbol{h}_k\|_{\mathrm{F}}^2 = N_r \forall k$ is used in sequel indicating CSIT is *not* known at the transmitter. Taking $\sum_{m>k}\rho_m = (K-k)\rho_a$ into consideration, there exists

$$\frac{\gamma_k}{\gamma'_k} = \frac{\rho_k N_r}{\rho_a(K_e - N_r)}\cdot\frac{(K-k)N_r\rho_a + 1}{N_r\rho_k} \approx N_r \tag{61}$$

where $\rho_a(K_e - N_r) \approx (K_e - 1)\rho_a + 1$ is used, which always holds true as long as $K_e \gg N_r$.

It is worth mentioning that the above reasoning assumes no prior knowledge regarding the small-scale fading $\boldsymbol{h}_k \forall k$, and only the convergence property $\|\boldsymbol{h}_k\|_{\mathrm{F}}^2 \overset{\text{a.s.}}{\to} N_r$ is exploited. ■

To guarantee rate-fairness among all devices, it is configured that $\bar{R}_k = R_{\mathrm{FEC}}/B \quad \forall k \in [1, K]$, and hence, each device experiences the same SINR expressed as

$$\gamma_k = \gamma = \mathcal{R}^{-1}(N, R_{\mathrm{FEC}}/B, \epsilon). \tag{62}$$

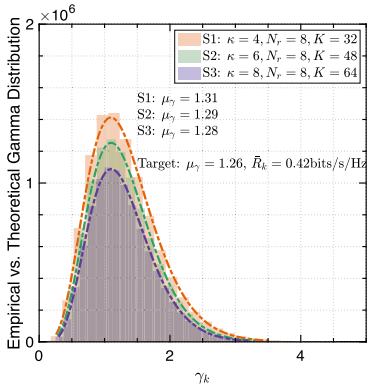

Fig. 1. Empirical histogram versus theoretical distribution of γ_k, given increasing κ.

Then there exists a geometric power allocation scheme, such that (61) holds for all devices, i.e.,

$$\rho_k = \rho_\Delta(1+\rho_\Delta)^{K-k} \tag{63}$$

where $\rho_\Delta = \gamma/N_r$. To verify, it is seen that

$$\rho_a = \frac{\sum_{m=k+1}^{K}\rho_m}{K-k} = \frac{1}{K-k}\cdot\frac{\rho_\Delta[1-(1+\rho_\Delta)^{K-k}]}{1-(1+\rho_\Delta)} \tag{64}$$

hence

$$\delta \approx \frac{1}{(K-k)\rho_\Delta}. \tag{65}$$

Therefore, $\mu_\gamma = 1/(\delta(\kappa-1)) = \gamma$ as required, and the power allocation is

$$p_k = \rho_k/g_k \quad \forall k \tag{66}$$

and the total transmit power is $p_{\mathsf{T}} := \sum_{k=1}^{K} p_k$.

Comments 1: It is now time to revisit the target problem (21) to see if (66) is a feasible solution. It is not too difficult to verify that $\gamma_k = \mathcal{R}^{-1}(N, R_{\mathrm{FEC}}/B, \epsilon)$ is equivalent to the optimality condition [34, Eq. (4)], from which (66) is derived. Hence, (66) is a feasible solution. However, it is noticed the total power actually required p'_{T} must exceed p_{T}, since the actual coded modulation is not Gaussian-distributed.

IV. Numerical Results and Discussions

In this section, the presented method is verified. The AP here is equipped with $N_r = 4, 6, 8$ antennas mainly; the user load K is configured, such that overloading factor $\kappa = K/N_r$ to indicate there are K data streams to be processed as LMMSE-SIC has not been implanted yet; the blocklength $N = 512, 2048$, and $g_k = 1 \forall k$ to make the results independent of path-loss. Two simple FECs are considered, REP with $R_{\mathrm{FEC}} = 1/3$ and $\mathrm{CC}(7,5)_8$ with $R_{\mathrm{FEC}} = 1/2$. In general CC is a stronger FEC than REP, but the latter is simpler. The target BLER ϵ and delay violation probability $p_d(w)$ are both 10^{-5}. The purpose is to verify the analysis presented in Section III, and to reveal the interplay among KPIs.

First in Figs. 1–3, the statistical behavior of SINR in terms of empirical and theoretical distribution is analyzed based on 10^7

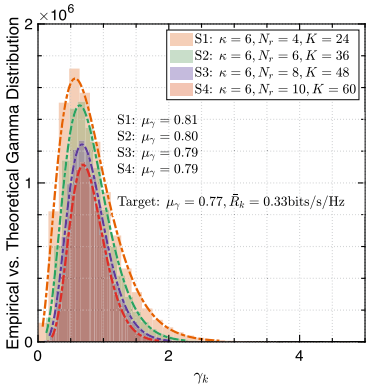

Fig. 2. Empirical histogram versus theoretical distribution of γ_k, given increasing N_r.

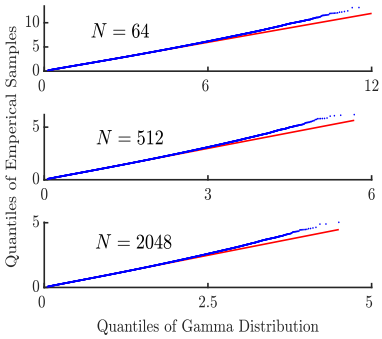

Fig. 3. Fitness of Gamma approximation of γ_k, given increasing N.

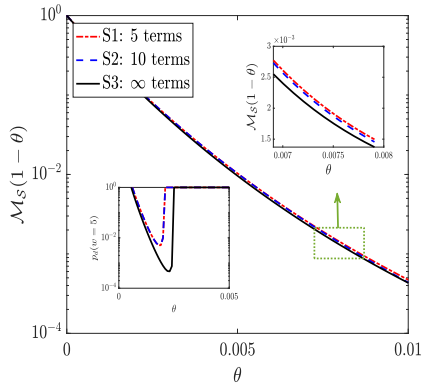

Fig. 4. Accuracy of $\mathcal{M}_\mathcal{S}(1-\theta)$ using different number of terms.

samples. In Fig. 1 the empirical and the theoretical distributions of γ_k are presented given three different sets of configurations: S1, S2, and S3. They are all configured with $N = 512$, $\bar{R}_k = 0.42$ bits/s/Hz, $N_r = 8$, and average $E_b/N_0 = 20$ dB, which can be obtained by combining MSK with $\text{CC}(7, 5)_8$. The only difference is κ, as a result of increasing connectivity K. In general, the empirical histograms are well aligned to the theoretical distributions and the target μ_γ is attainable, irrespective of the configurations. More importantly, the convergence behavior of γ_k with relation to N_r and κ is exposed, just as claimed in Section III-A. Specially, increasing κ while keeping N_r fixed facilitates the convergence of SINR, just as the tail behavior indicates.

In Fig. 2, the impact of N_r on the convergence of γ_k is evaluated given $\kappa = 6$. Four different sets of configurations are considered. The target SINR and SE are, respectively, $\mu_\gamma = 0.77$ and $\bar{R}_k = 0.33$ bits/s/Hz, which can be obtained by combining GMSK with REP. The blocklength here is $N = 2048$. It is observed that γ_k converges to μ_γ progressively as N_r increases from 4 to 10, just as the tail behavior indicates. The trend exemplified the result revealed in (31). More interestingly, as N_r increases from 8 to 10, μ_γ varies marginally suggesting that using $N_r \geq 10$ is perhaps fruitless in the considered system, hence, the following discussion is focused on $N_r < 10$.

When comparing Fig. 1 with Fig. 2, it seems that increasing the blocklength N from 512 to 2048 would facilitate the convergence, seeing the latter is less dispersive. To quantify this observation, quantile-quantile (Q-Q) plot [35] is employed and Fig. 3 is generated. In this figure, quantiles of Gamma distribution against quantiles of empirical samples are depicted. If empirical distribution is identical to the Gamma distribution, then the quantiles from the two underlying distributions are identical, hence, the result is the red line, otherwise the result is the blue dots. In general, the range of the quantiles is reduced suggesting improved convergence behavior. In all three cases, the blue dots form a line that is well aligned to the red line for small quantiles irrespective of the value of N, implying the empirical date set is also Gamma distributed. However, the two lines diverge as the quantile becomes very large. Such phenomenon suggests that the empirical distribution has heavier tail than the theoretical Gamma distribution does.

Fig. 4 demonstrates the number of terms n_M required for offering accurate enough approximation in (55). The system demonstrated in this figure is configured with $N = 2048$, $N_r = 8$, $K = 48$, $\text{CC}(7, 5)_8$, and GMSK modulation. The accuracy of $\mathcal{M}_\mathcal{S}(1-\theta)$ when $n_M = 5, 10, \infty$ are compared, where the result with $n_M = \infty$ is obtained by calculating the integral in (43) directly. It is observed that using $n_M = 10$ offers better accuracy than $n_M = 5$ does, though the improvement is negligible as the subfigure at top-right corner confirms. Increasing n_M further would not lead to noticeable improvement. This claim is also supported by the result presented at the bottom-left corner, where $p_d(w = 5)$ versus θ is presented given arrival rate $\eta = 1024$ and $E_b/N_0 = 10$ dB. Therefore, $n_M = 10$ is used in the sequel to construct $\mathcal{M}'_\mathcal{S}(1-\theta)$ in (55).

In Fig. 5, the analytical results regarding $p_d(w)$ are presented. The system is a GMSK based mNOMA with $N = 2048$, $N_r = 8$, $K = 40$, $R_{\text{FEC}} = 1/2$, and $E_b/N_0 = 12$ dB. First, the proposed upper-bounding technique, i.e., replacing $\mathcal{M}_\mathcal{S}(1-\theta)$ with $\mathcal{M}'_\mathcal{S}(1-\theta)$ that uses $n_M = 10$ terms is presented (marked with ✡). It is observed that the upper bound is negligibly larger than the exact results calculated with (50) or without $\mathcal{M}_2$ (52), which are, respectively, marked with ○ and □. It is reasonable to conclude that the upper-bounding technique and the proposed closed-form expression (52) are proved accurate numerically. Same claim can be made according to the delay violation probability $p_d(w)$ given $w = 3$ and $w = 6$. Similarly, these three methods are used to evaluate $p_d(w)$. The data arrival rate η

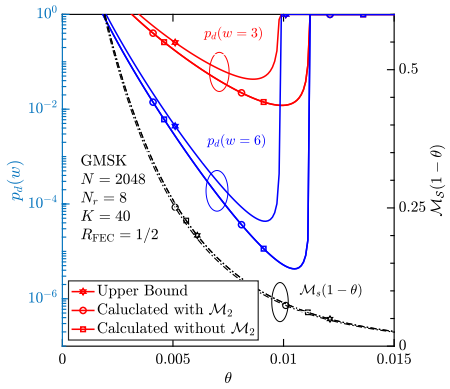

Fig. 5. $p_d(w)$ and $\mathcal{M}_S(1-\theta)$ versus θ.

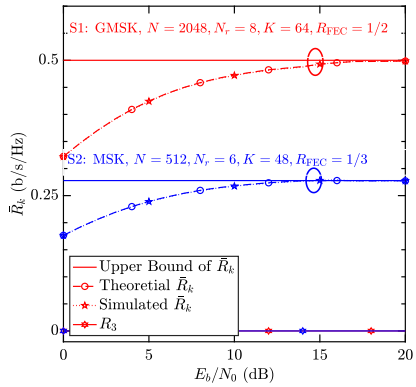

Fig. 6. Attainable per device SE of mNOMAs based on MSK and GMSK.

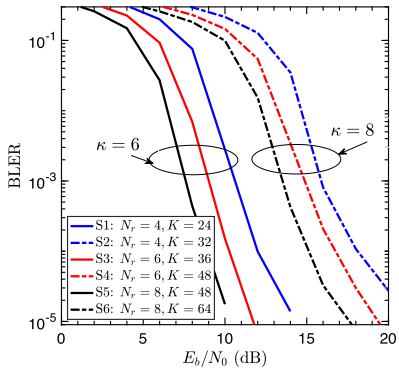

Fig. 7. BLER versus average E_b/N_0 with increasing connectivity K.

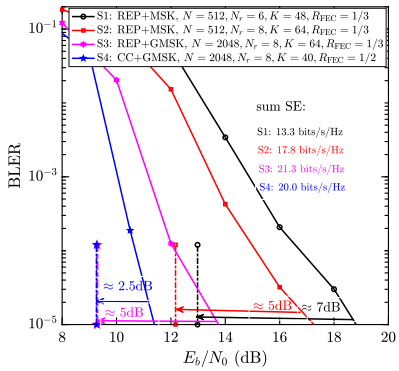

Fig. 8. Performance of mNOMA given full CSIT and partial CSIT.

is set to 1024. The accurate results using (50) are essentially identical to the results using (52), whereas the latter does not require $\mathcal{M}_2$. Further, the upper bound (54) using $\mathcal{M}'_S(1-\theta)$ is evaluated. It is observed that the upper bound captures the behaviors in both systems given different maximum allowable latency w. Though the gap to accurate result is noticeable, the upper bound is useful in simplifying the calculation as well as fast locating of the minimum $p_d(w)$, which can be then refined using accurate expression.

In Fig. 6, the analytical results regarding $\bar{R}_k$ are presented. The two systems S1 and S2 are of same κ, but of different $R_{\rm FEC}$ and modulation formats. As a result, the target ergodic SE is $\bar{R}_k = 0.5$ bits/s/Hz ($\bar{R}_k \approx 0.28$ bits/s/Hz), given GMSK (MSK) and $R_{\rm FEC}=1/2$ ($R_{\rm FEC}=1/3$). First, the upper bound (53) is evaluated for both sets of configurations, which turns out to be a constant, i.e., the target ergodic SE. Since the upper bound is calculated entirely based on μ_γ, the results are independent of actual power p_k. Hence, the upper bound is only accurate in high power domain. Fortunately, the theoretical ergodic SE using (51) (marked with ○) is demonstrated to attain the upper bound in high power domain. Here, the expression (51) instead of (41) is used, since the component R_3 contributes little to the result, as well confirmed in the figure. As a matter of fact, (51) coincides with the simulated result (marked with ✡) tightly, and they both attain the upper bound in high SINR domain disregarding the configurations. More interestingly, the behavior of $\bar{R}_k$ is accurately captured by the closed-form expression (51). As observed, S1 requires higher transmit power p_k than S2 does to attain the target SE. The reason is that the S1 targets higher SE than S2 does, and hence, more power is required as indicated in information theory. Nonetheless, the results presented in Fig. 6 validate the analysis in Section III-B1 and the claim made in Lemma 3.

In Fig. 7, the BLERs of the proposed mNOMAs with increasing connectivity K are presented. Six different configurations are considered, with $N=512$, MSK, and REP. These six configurations can be divided into two groups, with $\kappa=6$ and $\kappa=8$, respectively, and configurations within each group are separated according to N_r. The main trend is that mNOMA is interference-limited, hence, the performance in terms of BLER degrades as κ increases from 6 to 8. If κ is fixed, then, the performance degrades as N_r reduces from 8 to 4. There are two reasons mainly. First, small N_r means small diversity gain. Second, small N_r means significant randomness as discussed in Lemma 1 and further confirmed in Fig. 2.

In Fig. 8, the proposed design is exemplified by four systems S1, S2, S3, and S4 in terms of BLER versus E_b/N_0. Among these four schemes, S3 and S4 are designed to offer approximately identical sum SE but are configured differently. The vertical lines are the corresponding E_b/N_0 limits given *perfect* CSIT. The tradeoff among connectivity, reliability, and energy efficiency is observed. First, S1 and S2 are evaluated, which are essentially

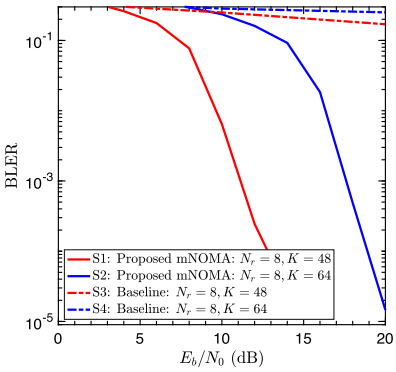

Fig. 9. Performance of mNOMA and existing scheme based on LDPC+BPSK.

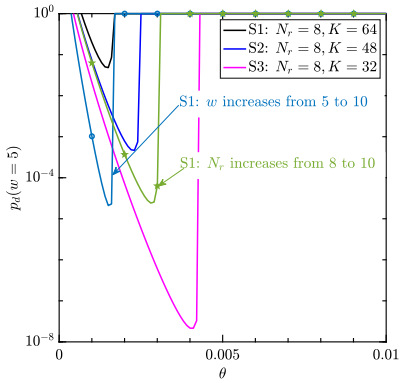

Fig. 10. $p_d(w)$ versus θ with increasing K.

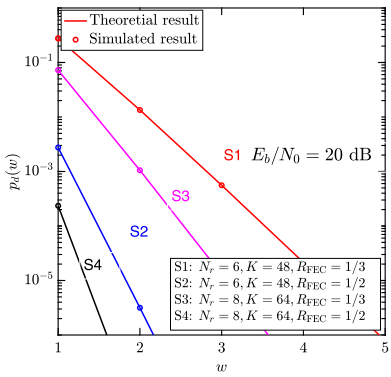

Fig. 11. κ versus R_{FEC} in reducing latency.

identical, except for N_r. This subtle difference leads to drastic results, i.e., S2 outperforms S1 by approximately 2 dB even the former accommodates much more devices and, hence, provides higher sum SE. The reason is the increased diversity gain and reduced randomness of γ_k, when N_r increases from 6 to 8, as revealed in (32). S3 and S4 increase the blocklength to 2048, and they are configured to offer similar sum SE around 20 bits/s/Hz, hence, the limits are almost identical. Though due to the increased blocklength, the required E_b/N_0 is significantly reduced in both systems, the gap between them is approximately 2.5 dB. This gap is mainly due to the reduced user load K and stronger FEC, while keeping $N_r = 8$. In general, it appears that N_r dominates the performance given the same κ, and smaller K is favorable if stronger FEC is employed. Nevertheless, the proposed design is able to offer verified ultra reliability. The gap obtained can then be used to calculate p'_{T} as commented in Comments 1. An important and interesting observation is that support of mMTC does not necessarily require very low rate FECs, which is prevalent in existing code domain NOMAs [3].

In Fig. 9, the proposed mNOMA is compared with the baseline scheme suggested in [20] and [21]. The baseline is a multiantenna system based on FEC scheme low-density parity-check (LDPC) and modulation scheme binary phase-shift keying (BPSK) and the coding parameters are optimized using density evolution technique. The code rate $R_{\text{FEC}} = 0.5$, and hence, the overall SE per device is 0.5 bits/s/Hz, since BPSK and GMSK occupy the same bandwidth. Apart from the FEC and modulation schemes, the blocklength N, antenna size N_r, and connectivity K remain the same in both the proposed mNOMA and the baseline design. The results are presented in terms of BLER versus E_b/N_0. It is observed that the proposed scheme outperforms the baseline design drastically. There are two reasons for this mainly. First of all, the baseline design is obtained using density evolution or similar techniques, which generates excellent performance only if $N \to \infty$. A typical setup in [20] and [21] is $N \geq 10^5$. Unfortunately, such method leads to significantly degraded performance inevitably in finite blocklength regime. Another reason is that power allocation is not employed in the recommended design, i.e., all devices are equally-powered, which prevents effective separation of different devices.

In Fig. 10, the latency with increasing connectivity K is evaluated. The system is configured with $N_r = 8$, $N = 2048$, $\eta = 1024$, GMSK, and $\text{CC}(7,5)_8$. As the connectivity K increases from 32 to 64, the value of $p_d(w)$ increases from $\approx 10^{-8}$ to $\approx 10^{-1}$. This result again indicates that mNOMA is interference-limited as expected. To support such high connectivity K while maintaining low $p_d(w)$, the allowable latency must increase. For example, the system S1 reduces $p_d(w)$ from $\approx 10^{-1}$ to $\approx 10^{-5}$, when w increases from 5 to 10. Based on this result, it is concluded that connectivity and latency are balanced in system design. Alternatively, increasing N_r to mitigate the interference and, hence, reduce the latency is also feasible. For example, keeping the rest of the parameters fixed while increasing N_r to 10 can attain $p_d(w) \approx 10^{-5}$ without incurring increased latency. The only drawback here is perhaps surged complexity since LMMSE-SIC is of complexity $O(N_r^3)$. Nevertheless, it is demonstrated that mNOMA can strike a good tradeoff among stringent KPIs if properly designed.

In Fig. 11, the latency of S1, S2, S3, and S4 are evaluated to reveal the interplay between N_r and R_{FEC} mainly. They are all configured with $\kappa = 8$, $N = 512$, $\eta = 128$, and GMSK modulation, but N_r and R_{FEC} vary. The simulated results (marked with $\circ$) and theoretical results (marked with $-$) using (52) and (18) are presented, which are demonstrated to coincide with each other. In general, all systems using the proposed design

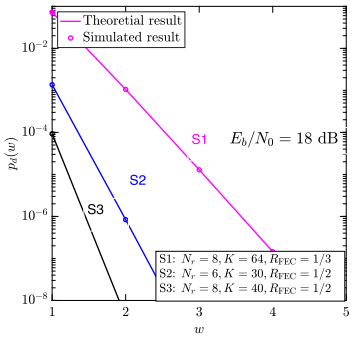

Fig. 12. κ versus N_r in reducing latency.

achieve the target $p_d(w)$. But it is again shown that N_r plays an important role. That is, the more N_r is the smaller w is, given $p_d(w) = 10^{-5}$, which can be derived from the results in Fig. 8. It is further revealed that higher code rate $R_{\rm FEC} = 1/2$ outperforms $R_{\rm FEC} = 1/3$ given the same setup. Intuitively, low rate mNOMA attempts more transmissions than high rate mNOMA does given the same amount of data, leading to longer latency. Quantitatively, the service process $\mathcal{M}_{\mathcal{S}}(1-\theta)$ of higher code rate offloads more data than lower code rate does given the same E_b/N_0, since the former reaches higher p_k, and hence, γ_k.

In Fig. 12, the latency of S1, S2, and S3 are evaluated to discuss the interplay between N_r and κ mainly, $N = 512$, $\eta = 128$, and GMSK modulation are used. They are configured to compare the delay performance of two cases: 1) Different $R_{\rm FEC}$, but identical sum SE, i.e., S1 versus S3; or 2) different N_r but identical κ, i.e., S2 versus S3. In case 1, high code rate prevails. The reason is that, as explained above, high code rate leads to less attempts of data offloading, and hence, lower probability of delay violation than low code rate system does. In case 2, large N_r prevails just like in Fig. 11. An interesting observation is that code rate plays a more important role than N_r does, as S2 versus S3 suggests, where S2 experiences smaller delay than S3 does, even though the latter is equipped with more antennas.

Comments 2: As the foregoing discussion has demonstrated extensively, the proposed mNOMA succeeds in QoS provision in terms of various KPIs in finite blocklength and overloaded regime without requiring small-scaling fading at the transmitter side. As expected, increasing N_r is beneficial in different scenarios. Interestingly, increasing user load K is provably favorable in terms of SINR convergence as long as N_r is sufficiently large while $N_r < K$. This result not only offers new method for designing multiantenna NOMA but also unifies two renowned tool sets, i.e., large dimensional analysis and random network calculus, to delivery guaranteed KPIs in IIoT.

However it is worth pointing out that the guaranteed KPIs are obtained presuming perfect CSI is accessible at AP, which is considered more or less optimistic in practice. Fortunately, there are some solutions available for overloaded mNOMA, such as using pilot [23] or deep-learning [24]. Both methods have been proved effective in estimating accurate enough CSI. Especially machine learning aided method can impressively reduce the normalized estimation error to -8 dB [24, Fig. 9], which is practically negligible. It appears that machine learning is a promising technique to resolve channel estimation for mNOMA. This technique is considered a following-up topic in the future.

V. Conclusion

The major concern of this article is to enable QoS provision for massive IIoT network using mNOMA given only partial CSIT, i.e., path-loss. To this end, the asymptotic performance using large dimensional analysis is developed to prove that SINR approaches a deterministic value and the small-scale fading is averaged out as the system is overloaded significantly. This property is then leveraged to derive performance analysis in terms of QoS KPIs including SE, delay, reliability, and connectivity, which are functions of SINR. The quantified interplay among KPIs is derived, which are used to design mNOMA based on joint power allocation and rate adaption in finite blocklength regime. The analysis and simulated results confirm that the proposed method is feasible in providing QoS even only given path-loss. The resultant design, i.e., mNOMA, can also find wide applications in deep space communications, where MSK family is pivotal as well.

References

[1] G. P. Hancke and V. Ç. Güngör, Eds., *Industrial Wireless Sensor Networks: Applications, Protocols, and Standards*, 1st ed., Boca Raton, FL, USA: CRC Press, Apr. 2013.
[2] A. Mahmood et al., "Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends," *IEEE Trans. Ind. Inf.*, vol. 18, no. 6, pp. 4122–4137, Jun. 2022.
[3] M. Mohammadkarimi, M. A. Raza, and O. A. Dobre, "Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 40–50, Dec. 2018.
[4] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tut.*, vol. 19, no. 2, pp. 721–742, Apr.–Jun. 2017.
[5] G. Ma, B. Ai, F. Wang, and Z. Zhong, "Joint design of coded tandem spreading multiple access and coded slotted ALOHA for massive machine-type communications," *IEEE Trans. Ind. Inf.*, vol. 14, no. 9, pp. 4064–4071, Sep. 2018.
[6] S. Hu et al., "Nonorthogonal interleave-grid multiple access scheme for industrial Internet of Things in 5G network," *IEEE Trans. Ind. Inf.*, vol. 14, no. 12, pp. 5436–5446, Dec. 2018.
[7] J. Zeng et al., "Achieving ultrareliable and low-latency communications in IoT by FD-SCMA," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 363–378, Jan. 2020.
[8] L. Bing, Y. Gu, T. Aulin, and J. Wang, "Design of auto-configurable random access NOMA for URLLC industrial IoT networking," *IEEE Trans. Ind. Inf.*, vol. 20, no. 1, pp. 190–200, Jan. 2024.
[9] W. U. Khan, F. Jameel, T. Ristaniemi, S. Khan, G. A. S. Sidhu, and J. Liu, "Joint spectral and energy efficiency optimization for downlink NOMA networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 645–656, Jun. 2020.
[10] Y. Gong, L. Zhang, R. Liu, K. Yu, and G. Srivastava, "Nonlinear MIMO for industrial Internet of Things in cyberphysical systems," *IEEE Trans. Ind. Inf.*, vol. 17, no. 8, pp. 5533–5541, Aug. 2021.
[11] X. Zhang, J. Wang, and H. V. Poor, "Statistical delay and error-rate bounded QoS provisioning for mURLLC over 6G CF M-MIMO mobile networks in the finite blocklength regime," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 652–667, Mar. 2021.
[12] J. Zhang, J. Fan, J. Zhang, D. W. K. Ng, Q. Sun, and B. Ai, "Performance analysis and optimization of NOMA-based cell-free massive MIMO for IoT," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9625–9639, Jun. 2022.
[13] H. R. Chi, C. K. Wu, N.-F. Huang, K.-F. Tsang, and A. Radwan, "A survey of network automation for industrial Internet-of-Things toward industry 5.0," *IEEE Trans. Ind. Inf.*, vol. 19, no. 2, pp. 2065–2077, Feb. 2023.

[14] H. Huang, J. Yang, H. Huang, Y. Song, and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, Sep. 2018.
[15] C.-J. Chun, J.-M. Kang, and I.-M. Kim, "Deep learning-based channel estimation for massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1228–1231, Aug. 2019.
[16] G. Lu, X. Dai, W. Zhang, Y. Yang, and F. Qin, "Nondata-aided Rician parameters estimation with redundant GMM for adaptive modulation in industrial fading channel," *IEEE Trans. Ind. Inf.*, vol. 18, no. 4, pp. 2603–2613, Apr. 2022.
[17] A. Perotti, S. Benedetto, and P. Remlein, "Spectrally efficient multiuser continuous-phase modulation systems," in *Proc. IEEE Int. Conf. Commun.*, 2010, pp. 1–5.
[18] N. Noels and M. Moeneclaey, "Iterative multiuser detection of spectrally efficient FDMA CPM," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5254–5267, Oct. 2012.
[19] L. Bing, T. Aulin, B. Bai, and H. Zhang, "Design and performance analysis of multiuser CPM with single user detection," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4032–4044, Jun. 2016.
[20] C. Schlegel and M. V. Burnashev, "The interplay between error control coding and iterative signal cancelation," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 3020–3031, Jun. 2017.
[21] N. Gao et al., "User-load-compatible masking schemes for raptor-like protograph-based LDPC codes in Gaussian multiple access channels," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 7652–7664, Aug. 2021.
[22] J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*. New York, NY, USA: Springer, 1986.
[23] C. Novak, G. Matz, and F. Hlawatsch, "IDMA for the multiuser MIMO-OFDM uplink: A factor graph framework for joint data detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4051–4066, Aug. 2013.
[24] H. Yu, Z. Fei, Z. Zheng, N. Ye, and Z. Han, "Deep learning-based user activity detection and channel estimation in grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2202–2214, Apr. 2023.
[25] Z. Li, W. Deng, W. Pei, Y. Xia, C. Zhu, and D. P. Mandic, "SINR analysis of MIMO systems with widely linear MMSE receivers for the reception of real-valued constellations," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, 2020, pp. 1–5.
[26] P. Popovski et al., "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.
[27] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surv. Tut.*, vol. 17, no. 1, pp. 92–105, Jan.–Mar. 2015.
[28] D. S. Bernstein, *Scalar, Vector, and Matrix Mathematics: Theory, Facts, and Formulas - Revised and Expanded Edition*, revised ed., Princeton, NJ, USA: Princeton Univ. Press, Feb. 2018.
[29] R. Muller and S. Verdu, "Design and analysis of low-complexity interference mitigation on vector channels," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 8, pp. 1429–1441, Aug. 2001.
[30] Z. Bai, Y. Chen, and Y.-C. Liang, Eds., *Random Matrix Theory and its Applications: Multivariate Statistics and Wireless Communications*. Singapore: World Scientific, 2009.
[31] D. N. C. Tse and O. Zeitouni, "Linear multiuser receivers in random environments," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 171–188, Jan. 2000.
[32] I. S. Gradshteĭn and D. Zwillinger, *Table of Integrals, Series, and Products*, 8th ed., Amsterdam, The Netherlands: Elsevier, 2015.
[33] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.
[34] 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, Oct. 2015.
[35] A. Law, *Simulation Modeling and Analysis*, 5th ed., New York, NY, USA: McGraw Hill, Jan. 2014.

Li Bing (Member, IEEE) received the Ph.D. degree in communication engineering from Xidian University, Xi'an, China, in 2014.

He then joined the School of Telecommunications, Xidian University. From 2010 to 2012, he was a Visiting Scholar with Chalmers University of Technology, Göteborg, Sweden. Since 2018, he has been an Associate Professor with Northwestern Polytechnical University, Xi'an, China.

Yating Gu (Graduate Student Member, IEEE) received the B.Eng. degree in automation from Nanjing Normal University, Nanjing, China, in 2021. She is currently working toward the M.Eng. degree in software engineering with Northwestern Polytechnical University, Xi'an, China.

Her research interests include Internet of Things and nonconvex optimization.

Lanke Hu received the B.Eng. degree in software engineering in 2021 from Northwestern Polytechnical University, Xi'an, China, where he is currently working toward the M.Eng. degree in electronic information.

His research interests include polar codes and nonorthogonal multiple access.

Tor Aulin (Life Fellow, IEEE) was born in Malmö, Sweden, in 1948. He received the M.S. degree in electrical engineering from the University of Lund, Lund, Sweden, in 1974, and the Dr. Techn. (Ph.D.) degree in electronic engineering from the Institute of Telecommunication Theory, University of Lund, in 1979.

In 1983, he was a Research Professor (Docent) of information theory with Chalmers University of Technology, Göteborg, Sweden. He has authored and coauthored nearly some 200 technical papers and has also authored the book Digital Phase Modulation (Plenum, 1986) as a result of his extensive research in this area at that time.

Dr. Aulin has two papers among the best (Best-of-the-Best) published during the first 50 years of the IEEE COMSOC, selected in connection with their 50th anniversary in 2002.

Yue Yin received the Ph.D. degree in microelectronics from Xidian University, Xi'an, China, in 2017.

Since then, he has been with Northwestern Polytechnical University, Xi'an, as an Associate Professor. His research interests include electronic science and technology, integrated circuit system design, and integrated circuit reliability analysis.

Jue Wang received the M.Eng. degree in communication engineering from Xidian University, Xi'an, China, in 2012.

Since 2012, he has been with the 20th Research Institute of China Electronic Technology Group Corporation (CETC), Xi'an,, China, where he is a Senior Research Scientist. He is currently a joint Ph.D. Scholar with Xidian University and CETC. His research interests include electronic and communication engineering.