Spatially-Correlated Rician-Faded Multi-Relay Multi-Cell Massive MIMO NOMA Systems

Dheeraj Naidu Amudala, Bibhor Kumar, and Rohit Budhiraja

Abstract

We consider the downlink of a relay-aided multi-cell massive multi-input multi-output (mMIMO) system, where in each cell the base station (BS) serves its users via multiple relays by employing non-orthogonal multiple access (NOMA). We model this system by considering spatially-correlated Rician-faded channels and their estimation errors. The users, consequently, perform imperfect successive interference cancellation (SIC). We derive a lower bound on the spectral efficiency (SE) of this system. We then optimize the non-convex global energy efficiency (GEE) metric, which is a fractional function of the optimization variables. We solve this problem by considering a low-complexity alternating minimization maximization approach, which splits a complex joint problem into multiple simpler convex surrogate sub-problems. We propose a novel surrogate function to exploit this framework, and analytically show that it satisfies the desirable properties of a valid surrogate function. We numerically show that i) reusing the pilots in each cell, when the channel has sufficiently hardened, provides higher SE than using orthogonal pilots in all cells; and ii) the proposed GEE algorithm provides similar GEE as that of an existing joint optimization framework, but with much less complexity.

Index Terms

Block optimization, energy efficiency, massive MIMO, multiple relays.

I. Introduction

Massive multi-input multi-output (mMIMO) is a key technology in the fifth-generation cellular systems that ensures high spectral efficiency (SE) and energy efficiency (EE) [1]. An mMIMO base station (BS), equipped with massive antenna arrays, leverages spatial diversity to serve multiple users on the same spectral resource [1]–[3]. Orthogonal multiple access (OMA) techniques, which allocates orthogonal time-frequency resources to various users, are employed in a multi-user mMIMO system to mitigate inter-user interference (IUI). Although OMA effectively suppresses IUI, it degrades the system SE when users observe poor channels [4], [5].

Non-orthogonal multiple access (NOMA) employs superposition coding to multiplex different users signals in power domain, and is shown to have a higher SE than OMA [4]–[6]. NOMA users employ successive interference cancellation (SIC) at the receiver to mitigate IUI. Integrating NOMA in mMIMO systems can significantly boost the system SE and EE, and is recently

The authors are with the Department of Electrical Engineering, IIT Kanpur, India. email: {dheeraja, bibhork, rohitbr}@iitk.ac.in. A part of this work is accepted in IEEE International Conference on Communications 2022 (ICC 2022).

investigated in [5]–[10]. The authors in [5] analyzed the outage probability and ergodic rate of a downlink mMIMO NOMA system with imperfect SIC, and showed that NOMA with reliable SIC outperforms OMA. Liu *et al.* in [6] proposed a capacity-achieving iterative low-complexity detector for a NOMA enabled multi-user MIMO system. Mandwaria *et al.* in [7] studied the weighted sum EE of a downlink mMIMO NOMA system with perfect SIC at the users. Both these works considered a *single-cell* mMIMO NOMA system, wherein orthogonal pilot based channel estimation and simple linear BS precoding schemes effectively mitigate IUI [5]. In a multi-cell scenario, due to limited coherence interval, pilot sequences are reused in the adjacent cells. The channel estimates of pilot-sharing users consequently get contaminated, which degrades the quality of SIC and eventually the SE [8], [10]. The authors in [8] proposed user-clustering and pilot assignment strategies for the downlink of spatially-correlated NOMA-aided multi-cell mMIMO system, and derived closed-form SE expression. Bashar *et al.* in [10] proposed per-user SE based NOMA/OMA mode switching strategy for NOMA-based cell-free mMIMO system.

Cooperative relaying, wherein the BS serves various users via relays, improves the system coverage area, along with its SE and EE [11]. Vaezi *et al.* in [11] provided an overview of the applications of NOMA in the downlink of single-hop mMIMO systems, cooperative relaying and relay-aided mMIMO systems. Single-cell multi-relay-assisted downlink mMIMO NOMA systems have recently been analyzed in [12]–[17]. The authors in [12] derived asymptotic SE expressions by assuming perfect channel state information (CSI) at the BS, and perfect SIC at the users. Chen *et al.* in [13] derived a closed-form SE expression for zero forcing precoding at the BS. Both these works assumed perfect CSI at the BS, and perfect SIC at the users. Mandawaria *et al.* in [14] extended the work in [13], and analyzed the SE and EE with CSI errors at the BS and imperfect SIC at the users. Li *et al.* in [15], in contrast to [12]–[14], considered multi-antenna relays, and derived closed-form SE expressions by assuming imperfect CSI, pilot contamination and maximal ratio (MR) transmission at the BS. The work in [15] was extended in [16] for zero forcing/null-space precoding scheme at the BS. Zhang *et al.* in [17] analyzed the SE for the system when users employ minimum mean square error successive interference cancellation receiver.

The aforementioned *single-cell* works that derived asymptotic SE results [12] or closed-form SE results [13], [14] assumed *uncorrelated* Rayleigh-faded channels, an assumption which is not entirely practical. An mMIMO BS, due to its closely-spaced antennas and the lack of rich scattering environment, observes spatially-correlated channels [2], [16]. Further, due to the

proximity of the users with relays, and relays with BS, their channels also have a deterministic line of sight (LoS) component [3], [18], which the existing works in [12]–[14] have ignored. Li *et al.* in [16] considered the downlink of a spatially-correlated relay-enabled mMIMO NOMA system with MR processing at the BS, and derived closed-form SE expression. This work, however, ignored the LoS channel component. Zhang *et al.* in [18] analyzed the SE of a single-hop downlink cell-free mMIMO NOMA system with spatially-correlated Rician faded channels. This work, however, did not consider relays. Both spatial correlation and LoS components are vital aspects which can impact the gains accrued by relays, mMIMO and NOMA technologies [2], [3]. The current work addresses these gaps by considering the downlink of a multi-cell multi-relay mMIMO NOMA system with spatially-correlated Rician-faded channels, which can model both LoS and non LoS (NLoS) components.

Design of optimal power allocation algorithms at the BS to optimize SE are recently investigated in mMIMO NOMA literature [5], [8], [13]. Sena *et al* in [5] designed a max-min SE based power allocation scheme to ensure fairness among different NOMA groups. Kudathanthirige *et al* in [8] optimized the sum SE of a spatially-correlated single-hop downlink multi-cell mMIMO NOMA system. Chen *et al* in [13] considered a single-cell relay-assisted mMIMO NOMA system, and optimized the weighted sum SE by assuming perfect CSI/SIC at the BS/users.

Proliferation in the number of active wireless devices, and the consequent increase in energy consumption, has necessitated the design of energy-efficient wireless systems. Global energy efficiency (GEE) metric, defined as the ratio of the network SE to its total power consumption, is recently optimized for downlink mMIMO NOMA systems [14], [19]–[21]. Zamani *et al.* in [19] optimized the GEE of a single-hop NOMA system with single-antenna BS and users. The authors in [20] considered the downlink of a wireless-powered single-cell mMIMO NOMA system, and optimized its GEE. The authors in [21] optimized the GEE of a downlink single-hop hybrid mMIMO NOMA system. Mandawaria *et al.* in [14] optimized the GEE of a multi-relay mMIMO NOMA system. This work, however, considered the downlink of a single-cell system with uncorrelated Rayleigh faded channels. The works that derived closed-form/asymptotic SE expressions and/or optimized SE/GEE for relay-aided mMIMO NOMA system are summarized in Table I. We infer from it that existing works have neither derived the closed-form SE lower bound nor optimized the GEE for the downlink of a *multi-cell multi-relay* mMIMO NOMA system with spatially-correlated Rician faded channels. The current work addresses these gaps with its **main contributions**, which are:

Table I: Summary of multi-relay-aided downlink mMIMO-NOMA literature focusing on SE and GEE.

Ref.	System	CSI/SIC	Channel	Correlation	SE expression	Optimization	Complexity	
							SIC	Optimization
[12]	single-cell	perfect	Rayleigh	✗	asymptotic	–	high	trivial
[13]	single-cell	perfect	Rayleigh	✗	approximate	sum SE	high	trivial
[16]	single-cell	imperfect	Rayleigh	✓	approximate	–	low	trivial
[14]	single-cell	imperfect	Rayleigh	✗	closed-form	GEE	low	$\mathcal{O}\left(n^4\right)^{\dagger}$
Current	**multi-cell**	**imperfect**	**Rician**	✓	**closed-form**	**GEE**	**low**	$\mathcal{O}\Big(\Big(\sum_{k=1}^{K} n_k^4\Big)\Big)^{\dagger}$

† Here n is the total number of optimization variables, which is split into K blocks of n_k variables each, such that $\sum_{k=1}^{K} n_k = n$.

1) We consider the downlink of a multi-cell multi-relay-aided mMIMO NOMA system with spatially-correlated Rician-faded channels. The BSs and the users, unlike [12], [13], do not have the CSI, and therefore estimate it by using the pilots transmitted by the relays. We address the derivation difficulty caused by the imperfect SIC/CSI, pilot contamination and spatially-correlated Rician channels, and derive a closed-form SE lower bound, which is a function of long-term channel statistics. Due to pilot contamination and spatially-correlated Rician-faded channels, the current SE analysis is a non-trivial extension of [14], [16].

2) We maximize GEE by optimally allocating the BS and relay transmit powers by using the derived SE lower bound. GEE optimization is a non-convex fractional programming (FP) problem. This is due to the sum SE in its numerator, which consists of fractional functions of optimization variables [14], [22]. The authors in [14] maximized the GEE by decoupling the scalar ratios by using the quadratic transform (QT) [23], and by approximating the non-convex product-of-variables as convex. This method jointly optimized the transmit power of the BS and multiple relays, and therefore has a extremely high computation complexity. Also, the current SE expression, due to correlated Rician fading, contains complicated coupled product terms, when compared with that in [14]. Their optimization, when applied directly to the current system, will have a high complexity. We will analytically and numerically validate this aspect later.

3) We develop a low-complexity iterative algorithm based on the alternating minorization maximization (AMM) to solve the non-convex GEE problem [24]. The proposed AMM approach first decomposes the joint GEE problem into multiple sub-problems by separating the optimization variables into multiple blocks, and then alternatingly optimizes a variable block by fixing the remaining ones. The resultant sub-problems are, however, still non-convex. The AMM approach then uses the MM technique to optimize the non-convex sub-problems by designing convex

surrogates functions [25]. We design novel surrogate functions and prove that they satisfy all the required surrogate function properties [25].

4) We numerically show that the proposed AMM-based GEE optimization provides a similar GEE as [14], but with a reduced complexity. We also numerically investigate the number of relays that the system need to deploy, and the users/relay it needs to serve, to maximize the SE. We also show that for a multi-cell multi-relay system that experiences pilot contamination, NOMA is able to provide higher SE fairness and power efficiency than its OMA counterpart.

II. System Model

We consider, as shown in Fig. 1, the downlink of an L-cell mMIMO system. In each cell, an N-antenna mMIMO BS serve clusters of cell-edge users which, due to high path loss, have an extremely weak direct link with it. The BS serves such coverage-limited users via single-antenna half-duplex amplify-and-forward (AF) relays. These relays are installed at a high altitude such that the BS-relay and users-relay channels have both LoS and NLoS components. These channels, therefore, have Rician probability density function (pdf). Further, the lth BS communicates with a cluster of $\mathcal{U}_{lk}$ single-antenna users via the relay R_{lk} by employing NOMA for $l = 1$ to L and $k = 1$ to K. The users associate with a particular relay based on their spatial locations i.e., users close to a relay form a cluster.

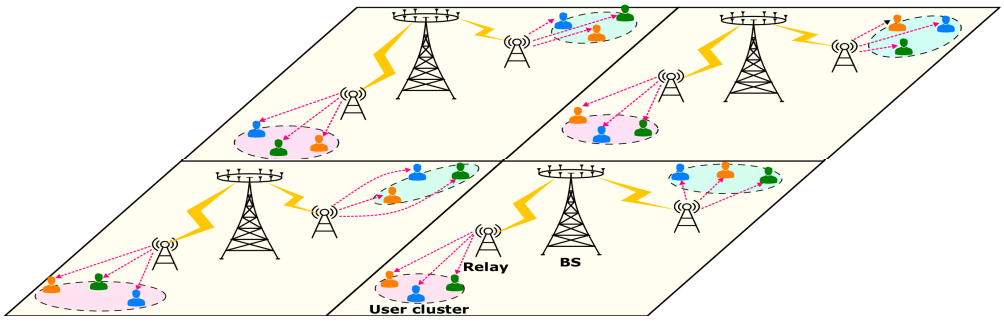

Fig. 1: Multi-cell relay-aided downlink mMIMO system model.

We next explain the communication protocol with all nodes operating in time division duplex mode. Here a τ_c symbol long channel coherence interval is divided into channel estimation (CE) and data transmission phases of $\tau \leq \tau_c$ and $(\tau_c$ - $\tau)$ symbols, respectively. In the CE phase, the relays will transmit pilots using which the BS and users estimate the BS-to-relay, and the relay-to-user channels, respectively. In the data transmission phase, the BS employs NOMA to serve users via relays. Before explaining the CE and data transmission phases, we model different channels in the system.

BS - relay channel: We denote the kth relay in the lth cell as R_{lk}. The channel from the jth BS to R_{lk} is denoted as $\mathbf{h}_{lk}^{j} \in \mathbb{C}^{N\times 1}$. Due to the lack of rich scattering around the BS, and the

limited antenna spacing, the channel $\mathbf{h}_{lk}^{j}$ is spatially-correlated [2]. The presence of LoS link between the BS and relays leads to $\mathbf{h}_{lk}^{j}$ having a Rician pdf. It is, accordingly, mathematically expressed as follows

$$\mathbf{h}_{lk}^{j} = \overline{\mathbf{h}}_{lk}^{j} + (\mathbf{R}_{lk}^{j})^{\frac{1}{2}} \mathbf{h}_{lk}^{j,\text{NLoS}}, \text{ where } \overline{\mathbf{h}}_{k} = \sqrt{\frac{K_{lk}^{j}\beta_{lk}^{j}}{1+K_{lk}^{j}}} \mathbf{h}_{lk}^{j,\text{LoS}} \text{ and } \mathbf{R}_{lk}^{j} = \frac{\beta_{lk}^{j}}{1+K_{lk}^{j}} \overline{\mathbf{R}}_{lk}^{j}. \quad (1)$$

Here K_{lk}^{j} is the Rician factor and β_{lk}^{j} is the large-scale fading coefficient. The vector $\mathbf{h}_{lk}^{j,\text{NLoS}}$ with pdf $\mathcal{CN}(0, \mathbf{I}_N)$, and $\mathbf{h}_{lk}^{j,\text{LoS}} = \left[1, e^{j2\pi d_\lambda \sin(\varphi_{lk}^{j})}, \ldots, e^{j2\pi d_\lambda (N-1)\sin(\varphi_{lk}^{j})}\right]^T$ model the NLoS and LoS components, respectively. Here φ_{lk}^{j} is the nominal angle between the jth BS and the relay R_{lk}, and d_λ is the inter-antenna distance (in fractions of wavelengths). The matrix $\overline{\mathbf{R}}_{lk}^{j}$ characterizes the spatial correlation of the NLoS component. This is unlike [14] which considered only uncorrelated Rayleigh-faded channels. The spatial correlation can dramatically alter the system insights, and is therefore crucial to consider [2], [26].

Relay - user channel: The channel vector from the relay $R_{jk'}$ to the nth user in the cluster $\mathcal{U}_{lk}$ is denoted as $g_{lk,n}^{jk'} \in \mathbb{C}^{1\times 1}$. The channels $g_{lk,n}^{jk'}$ are also Rician-faded such that

$$g_{lk,n}^{jk'} = \overline{g}_{lk,n}^{jk'} + (\gamma_{lk,n}^{jk'})^{\frac{1}{2}} g_{lk,n}^{jk',\text{NLoS}}, \text{ where } \overline{g}_{lk,n}^{jk'} = \sqrt{\frac{K_{lk,n}^{jk'}\beta_{lk,n}^{jk'}}{1+K_{lk,n}^{jk'}}} \text{ and } \gamma_{lk,n}^{jk'} = \frac{\beta_{lk,n}^{jk'}}{1+K_{lk,n}^{jk'}}. \quad (2)$$

Here $K_{lk,n}^{jk'}$ is the Rician factor and $\beta_{lk,n}^{jk'}$ is the large-scale fading coefficient. The scalar $g_{lk,n}^{jk',\text{NLoS}} \sim \mathcal{CN}(0,1)$ model the small-scale fading in the NLoS component.

Channel Estimation: Recall that the users in the current system, due to large path loss and shadowing, do not have a direct link with the BSs [14], [27]. This prohibits the users and the BSs to estimate the end-to-end downlink and uplink CSI, respectively. It is, therefore, crucial to design various precoders, combiners and optimization algorithms based on the local CSI available with the BSs and users. The K relays in each cell, which are connected to the BSs and users, can help them in estimating their first-hop BS-to-relay and the second-hop relay-to-users CSI, respectively. The proposed design uses only this CSI. To estimate the CSI, the kth relay in lth cell R_{lk}, transmits $\tau \geq K$-length pilot $\boldsymbol{\psi}_k$ to the BSs and users. The pilots are mutually orthogonal in each cell i.e., $\boldsymbol{\psi}_k^H \boldsymbol{\psi}_j = 0$ for $k \neq j$ and $\|\boldsymbol{\psi}_k\|^2 = \tau$. The kth relay in each cell shares the same pilot sequence, which causes pilot contamination. The lth BS uses this pilot to estimate $\mathbf{h}_{lk}^{l}$ i.e., its first-hop uplink CSI from the relay R_{lk}. The nth user associated with the kth relay in the lth cell, which belongs to the cluster $\mathcal{U}_{lk}$, uses the same pilot to estimate $g_{lk,n}^{lk}$, which is its second-hop downlink CSI from the relay R_{lk}. The pilot signals received by the lth BS and the nth user in cluster $\mathcal{U}_{lk}$ are given respectively as follows:

$$\mathbf{Y}^l = \sqrt{p_p}\sum_{l'=1}^{L}\sum_{k'=1}^{K}\mathbf{h}_{l'k'}^l\boldsymbol{\psi}_{k'}^T + \mathbf{N}^l \text{ and } \mathbf{y}_{lk,n}^p = \sqrt{p_p}\sum_{l'=1}^{L}\sum_{k'=1}^{K} g_{lk,n}^{l'k'}\boldsymbol{\psi}_{k'}^T + \mathbf{n}_{lk,n}^p. \tag{3}$$

Here p_p is the pilot transmit power and $\mathbf{N}^l$ (resp. $\mathbf{n}_{lk,n}^p$) is the additive white Gaussian noise (AWGN) at the lth BS (resp. nth user in cluster $\mathcal{U}_{lk}$) with independent and identically distributed (i.i.d) $\mathcal{CN}(0,1)$ elements. The BS estimates $\mathbf{h}_{lk}^l$ by projecting $\mathbf{Y}^l$ on to $\boldsymbol{\psi}_k$ as follows

$$\tilde{\mathbf{y}}_k^l = \mathbf{Y}^l\boldsymbol{\psi}_k^* = \sum_{l'=1}^{L}\sqrt{p_p}\tau\mathbf{h}_{l'k}^l + \mathbf{N}^l\boldsymbol{\psi}_k^*. \tag{4}$$

The MMSE estimate of the channel $\mathbf{h}_{lk}^l$ is therefore obtained using (4) as

$$\widehat{\mathbf{h}}_{lk}^l = \overline{\mathbf{h}}_{lk}^l + \sqrt{p_p}\mathbf{R}_{lk}^l\boldsymbol{\Psi}_{lk}\left[\tilde{\mathbf{y}}_k^l - \overline{\mathbf{y}}_k^l\right], \tag{5}$$

where $\boldsymbol{\Psi}_{lk} = \left(\mathbf{I}_N + \tau p_p\sum_{l'=1}^{L}\mathbf{R}_{l'k}^l\right)^{-1}$ and $\overline{\mathbf{y}}_k^l = \sum_{l'=1}^{L}\sqrt{p_p}\tau\overline{\mathbf{h}}_{l'k}^l$. The channel estimation error $\boldsymbol{\varepsilon}_{lk}^l = \mathbf{h}_{lk}^l - \widehat{\mathbf{h}}_{lk}^l$ has pdf $\mathcal{CN}(0, \mathbf{R}_{lk}^l - \widehat{\mathbf{R}}_{lk}^l)$, where $\widehat{\mathbf{R}}_{lk}^l = \tau p_p\mathbf{R}_{lk}^l\boldsymbol{\Psi}_{lk}\mathbf{R}_{lk}^l$ [28].

To estimate $g_{lk,n}^{lk}$, the nth user in cluster $\mathcal{U}_{lk}$ projects its received signal $\mathbf{y}_{lk,n}^p$ onto $\boldsymbol{\psi}_k$ as $\tilde{y}_{lk,n}^p = \mathbf{y}_{lk,n}^p\boldsymbol{\psi}_k^* = \sum_{l'=1}^{L}\sqrt{p_p}\tau g_{lk,n}^{l'k} + \mathbf{n}_{lk,n}^p\boldsymbol{\psi}_k^*$. The MMSE estimate of $g_{lk,n}^{lk}$ is [28]

$$\widehat{g}_{lk,n}^{lk} = \overline{g}_{lk,n}^{lk} + \frac{\sqrt{p_p}\gamma_{lk,n}^{l'k}}{1+\sum_{l'=1}^{L}\tau p_p\gamma_{lk,n}^{l'k}}\left[\tilde{y}_{lk,n}^p - \overline{y}_{lk,n}^p\right], \quad \text{where } \overline{y}_{lk,n}^p = \sum_{l'=1}^{L}\sqrt{p_p}\tau\overline{g}_{lk,n}^{l'k}. \tag{6}$$

The channel estimation error $e_{lk,n}^{lk} = g_{lk,n}^{lk} - \widehat{g}_{lk,n}^{lk}$ is statistically independent from the MMSE estimate $\widehat{g}_{lk,n}^{lk}$ and is distributed as $e_{lk,n}^{lk} \sim \mathcal{CN}(0, \gamma_{lk,n}^{lk} - \upsilon_{lk,n}^{lk})$ with $\upsilon_{lk,n}^{lk} = \frac{\tau p_p(\gamma_{lk,n}^{lk})^2}{\sum_{l'=1}^{L}\tau p_p\gamma_{lk,n}^{l'k}+1}$ [28].

Downlink data transmission phase: It is divided into two following time slots:

1) First time slot − BS to relay transmission: A BS first uses NOMA to superpose transmit signals of users in its cell, and then precodes and transmits it to the relays. Let $s_{jk,n}$ be the signal of the nth user in cluster $\mathcal{U}_{jk}$. The precoded NOMA signal transmitted by the jth BS is

$$\mathbf{x}^j = \sum_{k=1}^{K}\mathbf{w}_{jk}\sum_{n=1}^{\mathcal{U}_{jk}}\sqrt{p_{jk,n}}s_{jk,n} \triangleq \sum_{k=1}^{K}\mathbf{w}_{jk}x_{jk}. \tag{7}$$

Here $x_{jk} = \sum_{n=1}^{\mathcal{U}_{jk}}\sqrt{p_{jk,n}}s_{jk,n}$ is the NOMA signal for the users in cluster $\mathcal{U}_{jk}$, $\mathbf{w}_{jk} \in \mathbb{C}^{N\times 1}$ is the transmit precoder and $p_{jk,n}$ is the transmit power corresponding to the data of nth user in the cluster $\mathcal{U}_{jk}$, respectively. The precoder $\mathbf{w}_{jk}$ is designed based on MR transmission as $\mathbf{w}_{jk} = \frac{(\widehat{\mathbf{h}}_{jk}^j)^*}{\sqrt{\delta_{jk}}}$, with $\delta_{jk} = \mathbb{E}\left(\left\|\widehat{\mathbf{h}}_{jk}^j\right\|^2\right)$. The signal received by the kth relay in lth cell is given as

$$y_{R_{lk}} = \underbrace{\sum_{l''=1}^{L}\sum_{k''=1}^{K}(\mathbf{h}_{lk}^{l''})^T\mathbf{w}_{l''k''}x_{l''k''}}_{\tilde{y}_{R_{lk}}} + z_{R_{lk}}. \tag{8}$$

The scalar $z_{R_{lk}}$ with pdf $\mathcal{CN}(0,1)$ is the AWGN at the kth relay in lth cell.

2) Second time slot − relay to user transmission: The AF relay R_{lk} amplifies its received precoded NOMA signal as $x_{R_{lk}} = \mu_{lk} y_{R_{lk}}$, and then broadcasts it to the users in its cluster. Here μ_{lk} is the amplification factor designed to constrain the maximum relay transmit power to q_{lk} i.e., $\mathbb{E}[|x_{R_{lk}}|^2] = q_{lk}$. The expression for amplification factor is therefore given as

$$\mathbb{E}(|x_{R_{lk}}|^2) = q_{lk} \implies \mu_{lk}^2 \mathbb{E}(|y_{R_{lk}}|^2) = q_{lk} \implies \mu_{lk} = \sqrt{\frac{q_{lk}}{\mathbb{E}\left[\left|\tilde{y}_{R_{lk}} + z_{R_{lk}}\right|^2\right]}}. \tag{9}$$

The transmit signals of all the LK relays interfere with each other. The nth user associated with the relay R_{lk}, i.e., the user in the cluster $\mathcal{U}_{lk}$, receives a sum-signal $\hat{y}_{lk,n} = \sum_{l'=1}^{L} \sum_{k'=1}^{K} g_{lk,n}^{l'k'} x_{R_{l'k'}} + z_{lk,n}$, with $z_{lk,n}$ being the AWGN. The nth user in the cluster $\mathcal{U}_{lk}$ uses the estimated CSI $\widehat{g}_{lk,n}^{lk}$ to design an equalizer as $f_{lk,n} = (\widehat{g}_{lk,n}^{lk})^* / |\widehat{g}_{lk,n}^{lk}|$. It then equalizes its received signal as follows:

$$y_{lk,n} = f_{lk,n} \hat{y}_{lk,n} = \sum_{l'=1}^{L} \sum_{k'=1}^{K} f_{lk,n} g_{lk,n}^{l'k'} x_{R_{l'k'}} + f_{lk,n} z_{lk,n} \tag{10}$$

The equalized user signal $y_{lk,n}$ is re-expressed by substituting $x_{R_{l'k'}} = \mu_{l'k'} y_{R_{l'k'}}$ and using (8) as

$$\begin{aligned} y_{lk,n} = {} & \underbrace{f_{lk,n} g_{lk,n}^{lk} \mu_{lk} \mathbf{h}_{lk}^{l^T} \mathbf{w}_{lk} \sqrt{p_{lk,n}} s_{lk,n}}_{\text{desired signal}} + \underbrace{f_{lk,n} g_{lk,n}^{lk} \mu_{lk} \mathbf{h}_{lk}^{l^T} \mathbf{w}_{lk} \sum_{n' \neq n}^{\mathcal{U}_{lk}} \sqrt{p_{lk,n'}} s_{lk,n'}}_{\text{intra-relay interference}} \\ & + \underbrace{\sum_{l'' \neq l}^{L} \sum_{n'=1}^{\mathcal{U}_{lk}} f_{lk,n} g_{lk,n}^{lk} \mu_{lk} \mathbf{h}_{lk}^{l''^T} \mathbf{w}_{l''k} \sqrt{p_{l''k,n'}} s_{l''k,n'}}_{\text{1st hop PS inter-relay interference}} + \underbrace{\sum_{l''=1}^{L} \sum_{k'' \neq k}^{K} \sum_{n'=1}^{\mathcal{U}_{lk}} f_{lk,n} g_{lk,n}^{lk} \mu_{lk} \mathbf{h}_{lk}^{l''^T} \mathbf{w}_{l''k''} \sqrt{p_{l''k'',n'}} s_{l''k'',n'}}_{\text{1st hop nPS inter-relay interference}} \\ & + \underbrace{\sum_{l' \neq l}^{L} f_{k,n} g_{lk,n}^{l'k} \mu_{l'k} \tilde{y}_{R_{l'k}}}_{\substack{\text{2nd hop PS} \\ \text{inter-relay interference}}} + \underbrace{\sum_{l'=1}^{L} \sum_{k' \neq k}^{K} f_{lk,n} g_{lk,n}^{l'k'} \mu_{l'k'} \tilde{y}_{R_{l'k'}}}_{\substack{\text{2nd hop nPS} \\ \text{inter-relay interference}}} + \underbrace{\sum_{l'=1}^{L} \sum_{k'=1}^{K} f_{lk,n} g_{lk,n}^{l'k'} \mu_{l'k'} z_{R_{l'k'}}}_{\text{forwarding noise}} + \underbrace{f_{k,n} z_{k,n}}_{\text{receiver noise}}. \end{aligned} \tag{11}$$

In (11), i) intra-relay interference is caused by the data signals of other users served by the relay R_{lk}; ii) 1st hop pilot-shared (PS)/non pilot-sharing (nPS) inter-relay interference is because the BS uses MR precoding for NOMA signals of multiple relays, and the relay R_{lk} amplifies the NOMA signal of the PS/nPS relays; and ii) 2nd hop PS/nPS inter-relay interference is caused by the amplified transmit signal of the PS/nPS relays that serve their respective user clusters.

In a NOMA system, users served by the kth relay in lth cell mitigate the intra-relay interference by performing SIC. To enable successful SIC, we assume that the users in cluster $\mathcal{U}_{lk}$ are ordered in the descending order of their channel statistics. The nth user associated in the cluster first cancels the intra-relay interference from $\forall\ n' > n$ users by employing SIC [14], [16], and then decodes its own signal while treating the signal from the first $n-1$ users as inherent intra-relay

interference [14], [16]. We observe from (11) that for a user to perform SIC and cancel intra-relay interference, it should also have instantaneous CSI $(\mathbf{h}_{lk}^{l})^T\mathbf{w}_{lk}$ along with $\hat{g}_{lk,n}^{lk}$. It is difficult for a user to have $(\mathbf{h}_{lk}^{l})^T\mathbf{w}_{lk}$. The user, therefore, uses $\mathbb{E}[(\mathbf{h}_{lk}^{l})^T\mathbf{w}_{lk}]$ to perform SIC [16]. We, similar to [14], assume that the users employ statistical value $\mathbb{E}[(\mathbf{h}_{lk}^{l})^T\mathbf{w}_{lk}]$ and its channel estimate $\widehat{g}_{lk,n}^{lk}$ to perform SIC. The intra-relay interference in (11) after performing SIC is given as

$$\underbrace{f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\sum_{n'=1}^{n-1}\sqrt{p_{k,n'}}s_{k,n'}}_{\text{inherent intra-relay interference}}+\underbrace{\sum_{n'=n+1}^{\mathcal{U}_{lk}}\mu_{lk}\left[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}-f_{lk,n}\widehat{g}_{lk,n}^{lk}\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right]\sqrt{p_{k,n'}}s_{k,n'}}_{\text{residual intra-relay interference due to imperfect SIC}}.$$

The first-term is the inherent intra-relay interference, and the second-term is the residual intra-relay interference due to imperfect SIC. The post-SIC receive signal at the nth user associated with the relay R_{lk}, denoted as $\bar{y}_{lk,n}$, can be derived as

$$\bar{y}_{lk,n}=f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\sqrt{p_{lk,n}}s_{lk,n}+f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\sum_{n'=1}^{n-1}\sqrt{p_{lk,n'}}s_{lk,n'} \tag{12}$$

$$+\sum_{n'=n+1}^{\mathcal{U}_{lk}}\mu_{lk}\left[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}-f_{lk,n}\widehat{g}_{lk,n}^{lk}\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right]\sqrt{p_{lk,n'}}s_{lk,n'}$$

$$+\sum_{l''\neq l}^{L}\sum_{n'=1}^{\mathcal{U}_{lk}}f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k}\sqrt{p_{l''k,n'}}s_{l''k,n'}+\sum_{l''=1}^{L}\sum_{k''\neq k}^{K}\sum_{n'=1}^{\mathcal{U}_{lk}}f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k''}\sqrt{p_{l''k'',n'}}s_{l''k'',n'}$$

$$+\sum_{l'\neq l}^{L}f_{k,n}g_{lk,n}^{l'k}\mu_{l'k}\tilde{y}_{R_{l'k}}+\sum_{l'=1}^{L}\sum_{k'\neq k}^{K}f_{lk,n}g_{lk,n}^{l'k'}\mu_{l'k'}\tilde{y}_{R_{l'k'}}+\sum_{l'=1}^{L}\sum_{k'=1}^{K}f_{lk,n}g_{lk,n}^{l'k'}\mu_{l'k'}z_{R_{l'k'}}+f_{k,n}z_{k,n}.$$

III. Achievable SE analysis

We now use *hardening bound* technique to derive a closed-form SE expression for the multi-cell relay-aided mMIMO NOMA system. This technique splits the user receive signal in (12) into signal received over *hardened* channel $\mathbb{E}[\mu_{lk}f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}]$ plus effective noise as [26]:

$$\bar{y}_{lk,n}=\underbrace{\mu_{lk}\mathbb{E}\Big[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\Big]\sqrt{p_{lk,n}}s_{lk,n}}_{\text{true desired signal }(\mathrm{D}_{lk,n})}+\bar{z}_{lk,n},\ \text{where} \tag{13}$$

$$\bar{z}_{lk,n}=\underbrace{\mu_{lk}\left(f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}-\mathbb{E}\Big[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\Big]\right)\sqrt{p_{k,n}}s_{k,n}}_{\text{beamforming uncertainty }(\mathrm{BU}_{lk,n})}+f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\sum_{n'=1}^{n-1}\sqrt{p_{lk,n'}}s_{lk,n'}$$

$$+\sum_{n'=n+1}^{\mathcal{U}_{lk}}\mu_{lk}\Big(f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}-f_{lk,n}\widehat{g}_{lk,n}^{lk}\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\Big)\sqrt{p_{lk,n'}}s_{lk,n'}+$$

$$\sum_{l''\neq l}^{L}\sum_{n'=1}^{\mathcal{U}_{lk}} f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k}\sqrt{p_{l''k,n'}}s_{l''k,n'} + \sum_{l''=1}^{L}\sum_{k''\neq k}^{K}\sum_{n'=1}^{\mathcal{U}_{lk}} f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k''}\sqrt{p_{l''k'',n'}}s_{l''k'',n'}$$

$$+\sum_{l'\neq l}^{L} f_{k,n}g_{lk,n}^{l'k}\mu_{l'k}\tilde{y}_{R_{l'k}} + \sum_{l'=1}^{L}\sum_{k'\neq k}^{K} f_{lk,n}g_{lk,n}^{l'k'}\mu_{l'k'}\tilde{y}_{R_{l'k'}} + \sum_{l'=1}^{L}\sum_{k'=1}^{K} f_{lk,n}g_{lk,n}^{l'k'}\mu_{l'k'}z_{R_{l'k'}} + f_{k,n}z_{k,n}. \quad (14)$$

The term $\mathrm{BU}_{lk,n}$ in (14) is uncorrelated with $\mathrm{D}_{lk,n}$ i.e.,

$$\mathbb{E}[(\mathrm{D}_{lk,n})^{*}\mathrm{BU}_{lk,n}] = \mu_{lk}^2\left\{\left|\mathbb{E}\left[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right|^2 - \left|\mathbb{E}\left[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right|^2\right\} = 0.$$

It is easy to note that the effective noise $\bar{z}_{lk,n}$ is uncorrelated with the true desired signal.We now similar to [8], [26], use central limit theorem to approximate it as the worst case Gaussian noise. The resultant sum SE of the system is given as

$$\overline{R}_{\text{sum}} = \frac{1}{2}\Big(1-\frac{\tau}{\tau_c}\Big)\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_k}\left[\log_2\left(1+\frac{\Delta_{lk,n}}{\Omega_{lk,n}}\right)\right], \text{ where} \quad (15)$$

$\Delta_{lk,n} = \left|\mathbb{E}\left[\Theta_{lk,n}^{lk}\right]\right|^2 p_{lk,n}$, $\Theta_{lk,n}^{lk} = \left[f_{lk,n}g_{lk,n}^{lk}\theta_{lk}^{lk}\right]$, $\theta_{lk}^{l''k''} = \left[\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k''}\right]$ and $\Omega_{lk,n} = \sum_{m=0}^{7}\overline{I}_{lk,n}^{(m)} + 1$, with

$$\overline{I}_{lk,n}^{(0)} = \mathbb{E}\left[\left|\mu_{lk}\left(\Theta_{lk,n}^{lk} - \mathbb{E}\left[\Theta_{lk,n}^{lk}\right]\right)\right|^2\right]p_{lk,n},\ \overline{I}_{lk,n}^{(1)} = \mu_{lk}^2\mathbb{E}\left[\left|\Theta_{lk,n}^{lk}\right|^2\right]\sum_{n'=1}^{n-1}p_{lk,n'},$$

$$\overline{I}_{lk,n}^{(2)} = \sum_{n'=n+1}^{\mathcal{U}_{lk}}\mathbb{E}\left[\left|\Theta_{lk,n}^{lk} - f_{lk,n}\widehat{g}_{lk,n}^{lk}\mathbb{E}\left[\theta_{lk}^{lk}\right]\right|^2\right]\mu_{lk}^2 p_{lk,n'},\ \overline{I}_{lk,n}^{(3)} = \sum_{l''\neq l}^{L}\sum_{n'=1}^{\mathcal{U}_{l''k}}\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\theta_{lk}^{l''k}\right|^2\right]p_{l''k,n'},$$

$$\overline{I}_{lk,n}^{(4)} = \sum_{l''=1}^{L}\sum_{k''\neq k}^{K}\sum_{n'=1}^{\mathcal{U}_{l''k''}}\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\theta_{lk}^{l''k''}\right|^2\right]p_{l''k'',n'},\ \overline{I}_{lk,n}^{(5)} = \sum_{l'\neq l}^{L}\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{l'k}\mu_{l'k}\tilde{y}_{R_{l'k}}\right|^2\right],$$

$$\overline{I}_{lk,n}^{(6)} = \sum_{l'=1}^{L}\sum_{k'\neq k}^{K}\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{l'k'}\mu_{l'k'}\tilde{y}_{R_{l'k'}}\right|^2\right] \quad \text{and} \quad \overline{I}_{lk,n}^{(7)} = \sum_{l'=1}^{L}\sum_{k'=1}^{K}\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{l'k'}\mu_{l'k'}\right|^2\right]. \quad (16)$$

Here $\overline{I}_{lk,n}^{(0)}, \cdots, \overline{I}_{lk,n}^{(7)}$ are the powers of the beamforming uncertainty (BU), inherent intra-relay interference, residual intra-relay interference, 2nd hop PS and nPS inter-relay interferences, 1st hop PS and nPS inter-relay interferences and amplified relay noise, respectively. We next calculate the expectations in (16) to obtain a closed-form SE expression. The inclusion of spatially-correlated Rician channels, pilot-contamination and imperfect CSI and SIC errors significantly complicates the closed-form SE derivation when compared with [14], [16] and requires calculation of moments of Rice distribution [29, Table 1].

Theorem 1. The closed-form SE of the nth user in the cluster $\mathcal{U}_{lk}$ for a finite number of BS antennas relying on MMSE channel estimation, and with imperfect user SIC, is given as

$$\widetilde{R}_{lk,n} = \frac{1}{2}\left(1-\frac{\tau}{\tau_c}\right)\log_2\left(1+\frac{\overline{\Delta}_{lk,n}}{\overline{\Omega}_{lk,n}}\right), \text{ where } \overline{\Omega}_{lk,n} = \sum_{m=0}^{7}\overline{I}_{lk,n}^{(m)} + 1, \quad (17)$$

$$\overline{\Delta}_{lk,n}=A_{lk,n}\tilde{\mu}_{lk}^{2}p_{lk,n},\ \overline{I}_{lk,n}^{(0)}=C_{lk,n}^{(0)}p_{lk,n}\tilde{\mu}_{lk}^{2},\ \overline{I}_{lk,n}^{(1)}=C_{lk,n}^{(1)}\sum_{n'=1}^{n-1}\tilde{\mu}_{lk}^{2}p_{lk,n'},\ \overline{I}_{lk,n}^{(2)}=C_{lk,n}^{(2)}\sum_{n'=n+1}^{\mathcal{U}_{lk}}p_{lk,n'}\tilde{\mu}_{lk}^{2},$$

$$\sum_{m=3}^{6}\overline{I}_{lk,n}^{(m)}=\sum_{(l',k')\neq(l,k)}\sum_{n'=1}^{\mathcal{U}_{l'k'}}C_{l'k',lk,n}^{(3)}p_{l'k',n'}\mu_{l'k'}^{2}+\sum_{(l',k')}\sum_{(l'',k'')\neq(l',k')}\sum_{n'=1}^{\mathcal{U}_{l''k''}}C_{\substack{l''k'',l'k',\\lk,n}}^{(4)}p_{l''k'',n'}\tilde{\mu}_{l'k'}^{2},$$

$$\overline{I}_{k,n}^{(7)}=\sum_{l'=1}^{L}\sum_{k'=1}^{K}C_{l'k',lk,n}^{(5)}\tilde{\mu}_{l'k'}^{2},\ \text{and}\ \tilde{\mu}_{lk}=\sqrt{\frac{q_{lk}}{\left(\sum_{l''=1}^{L}\sum_{k''=1}^{K}\rho_{l''k'',lk}p_{l''k''}+\sum_{l''=1}^{K}\xi_{l''k,lk}p_{l''k}+1\right)}}. \quad (18)$$

Here $p_{lk}=\sum_{n=1}^{\mathcal{U}_{lk}}p_{lk,n}$, $A_{lk,n}=\frac{\pi\upsilon_{lk,n}^{lk}\delta_{lk}}{4}\left[L_{1/2}\left(-\left|\overline{g}_{lk,n}^{lk}\right|^{2}/\upsilon_{lk,n}^{lk}\right)\right]^{2}$, with $L_{\frac{1}{2}}(\cdot)$ being Laguerre polynomial [30]. The terms $C_{lk,n}^{(0)}$, $C_{lk,n}^{(1)}$, $C_{lk,n}^{(2)}$, $C_{l'k',lk,n}^{(3)}$, $C_{\substack{l''k'',l'k',\\lk,n}}^{(4)}$ and $C_{l'k',lk,n}^{(5)}$ are functions of long term channel statistics, which are given in Appendix A.

Corollary 1. We now use the derived lower bound to obtain the asymptotic SE expression for a single-cell system, when the Rician factors $K_{lk}^{l'}=K_{lk,n}^{l'k'}=\widetilde{K}$ and BS antennas N jointly tend to infinity. *This study helps to characterize the impact of high Rician factors and a large number of BS antennas on the performance of mMIMO NOMA system, an aspect which [14] crucially ignored*. We consider the term $\overline{\Delta}_{lk,n}$ in Theorem 1, and as $(\widetilde{K},N)\to\infty$, it simplifies to

$$\overline{\Delta}_{lk,n}=A_{lk,n}\tilde{\mu}_{lk}^{2}\overset{(a)}{=}\frac{\pi\upsilon_{lk,n}^{lk}\delta_{lk}}{4}\frac{4|\bar{g}_{lk,n}^{lk}|^{2}}{\pi\upsilon_{lk,n}^{lk}}\frac{q_{lk}p_{lk,n}}{\xi_{lk,lk}p_{lk}+\sum_{k'=1}^{K}\rho_{lk,lk'}p_{lk'}+1}\overset{(b)}{=}\frac{p_{lk,n}q_{lk}}{p_{lk}}\beta_{lk,n}^{lk}. \quad (19)$$

Equality (a) holds because of the upper limit on Laguerre polynomial $L_{1/2}\left(\frac{-\left|\bar{g}_{lk,n}^{lk}\right|^{2}}{\upsilon_{lk,n}^{lk}}\right)\xrightarrow{\widetilde{K}\to\infty}\frac{|\bar{g}_{lk,n}^{lk}|^{2}}{\upsilon_{lk,n}^{lk}}\frac{4}{\pi}$ [30, 13.5.1]. Equality (b) is obtained by substituting $\bar{\mathbf{h}}_{lk}^{l}=\sqrt{\beta_{lk}^{l}}\mathbf{h}_{lk}^{l,\mathrm{LoS}}$, $\bar{g}_{lk,n}^{lk}=\sqrt{\beta_{lk,n}^{lk}}$, $\overline{\mathbf{R}}_{lk}^{l}=\mathbf{0}$ and use the result $\bar{\mathbf{h}}_{lk}^{lH}\bar{\mathbf{h}}_{lk'}^{l}=N$, if $k'=k$ and ; $=0$, otherwise. The other terms in Theorem 1, on similar lines, can be calculated as $\overline{I}_{k,n}^{(1)}\xrightarrow{\widetilde{K},N\to\infty}\sum_{n'=1}^{n-1}\frac{p_{lk,n'}q_{lk}}{p_{lk}}$, $\sum_{m=3}^{6}\overline{I}_{lk,n}^{(m)}\xrightarrow{\widetilde{K},N\to\infty}\sum_{k'\neq k}^{K}\beta_{lk',n}^{lk'}q_{lk'}$ and $\left\{\overline{I}_{k,n}^{(0)},\overline{I}_{k,n}^{(2)},\cdots,\overline{I}_{k,n}^{(5)},\overline{I}_{k,n}^{(7)}\right\}\xrightarrow{\widetilde{K},N\to\infty}0$. The final SE expression is therefore $\overline{R}_{lk,n}^{\infty}=\frac{1}{2}\left(1-\frac{\tau}{\tau_c}\right)\log_2\left(1+\frac{\frac{p_{lk,n}q_{lk}}{p_{lk}}\beta_{lk,n}^{lk}}{\beta_{lk,n}^{lk}\sum_{n'=1}^{n-1}\frac{q_{lk}p_{lk,n'}}{p_{lk}}+\sum_{k'\neq k}^{K}\beta_{lk',n}^{lk'}q_{lk'}+1}\right)$. We infer that as the Rician factors and BS antennas jointly tend to infinity, the

- beamforming uncertainty power $\overline{I}_{k,n}^{(0)}$, residual intra-relay interference power $\overline{I}_{k,n}^{(2)}$ and the 2nd hop nPS inter-relay interference power $\overline{I}_{k,n}^{(4)}$ vanish. This is because the Rician factor and massive antennas together increase the channel hardening and favourable propagation properties of an mMIMO system [1]. This further improves the SIC capability of the users.
- SE $\overline{R}_{lk,n}^{\infty}$ i) is independent of the BS antennas and the relay-to-BS channel large-scale coeffi-

cients and; ii) depends only the relay-user channel strengths. This suggests that with increase in BS antennas, the SE saturates to a non-zero finite value given in $\overline{R}^{\infty}_{lk,n}$.

Corollary 2. Recall that the existing multi-cell multi-relay mMIMO NOMA literature has not yet derived closed-form SE expression for correlated Rayleigh channels with perfect/imperfect CSI/SIC and Rician channels with perfect CSI/SIC. The derived closed-form SE expression can be reduced to the aforementioned cases by doing the changes provided in Table II.

Table II: Changes to be made in Theorem 1 to analyze Rayleigh/Rician channels with perfect/imperfect CSI.

Channel	SIC/CSI	Changes to be made in Theorem 1
Rayleigh	imperfect	Substitute $\bar{\mathbf{h}}^{j}_{lk} = \mathbf{0}_{N\times 1}$,$\bar{g}^{lk}_{lk,n} = 0$ and $K^{j}_{lk} = K^{lk}_{lk,n} = 0$, $\forall l, k, j, n$ in Theorem 1.
Rayleigh	perfect	Substitute $\bar{\mathbf{h}}^{j}_{lk} = \mathbf{0}_{N\times 1}$,$\bar{g}^{lk}_{lk,n} = 0$, $K^{j}_{lk} = K^{lk}_{lk,n} = 0$, $\forall l, k, j, n$, $\widehat{\mathbf{R}}^{j}_{lk} = \mathbf{R}^{j}_{lk}$, $\upsilon^{lk}_{lk,n} = \gamma^{lk}_{lk,n}$ and $\overline{I}^{(2)}_{k,n} = 0$ in Theorem 1.
Rician	perfect	Substitute $\widehat{\mathbf{R}}^{j}_{lk} = \mathbf{R}^{j}_{lk}$, $\upsilon^{lk}_{lk,n} = \gamma^{lk}_{lk,n}$ and $\overline{I}^{(2)}_{k,n} = 0$ in Theorem 1.

IV. Global energy efficiency definition and optimization

We first define the GEE of the system and then design a low-complexity AMM-based power allocation scheme to optimally allocate the BS and the relay powers to maximize it.

1) GEE definition: The GEE, which is the ratio of the network SE to its power consumption, characterizes the effective number of bits transmitted per unit joule of energy [26]. The GEE expression, for the considered system, can be obtained using Theorem 1 as follows [22]:

$$\text{GEE}\left(p_{lk,n}, q_{lk}\right) = \frac{B\left(\frac{\tau_c-\tau}{2\tau_c}\right)\sum\limits_{l=1}^{L}\sum\limits_{k=1}^{K}\sum\limits_{n=1}^{\mathcal{U}_{lk}}\log_2\left(1+\frac{\overline{\Delta}_{lk,n}(p_{lk,n},q_{lk})}{\overline{\Omega}_{lk,n}(p_{lk,n},q_{lk})}\right)}{P_{tot}(p_{lk,n}, q_{lk})}. \tag{20}$$

Here, B is the bandwidth and $\left(\frac{\tau_c-\tau}{2\tau_c}\right)$ is the overhead due to channel estimation and two-hop data transmission. The term $P_{tot}(p_{lk,n}, q_{lk})$ is the network power consumption, and is given as:

$$P_{tot}(p_{lk,n}, q_{lk}) = \frac{(\tau_c-\tau)}{2\tau_c}\eta_b\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}}p_{lk,n} + \frac{(\tau_c-\tau)}{2\tau_c}\eta_r\sum_{l=1}^{L}\sum_{k=1}^{K}q_{lk} + \frac{\tau}{\tau_c}\eta_r\sum_{l=1}^{L}\sum_{k=1}^{K}p_p + P_c. \tag{21}$$

The scalars η_b and η_r denote the power amplifier (PA) inefficiency at the BSs and relays, respectively [22]. The first and second terms in P_{tot} denote the power consumed by the BSs and the relays in the data transmission phase respectively, while the third term is the power consumed by the relays in CE phase. The term P_c is the circuit power consumed by the entire network, and is modelled as [26]

$$P_c = L\left(P_{fix} + NP_{bs} + KP_{rel} + K\mathcal{U}_{lk}P_{user} + P_{ce} + P_{sig}\right). \tag{22}$$

Here P_{fix} is the fixed circuit power, and the terms P_{bs}, P_{rel} and P_{user} denote the per-antenna circuit power consumption at BS, relay and user, respectively. The terms P_{ce} and P_{sig} model

the power consumed to estimate channel at the BS and users, and the load dependent signal processing at BS, respectively. They are modelled as [26]

$$P_{ce} = \frac{3B}{\tau_c \zeta} K(N\tau + N^2 + K\mathcal{U}_{lk}(\tau + 1)) \text{ and } P_{sig} = \frac{3B}{\zeta} NK(\tau_c - \tau + 1). \tag{23}$$

Here ζ is the computational efficiency in (flops/W) of the BS and users.

A. GEE problem formulation and optimization

The GEE optimization can be cast by ignoring the constant $B\frac{(\tau_c - \tau)}{2\tau_c}$ in (20) as:

$$\mathbf{P1}: \quad \underset{\overline{\mathbf{P}}, \mathbf{Q}}{\text{Maximize}} \ \frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}} \log_2\left(1 + \frac{\overline{\Delta}_{lk,n}(\overline{\mathbf{P}}, \mathbf{Q})}{\overline{\Omega}_{lk,n}(\overline{\mathbf{P}}, \mathbf{Q})}\right)}{P_{tot}(\overline{\mathbf{P}}, \mathbf{Q})} \triangleq f_{\text{GEE}}(\overline{\mathbf{P}}, \mathbf{Q})$$

$$\text{s.t.} \quad \sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}} p_{lk,n} \leq P_T \ \forall l, \ q_{lk} \leq Q_T \ \forall l, \ \forall k \text{ and } p_{lk,n}, q_{lk} \geq 0 \tag{24a}$$

Here the matrix $\overline{\mathbf{P}} = [\mathbf{P}_1, \cdots, \mathbf{P}_L] \in \mathbb{R}^{\mathcal{U}_{lk} \times LK}$ with $\mathbf{P}_l = [\mathbf{p}_{l1}, \cdots, \mathbf{p}_{lK}] \in \mathbb{R}^{\mathcal{U}_{lk} \times K}$ and $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_L] \in \mathbb{R}^{K \times L}$. The first two constraints in (24a) bound the transmit power of the BS and the relays in each cell, respectively.

The GEE optimization is a single ratio fractional program (FP), with the network SE in the numerator and its power consumption in the denominator. The network SE further contains joint functions involving product and/or ratios of the optimization variables, which renders the GEE objective non-convex. The authors in [14] handled this non-convexity by approximating the objective as convex, and then jointly optimized the BS and relays transmit powers. This work, as discussed earlier, considered only single-cell system uncorrelated Rayleigh faded channels, *and thus also has a relatively simpler SE expression*. The GEE optimization has $L(K\mathcal{U}_{lk} + K)$ variables, where K and $\mathcal{U}_{lk}$ are the number of relays and users per cluster. For a large K and $\mathcal{U}_{lk}$, the joint optimization in [14], *will not only have enormous complexity but also require large number of approximations, which will affect the optimization performance*.

To reduce the number of approximations and to solve the GEE optimization with a reduced complexity, we now develop an efficient AMM algorithm, which leverages the benefits of alternating maximization [31] and MM frameworks [25]. It tackles a complex non-convex problem by partitioning it into multiple sub-problems, wherein each sub-problem is optimized over a block of variables using the MM technique [24]. We note that the sub-problems in the current work are still non-convex, and the AMM framework requires construction of convex surrogate functions to handle the non-convexity [25]. We construct a novel surrogate function, and prove that it satisfies the required properties of a valid surrogate function.

We now explain the AMM framework in the following proposition.

Proposition 1. Consider the following maximization problem

$$\mathbf{P}: \quad \underset{\mathbf{x}}{\text{Maximize}} \;\; f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X}, \tag{25}$$

where $f(\cdot)$ is a continuous non-concave function, $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_L$ is a constraint set with $\mathcal{X}_l \in \mathbb{R}^{\frac{N}{L}}$ and $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L]^T \in \mathbb{R}^N$ is the optimization variable that can be partitioned into L blocks such that $\mathbf{x}_l \in \mathcal{X}_l$. The AMM algorithm first splits the original problem into L sub-problems, which are alternatively optimized one at a time, while keeping the other variable blocks fixed. The ith sub-problem is solved over the variable block $\mathbf{x}_i$ as:

$$\mathbf{P}^{(i)}: \quad \underset{\mathbf{x}_i}{\text{Maximize}} \;\; f(\mathbf{x}_i; \mathbf{x}_{-i}) \quad \text{s.t. } \mathbf{x}_i \in \mathcal{X}_i. \tag{26}$$

Here $\mathbf{x}_{-i} = [\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_L]$. The AMM framework, next solves the ith sub-problem $\mathbf{P}^{(i)}$ by generating a series of feasible points $\mathbf{x}_i^{t+1}$ by solving the following problem:

$$\mathbf{P}^{(i)}_{\text{AMM}}: \quad \hat{\mathbf{x}}_i^{t+1} = \underset{\mathbf{x}_i}{\text{argmax}} \;\; \tilde{g}_i(\mathbf{x}_i; \underbrace{\hat{\mathbf{x}}_1^{t+1}, \hat{\mathbf{x}}_2^{t+1}, \cdots, \hat{\mathbf{x}}_{i-1}^{t+1}}_{\hat{\mathbf{x}}_{i-}^{t+1}}, \hat{\mathbf{x}}_i^t, \underbrace{\hat{\mathbf{x}}_{i+1}^t, \cdots, \hat{\mathbf{x}}_L^t}_{\hat{\mathbf{x}}_{i+}^t}) \triangleq \tilde{g}_i(\mathbf{x}_i; \hat{\mathbf{x}}_{i-}^{t+1}, \hat{\mathbf{x}}_i^t, \hat{\mathbf{x}}_{i+}^t)$$
$$\text{s.t. } \mathbf{x}_i \in \mathcal{X}_i. \tag{27}$$

Here $\hat{\mathbf{x}}_i^t$ is the feasible point of the ith problem $\mathbf{P}^{(i)}_{\text{AMM}}$ at tth iteration, and $\tilde{g}_i(\mathbf{x}_i; \hat{\mathbf{x}}_{i-}^{t+1}, \hat{\mathbf{x}}_i^t, \hat{\mathbf{x}}_{i+}^t)$ is the surrogate function in $\mathbf{x}_i$ constructed around the point $\hat{\mathbf{x}}_i^t$. The $\tilde{g}_i$ is a tight lower bound on the original objective $f(\mathbf{x})$ over the variable block $\mathbf{x}_i$, and satisfies the following properties.

$$\text{C1: } \tilde{g}_i(\hat{\mathbf{x}}_i^t; \hat{\mathbf{x}}_i^t, \tilde{\mathbf{x}}_{-i}) = f(\hat{\mathbf{x}}_i^t, \tilde{\mathbf{x}}_{-i}) \text{ and; } \;\; \text{C2: } \nabla_{\mathbf{x}_i} g(\mathbf{x}_i; \hat{\mathbf{x}}_i^t, \tilde{\mathbf{x}}_{-i})|_{\mathbf{x}_i = \hat{\mathbf{x}}_i^t} = \nabla_{\mathbf{x}_i} f(\mathbf{x}_i; \tilde{\mathbf{x}}_{-i})|_{\mathbf{x}_i = \hat{\mathbf{x}}_i^t}. \tag{28}$$

Here $\tilde{\mathbf{x}}_{-i} = [\hat{\mathbf{x}}_{i-}^{t+1}, \hat{\mathbf{x}}_{i+}^t]$. By iteratively constructing the surrogate function $\tilde{g}_i(\mathbf{x}_i; \hat{\mathbf{x}}_{i-}^{t+1}, \hat{\mathbf{x}}_i^t, \hat{\mathbf{x}}_{i+}^t)$ and by solving problem $\mathbf{P}^{(i)}_{\text{AMM}}$ for $1 \leq i \leq L$, the proposed AMM algorithm converges to a stationary point of the original problem $\mathbf{P}$ [24].

We now use the above AMM framework to optimize the non-convex GEE metric in problem $\mathbf{P1}$, which is a joint optimization in $L(K\mathcal{U}_{lk} + K)$ variables. To use the AMM framework, we first separate the optimization variables into $L(K+1)$ blocks as:

$$\overline{\mathbf{P}} = \big[\underbrace{\mathbf{p}_{11}, \mathbf{p}_{12}, \cdots, \mathbf{p}_{1K}}_{1^{\text{st}} \text{ set of } K \text{ blocks}}, \underbrace{\mathbf{p}_{21}, \mathbf{p}_{22}, \cdots, \mathbf{p}_{2K}}_{2^{\text{nd}} \text{ set of } K \text{ blocks}}, \cdots, \underbrace{\mathbf{p}_{L1}, \mathbf{p}_{L2}, \cdots, \mathbf{p}_{LK}}_{L^{\text{th}} \text{ set of } K \text{ blocks}}\big] \text{ and } \mathbf{Q} = \big[\underbrace{\mathbf{q}_1, \cdots, \mathbf{q}_L}_{L \text{ blocks}}\big]. \tag{29}$$

Here $\mathbf{p}_{lk} = [p_{lk,1}, \cdots, p_{lk,\mathcal{U}_{lk}}]^T \in \mathbb{R}^{\mathcal{U}_{lk} \times 1}$ and $\mathbf{q}_l = [q_{l1}, \cdots, q_{lK}] \in \mathbb{R}^{K \times 1}$. Problem $\mathbf{P1}$, using the above block structure, is split into $LK + L$ sub-problems wherein the

- (l,k)th sub-problem in the first LK sub-problems, for $1 \leq k \leq K$ and $1 \leq l \leq L$, is solved

for $\mathbf{p}_{lk}$ by fixing other blocks $\left(\{\{\mathbf{p}_{ji}\}_{j\neq l}^{L}\}_{i\neq k}^{K}, \mathbf{Q}\right)$ as

$$\mathbf{P2}_{lk}: \mathbf{p}_{lk} = \underset{\mathbf{p}_{lk}}{\text{argmax}} \quad f_{\text{GEE}}(\mathbf{p}_{lk}; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle}) \quad \text{s.t.} \quad \mathbf{1}^T\mathbf{p}_{lk} \leq P_T - \sum_{i\neq k}^{K} \mathbf{1}^T\mathbf{p}_{li}^{\blacktriangle} \text{ and } p_{lk,n} \geq 0. \quad (30)$$

Here $\mathbf{1}$ is an all-ones vector of length $\mathcal{U}_{lk}$ and $\overline{\mathbf{P}}_{-lk} = [\mathbf{P}_1, \cdots, \mathbf{P}_{l-1}, \widetilde{\mathbf{P}}_l, \mathbf{P}_{l+1}, \cdots, \mathbf{P}_L] \in \mathbb{R}^{\mathcal{U}_{lk} \times L(K-1)}$, with $\widetilde{\mathbf{P}}_l = [\mathbf{p}_{l1}, \cdots, \mathbf{p}_{l(k-1)}, \mathbf{p}_{l(k+1)}, \cdots, \mathbf{p}_{lK}] \in \mathbb{R}^{\mathcal{U}_{lk} \times K-1}$.

- last L sub-problems are solved over variable blocks $\mathbf{q}_1, \cdots, \mathbf{q}_L$ for a fixed $\overline{\mathbf{P}}$ as

$$\mathbf{P2}_{LK+l}: \mathbf{q}_l = \underset{\mathbf{q}_l}{\text{argmax}} \quad f_{\text{GEE}}(\mathbf{q}_l; \mathbf{Q}_{-l}^{\blacktriangle}, \overline{\mathbf{P}}^{\blacktriangle}) \quad \text{s.t.} \quad 0 \leq q_{lk} \leq Q_T, \forall k. \quad (31)$$

Here $\mathbf{Q}_{-l} = [\mathbf{q}_1, \cdots, \mathbf{q}_{l-1}, \mathbf{q}_{l+1}, \cdots, \mathbf{q}_L] \in \mathbb{R}^{K \times (L-1)}$.

The notation '$\blacktriangle$' in the superscript of variables $(\overline{\mathbf{P}}_{-lk}, \mathbf{Q})$ (resp. $(\mathbf{Q}_{-l}, \overline{\mathbf{P}})$) in $\mathbf{P2}_{lk}$ (resp. $\mathbf{P2}_{LK+l}$) indicates that their values are fixed during optimization. We now cyclically solve each sub-problem using the AMM framework. In each sub-problem, we iterate between the following steps:

▷ construct a surrogate function around a feasible point for the original objective function.

▷ maximize the surrogate function to obtain the next feasible point.

We first design the surrogate function for each of the subproblems in $\mathbf{P2}_i$ for $1 \leq i \leq L(K+1)$, and then explain the AMM-based solution to the GEE optimization.

1) Surrogate function for the first LK sub-problems: The objective function of problem $\mathbf{P2}_{lk}$ is re-expressed in terms of variable block $\mathbf{p}_{lk}$ using (20) as

$$f_{\text{GEE}}(\mathbf{p}_{lk}; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle}) = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}} \log_2\left(1 + \frac{\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})}{\overline{\Omega}_{lk,n}(\mathbf{p}_k; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})}\right)}{\frac{(\tau_c-\tau)}{2\tau_c}\eta_b \mathbf{1}^T\mathbf{p}_{lk} + \overline{P}_{tot}(\overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})}, \quad (32)$$

where $\overline{P}_{tot}(\overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle}) = \frac{(\tau_c-\tau)}{2\tau_c}\eta_b \sum_{(j,i)\neq(l,k)}^{(L,K)} \mathbf{1}^T\mathbf{p}_{ji}^{\blacktriangle} + \frac{(\tau_c-\tau)}{2\tau_c}\eta_r \sum_{l=1}^{L}\sum_{k=1}^{K} q_{lk}^{\blacktriangle} + \frac{\tau}{\tau_c}\eta_r \sum_{l=1}^{L}\sum_{k=1}^{K} p_p + P_c$ is a constant independent of $\mathbf{p}_{lk}$. We note that $f_{\text{GEE}}(\mathbf{p}_{lk}; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})$ is a *single-ratio* with sum SE in the numerator and the power consumption in denominator. The numerator of the *single-ratio*, further, consists of multiple ratio terms $\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})/\overline{\Omega}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})$. To lower bound the objective function $f_{\text{GEE}}(\mathbf{p}_k; \overline{\mathbf{P}}_{-lk}^{\blacktriangle}, \mathbf{Q}^{\blacktriangle})$, we need to find an equivalent lower bound representation for each of these ratios, which we next do in the following proposition.

Proposition 2. Let $A_k(\mathbf{x}): \mathbb{R}^N \to \mathbb{R}_+$ and $B_k(\mathbf{x}): \mathbb{R}^N \to \mathbb{R}_{++}$, denote the set of non-negative and positive functions in $\mathbf{x}$, respectively. For any feasible point $\mathbf{x} = \mathbf{x}^t$, a surrogate function that lower bounds the fraction $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$ can be designed as

$$\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} \geq 2y_k\sqrt{A_k(\mathbf{x})} - y_k^2 B_k(\mathbf{x}) \triangleq h_k^{lb}(\mathbf{x}, y_k) \text{ where } y_k = \frac{\sqrt{A_k(\mathbf{x}^t)}}{B_k(\mathbf{x}^t)}. \quad (33)$$

The function $h_k^{lb}(\mathbf{x}, y_k)$ satisfies the properties C1 and C2 and is therefore a valid surrogate function.

Proof. Refer Appendix B. □

We now use propostion 2 twice. First to decouple the single-ratio, and second to decouple multiple fractions in the numerator of the single-ratio. The equivalent lower bound on $f_{\text{GEE}}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})$, for a feasible point $\mathbf{p}_{lk} = \mathbf{p}^t_{lk}$, is given as:

$$\begin{aligned} f_{\text{GEE}}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle}) &\geq 2u\sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}} \log_2\left(1 + \frac{\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})}{\overline{\Omega}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})}\right)} \\ &\quad - u^2\left[\frac{(\tau_c - \tau)}{\tau_c}\eta_b \mathbf{1}^T \mathbf{p}_{lk} + \overline{P}_{tot}(\overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})\right] \\ &\geq 2u\sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}} \log_2(1 + 2t_{lk,n}\sqrt{\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})} - t^2_{lk,n}\overline{\Omega}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle}))} \\ &\quad - u^2\left[\frac{(\tau_c - \tau)}{2\tau_c}\eta_b \mathbf{1}^T \mathbf{p}_{lk} + \overline{P}_{tot}(\overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})\right] \\ &\triangleq \tilde{g}_{lk}(\mathbf{p}_{lk}; \mathbf{p}^t_{lk}, \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle}). \end{aligned} \tag{34}$$

The scalars u and $t_{lk,n}$ are functions of $\mathbf{p}^t_k$, and are calculated using Proposition 2 as

$$u = \frac{\sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}} \log_2\left(1 + \frac{\overline{\Delta}_{lk,n}(\mathbf{p}^t_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})}{\overline{\Omega}_{lk,n}(\mathbf{p}^t_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})}\right)}}{\frac{(\tau_c - \tau)}{2\tau_c}\eta_b \mathbf{1}^T \mathbf{p}^t_{lk} + \overline{P}_{tot}(\overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})} \text{ and } t_{lk,n} = \frac{\sqrt{\overline{\Delta}_{lk,n}(\mathbf{p}^t_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})}}{\overline{\Omega}_{lk,n}(\mathbf{p}^t_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})}. \tag{35}$$

The function $\tilde{g}_{lk}(\mathbf{p}_{lk}; \mathbf{p}^t_{lk}, \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})$ is concave in $\mathbf{p}_{lk}$ provided its constituent $\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})$ and $\overline{\Omega}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})$ are concave and convex in $\mathbf{p}_{lk}$, respectively. We now investigate their concavity/convexity.

- The term $\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})$ is rewritten in terms of $\mathbf{p}_{lk}$ as

$$\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle}) = \tilde{A}_{lk,n}\frac{p_{lk,n}}{(\xi_{lk,lk} + \rho_{lk,lk})\mathbf{1}^T \mathbf{p}_{lk} + \varepsilon_{lk}}. \tag{36}$$

Here $\tilde{A}_{lk,n} = A_{lk,n}q^{\blacktriangle}_{lk}$ and $\varepsilon_{lk} = \sum_{(l'',k'')\neq(l,k)} \rho_{l''k'',lk}\mathbf{1}^T \mathbf{p}^{\blacktriangle}_{l''k''} + \sum_{l''\neq l}^{L} \xi_{l''k,lk}\mathbf{1}^T \mathbf{p}^{\blacktriangle}_{l''k} + 1$ are constants, which are independent of the block $\mathbf{p}_{lk}$. The term $\overline{\Delta}_{lk,n}(\mathbf{p}_{lk}; \overline{\mathbf{P}}^{\blacktriangle}_{-lk}, \mathbf{Q}^{\blacktriangle})$ contains non-concave fractional term $\frac{p_{lk,n}}{(\xi_{lk,lk}+\rho_{lk,lk})\mathbf{1}^T \mathbf{p}_{lk}+\varepsilon_{lk}}$, for which we need to find a concave approximation. We now use (33) in Proposition 2 to obtain a tight lower bound on the fractional term as:

$$\frac{p_{lk,n}}{(\xi_{lk,lk}+\rho_{lk,lk})\mathbf{1}^T \mathbf{p}_{lk}+\varepsilon_{lk}} \geq 2a_{lk,n}\sqrt{p_{lk,n}} - a^2_{lk,n}\Big((\xi_{lk,lk} + \rho_{lk,lk})\mathbf{1}^T \mathbf{p}_{lk} + \varepsilon_{lk}\Big) \triangleq \overline{\Phi}^{(1)}_{lk,n}(\mathbf{p}_{lk}; \mathbf{p}^t_{lk}). \tag{37}$$

The variable $a_{lk,n}$, for a fixed $\mathbf{p}_{lk} = \mathbf{p}^t_{lk}$, is given using Proposition 2 as

$$a_{lk,n} = \frac{\sqrt{p^t_{lk,n}}}{(\xi_{lk,lk} + \rho_{lk,lk})\mathbf{1}^T \mathbf{p}^t_{lk} + \varepsilon_{lk}}. \tag{38}$$

- We now give the expression of $\overline{\Omega}_{lk,n}(\mathbf{p}_{lk};\overline{\mathbf{P}}^{\blacktriangle}_{-lk},\mathbf{Q}^{\blacktriangle})$ in terms of $\mathbf{p}_{lk}$ as

$$\overline{\Omega}_{lk,n}(\mathbf{p}_{lk};\overline{\mathbf{P}}^{\blacktriangle}_{-lk},\mathbf{Q}^{\blacktriangle}) = \widetilde{C}^{(0)}_{lk,n}\Phi^{(2)}_{lk,lk,n}(\mathbf{p}_{lk}) + \sum_{n'=1}^{n-1}\widetilde{C}^{(1)}_{lk,n}\Phi^{(2)}_{lk,lk,n'}(\mathbf{p}_{lk}) + \sum_{n'=n+1}^{\mathcal{U}_{lk}}\widetilde{C}^{(2)}_{lk,n}\Phi^{(2)}_{lk,lk,n'}(\mathbf{p}_{lk})$$
$$+ \sum_{(l',k')\neq(l,k)}\sum_{n'=1}^{\mathcal{U}_{l'k'}}\widetilde{C}^{(3)}_{l'k',lk,n}\Phi^{(2)}_{l'k',l'k',n'}(\mathbf{p}_{lk}) + \sum_{(l',k')}\sum_{(l'',k'')\neq(l',k')}\sum_{n'=1}^{\mathcal{U}_{l''k''}}\widetilde{C}^{(4)}_{\substack{l''k'',l'k'\\lk,n}}\Phi^{(2)}_{l'k',l''k'',n'}(\mathbf{p}_{lk})$$
$$+ \sum_{l'=1}^{L}\sum_{k'=1}^{K}\widetilde{C}^{(5)}_{l'k',lk,n}\frac{1}{\xi_{l'k',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l'k'} + \rho_{l'k',lk}\mathbf{1}^T\mathbf{p}_{lk} + \varepsilon_{l'k'}} + 1, \tag{39}$$

where $\Phi^{(2)}_{l'k',lk,n'}(\mathbf{p}_{lk}) = \dfrac{p_{lk,n'}}{\xi_{l'k',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l'k'}+\rho_{lk,l'k'}\mathbf{1}^T\mathbf{p}_{lk}+\varepsilon_{l'k'}}$. The constants $\widetilde{C}^{(i)}_{lk,n} = C^{(i)}_{lk,n}q^{\blacktriangle}_{k}$, for $i = 0,1,2$ and $\widetilde{C}^{(3)}_{l'k',lk,n} = C^{(3)}_{l'k',lk,n}q^{\blacktriangle}_{l'k'}$, $\widetilde{C}^{(4)}_{\substack{l''k'',l'k'\\lk,n}} = C^{(4)}_{\substack{l''k'',l'k'\\lk,n}}q^{\blacktriangle}_{l'k'}$ and $\widetilde{C}^{(5)}_{l'k',lk,n} = C^{(5)}_{l'k',lk,n}q^{\blacktriangle}_{l'k'}$, where $C^{(0)}_{lk,n},\cdots,C^{(5)}_{l'k',lk,n}$ are defined in Appendix A. The function $\overline{\Omega}_{lk,n}(\mathbf{p}_{lk};\overline{\mathbf{P}}^{\blacktriangle}_{-lk},\mathbf{Q}^{\blacktriangle})$, due to the fractional terms $\Phi^{(2)}_{l'k',lk,n'}$, $\forall l,k,l',k',n'$ in the first five terms, is non-convex in $\mathbf{p}_{lk}$. To make it convex, we need to derive a surrogate function that upper bounds the fractional terms, which we do in the following proposition. Its proof is relegated to Appendix C.

Proposition 3. Let $A_k(\mathbf{x}):\mathbb{R}^N\to\mathbb{R}_+$ and $B_k(\mathbf{x}):\mathbb{R}^N\to\mathbb{R}_{++}$, denote the set of non-negative and positive functions, which are convex in $\mathbf{x}$, respectively. For any feasible point $\mathbf{x}=\mathbf{x}^t$, an upper bound on the fraction $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$ can be designed as

$$\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} \leq \frac{1}{2}z_k\big(A_k(\mathbf{x})\big)^2 + \frac{1}{2z_k}\big(B_k(\mathbf{x})\big)^{-2} \triangleq h^{ub}_k(\mathbf{x},z_k), \text{ where } z_k = \frac{1}{A_k(\mathbf{x}^t)B_k(\mathbf{x}^t)}. \tag{40}$$

The function $h^{ub}_k(\mathbf{x},z_k)$, similar to the lower bound in (33), satisfies the properties C1 and C2, and is therefore a valid surrogate function.

Using this result, the fractional term $\Phi^{(2)}_{l'k',lk,n'}(\mathbf{p}_{lk})$ at a feasible point $\mathbf{p}_{lk}=\mathbf{p}^t_{lk}$, is upper bounded as

$$\Phi^{(2)}_{l'k',lk,n'}(\mathbf{p}_{lk}) \leq \frac{b_{l'k',lk,n'}}{2}p^2_{lk,n'} + \frac{1}{2b_{l'k',lk,n'}}\left(\xi_{l'k',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l'k'}+\rho_{lk,l'k'}\mathbf{1}^T\mathbf{p}_{lk}+\varepsilon_{l'k'}\right)^{-2} \triangleq \overline{\Phi}^{(2)}_{l'k',lk,n'}(\mathbf{p}_{lk};\mathbf{p}^t_{lk}), \tag{41}$$

where $b_{l'k',lk,n'}$ is calculated from the fixed $\mathbf{p}_{lk}=\mathbf{p}^t_{lk}$ using Proposition 3 as

$$b_{l'k',lk,n'} = (p^t_{lk,n'})^{-1}\left(\xi_{l'k',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l'k'}+\rho_{lk,l'k'}\mathbf{1}^T\mathbf{p}^t_{lk}+\varepsilon_{l'k'}\right)^{-1}. \tag{42}$$

We now replace the non-concave terms in $\overline{\Delta}_{lk,n}(\mathbf{p}_{lk};\overline{\mathbf{P}}^{\blacktriangle}_{-lk},\mathbf{Q}^{\blacktriangle})$ with their concave lower bound expression in (37). We similarly replace the non-convex terms in $\overline{\Omega}_{lk,n}(\mathbf{p}_{lk};\overline{\mathbf{P}}^{\blacktriangle}_{-lk},\mathbf{Q}^{\blacktriangle})$ with their convex upper bound expression in (41).

The final expression of the surrogate function at the feasible point $\mathbf{p}^t_k$, is therefore obtained

by using (36), (39) and (34) as:

$$\breve{g}_{lk}(\mathbf{p}_{lk};\mathbf{p}_{lk}^t,\overline{\mathbf{P}}_{-lk}^{\blacktriangle},\mathbf{Q}^{\blacktriangle})$$
$$=2u\sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}}\log_2(1+2t_{lk,n}\sqrt{\widetilde{\Delta}_{lk,n}(\mathbf{p}_{lk};\mathbf{p}_{lk}^t,\overline{\mathbf{P}}_{-lk}^{\blacktriangle},\mathbf{Q}^{\blacktriangle})}-t_{lk,n}^2\widetilde{\Omega}_{lk,n}(\mathbf{p}_{lk};\mathbf{p}_{lk}^t,\overline{\mathbf{P}}_{-lk}^{\blacktriangle},\mathbf{Q}^{\blacktriangle}))}$$
$$-u^2\left[\frac{(\tau_c-\tau)}{2\tau_c}\eta_b\mathbf{1}^T\mathbf{p}_{lk}+\overline{P}_{tot}(\overline{\mathbf{P}}_{-lk}^{\blacktriangle},\mathbf{Q}^{\blacktriangle})\right]. \tag{43}$$

Here $\widetilde{\Delta}_{lk,n}\big(\mathbf{p}_{lk};\mathbf{p}_{lk}^t,\overline{\mathbf{P}}_{-lk}^{\blacktriangle},\mathbf{Q}^{\blacktriangle}\big)=\widetilde{A}_{lk,n}\overline{\Phi}_{lk,n}^{(1)}(\mathbf{p}_{lk};\mathbf{p}_{lk}^t)$ and $\widetilde{\Omega}_{lk,n}(\mathbf{p}_{lk};\mathbf{p}_{lk}^t,\overline{\mathbf{P}}_{-lk}^{\blacktriangle},\mathbf{Q}^{\blacktriangle})$ is obtained by replacing the non-convex terms $\Phi_{l'k',lk,n'}^{(2)}(\mathbf{p}_{lk})$ with the convex upper bound $\overline{\Phi}_{l'k',lk,n'}^{(2)}(\mathbf{p}_{lk};\mathbf{p}_{lk}^t)$.

2) Surrogate function for the last L sub-problems: We now construct surrogate function for the objective of sub-problem $\mathbf{P3}_{LK+l}$ for $1\leq l\leq L$. The function $f_{\text{GEE}}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})$ is given using (20) as

$$f_{\text{GEE}}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})=\frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}}\log_2\left(1+\frac{\overline{\Delta}_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle}\overline{\mathbf{P}}^{\blacktriangle})}{\overline{\Omega}_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}\right)}{\frac{(\tau_c-\tau)}{2\tau_c}\eta_r\mathbf{1}^T\mathbf{q}_l+\widehat{P}_{tot}(\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}, \tag{44}$$

where the constant $\widehat{P}_{tot}(\mathbf{Q}_{-l},\overline{\mathbf{P}}^{\blacktriangle})=\frac{(\tau_c-\tau)}{2\tau_c}\eta_r\sum_{j\neq l}^{L}\mathbf{1}^T\mathbf{q}_j^{\blacktriangle}+\frac{(\tau_c-\tau)}{2\tau_c}\eta_b\sum_{l=1}^{L}\sum_{k=1}^{K}\mathbf{1}^T\mathbf{p}_{lk}^{\blacktriangle}+\frac{\tau}{\tau_c}\eta_r\sum_{l=1}^{L}\sum_{k=1}^{K}p_p+P_c$. The function $f_{\text{GEE}}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})$, similar to the objective function of problem $\mathbf{P3}_{lk}$, is a *single-ratio*, which has multiple fractional terms $\frac{\overline{\Delta}_{k,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}{\overline{\Omega}_{k,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}$ in the numerator. Similar to the first LK sub-problems, a lower bound on $f_{\text{GEE}}(\mathbf{q}_l;\mathbf{Q}_{-l},\overline{\mathbf{P}}^{\blacktriangle})$ is obtained using (33) in Proposition 2 as

$$f_{\text{GEE}}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})\geq\breve{g}_{LK+l}(\mathbf{q}_l;\mathbf{q}_l^t,\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle}), \tag{45}$$

where

$$\breve{g}_{LK+l}(\mathbf{q}_l;\mathbf{q}_l^t,\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})$$
$$=2v\sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}}\log_2(1+2w_{lk,n}\sqrt{\overline{\Delta}_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}-w_{lk,n}^2\overline{\Omega}_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle}))}$$
$$-v^2\left[\frac{(\tau_c-\tau)}{2\tau_c}\eta_r\mathbf{1}^T\mathbf{q}_l+\widehat{P}_{tot}(\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})\right]. \tag{46}$$

Here the scalars v and $w_{lk,n}$, for a fixed $\mathbf{q}_l=\mathbf{q}_l^t$ are calculated as

$$v=\frac{\sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{\mathcal{U}_{lk}}\log_2\left(1+\frac{\overline{\Delta}_{lk,n}(\mathbf{q}_l^t;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}{\overline{\Omega}_{lk,n}(\mathbf{q}_l^t;\mathbf{Q}_{-l}^{\blacktriangle}\overline{\mathbf{P}}^{\blacktriangle})}\right)}}{\frac{(\tau_c-\tau)}{2\tau_c}\eta_r\mathbf{1}^T\mathbf{q}_l+\widehat{P}_{tot}(\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}\text{ and }w_{lk,n}=\frac{\sqrt{\overline{\Delta}_{lk,n}(\mathbf{q}_l^t;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}}{\overline{\Omega}_{lk,n}(\mathbf{q}_l^t;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})}. \tag{47}$$

The function $\breve{g}_{LK+l}(\mathbf{q}_l;\mathbf{q}_l^t,\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})$ is concave in $\mathbf{q}_l$, if the terms $\Delta_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l}^{\blacktriangle},\overline{\mathbf{P}}^{\blacktriangle})$ and

$\Omega_{lk,n}(\mathbf{q}_l;\mathbf{Q}^{\blacktriangle}_{-l},\overline{\mathbf{P}}^{\blacktriangle})$ are concave and convex in $\mathbf{q}_l$, respectively. These two functions are given as

$$\overline{\Delta}_{lk,n}(\mathbf{q}_l;\mathbf{Q}^{\blacktriangle}_{-l},\overline{\mathbf{P}}^{\blacktriangle}) = A_{lk,n}\tilde{\Phi}^{(2)}_{lk,lk,n}q_{lk} \text{ and}$$

$$\overline{\Omega}_{lk,n}(\mathbf{q}_l;\mathbf{Q}^{\blacktriangle}_{-l},\overline{\mathbf{P}}^{\blacktriangle}) = C^{(0)}_{lk,n}\tilde{\Phi}^{(2)}_{lk,lk,n}q_{lk}+\sum_{n'=1}^{n-1}C^{(1)}_{lk,n}\tilde{\Phi}^{(2)}_{lk,lk,n'}q_{lk}+\sum_{n'=n+1}^{\mathcal{U}_{lk}}C^{(2)}_{lk,n}\tilde{\Phi}^{(2)}_{lk,lk,n'}q_{lk}$$

$$+\sum_{(l',k')\neq(l,k)}\sum_{n'=1}^{\mathcal{U}_{l'k'}}C^{(3)}_{lk',lk,n}\tilde{\Phi}^{(2)}_{l'k',l'k',n'}q^{\blacktriangle}_{l'k'}+\sum_{(l',k')}\sum_{(l'',k'')\neq(l',k')}\sum_{n'=1}^{\mathcal{U}_{l''k''}}C^{(4)}_{\substack{l''k'',l'k'\\lk,n}}\tilde{\Phi}^{(2)}_{l'k',l''k'',n'}q_{l'k'}$$

$$+\sum_{l'=1}^{L}\sum_{k'=1}^{K}C^{(5)}_{lk,n}\frac{q_{l'k'}}{\sum\limits_{l''=1}^{L}\sum\limits_{k''=1}^{K}\rho_{l''k'',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l''k''}+\sum\limits_{l''=1}^{K}\xi_{l''k',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l''k'}+1}+1. \tag{48}$$

Here $\tilde{\Phi}^{(2)}_{l'k',lk,n'}=\dfrac{p^{\blacktriangle}_{lk,n'}}{\xi_{l'k',l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{l'k'}+\rho_{lk,l'k'}\mathbf{1}^T\mathbf{p}^{\blacktriangle}_{lk}+\varepsilon_{l'k'}}$ is a positive constant, which is independent of the variable block $\mathbf{q}_l$. We infer from (48), that both $\overline{\Delta}_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l},\overline{\mathbf{P}}^{\blacktriangle})$ and $\overline{\Omega}_{lk,n}(\mathbf{q}_l;\mathbf{Q}_{-l},\overline{\mathbf{P}}^{\blacktriangle})$ are affine functions in $\mathbf{q}_l$. The surrogate function $\breve{g}_{LK+l}(\mathbf{q}_l;\mathbf{q}^t_l,\mathbf{Q}^{\blacktriangle}_{-l},\overline{\mathbf{P}}^{\blacktriangle})$, is therefore concave in $\mathbf{q}_l$.

3) AMM-based GEE optimization: The sub-problems $\mathbf{P2}_{lk}$ and $\mathbf{P2}_{LK+l}$ for $1\leq l\leq L$ and $1\leq k\leq K$, using Proposition 1 and the surrogates designed in (43) and (46), are recast as:

$$\mathbf{P3}_{lk}:\mathbf{p}^{t+1}_{lk}=\underset{\mathbf{p}_{lk}}{\text{argmax}}\quad \breve{g}_{lk}(\mathbf{p}_{lk};\mathbf{p}^t_{lk},\overline{\mathbf{P}}^{(t+1,t)}_{-lk},\mathbf{Q}^t)$$

$$\text{s.t. } \mathbf{1}^T\mathbf{p}_{lk}\leq P_T-\sum_{i=1}^{k-1}\mathbf{1}^T\mathbf{p}^{t+1}_{li}-\sum_{i=k+1}^{K}\mathbf{1}^T\mathbf{p}^t_{li} \text{ and } p_{lk,n}\geq 0. \tag{49}$$

and

$$\mathbf{P3}_{LK+l}:\mathbf{q}^{t+1}_l=\underset{\mathbf{q}}{\text{argmax}}\quad \breve{g}_{LK+l}(\mathbf{q}_l;\mathbf{q}^t_l,\mathbf{Q}^{(t+1,t)}_{-l}\overline{\mathbf{P}}^{t+1})\quad \text{s.t.}\quad 0\leq q_{lk}\leq Q_T,\forall k. \tag{50}$$

Here $\overline{\mathbf{P}}^{(t+1,t)}_{-lk}=[\mathbf{P}^{t+1}_1,\cdots,\mathbf{P}^{t+1}_{l-1},\widetilde{\mathbf{P}}^{(t+1,t)}_l,\mathbf{P}^t_{l+1},\cdots,\mathbf{P}^t_L]$ and $\mathbf{Q}^{(t+1,t)}_{-l}=[\mathbf{q}^{t+1}_1,\cdots,\mathbf{q}^{t+1}_{l-1},\mathbf{q}^t_{l+1},\cdots,\mathbf{q}^t_L]$, with $\widetilde{\mathbf{P}}^{(t+1,t)}_l=[\mathbf{p}^{t+1}_{l1},\cdots,\mathbf{p}^{t+1}_{l(k-1)},\mathbf{p}^t_{l(k+1)},\cdots,\mathbf{p}^t_{lK}]$. We now cyclically solve each sub-problem over a variable block while fixing the remaining variables. In each sub-problem, we first construct a surrogate function over the variable block at some feasible point, and then maximize the surrogate function to obtain next feasible point. We repeat this process until the it converges to a stationary point. The GEE optimization is summarized in Algorithm 1.

Convergence of AMM: The iterative solution in Algorithm 1 solves a series of MM problems to optimize the non-convex GEE metric. We also see that the derived surrogate functions in (43) and (46) satisfy properties C1 and C2, and therefore tightly lower bound the GEE objective. Due to these properties, the GEE objective monotonically decreases with number of iterations, and ultimately converges to a stationary point of the original non-convex GEE problem, which is not necessarily optimal [24], [25]. Its proof can be established on lines similar to [24], [25].

Algorithm 1: AMM-based GEE optimization algorithm

Input: Given a tolerance $\epsilon > 0$, the maximum number of iterations I, maximum power constraint at BS P_T and at relay Q_T. The initial transmit power is chosen as $p_{lk,n}^{(0)} = P_T/(K\mathcal{U}_{lk})$ and $q_{lk}^{(0)} = Q_T/2$.

Output: $\overline{\mathbf{P}}^*$, $\mathbf{Q}^*$.

for $t \leftarrow 0$ ***to*** I **do**
  for $l \leftarrow 1$ ***to*** L **do**
    for $k \leftarrow 1$ ***to*** K **do**
      For the feasible point $(\mathbf{p}_{lk}^t, \overline{\mathbf{P}}_{-lk}^{(t+1,t)}, \mathbf{Q}^t)$, construct surrogate $\tilde{g}_{lk}(\mathbf{p}_{lk}; \mathbf{p}_{lk}^t, \overline{\mathbf{P}}_{-lk}^{(t+1,t)}, \mathbf{Q}^t)$, with auxiliary variables $\{a_{lk,n}, b_{l'k',lk,n'}, t_{lk,n}, u\}$ as in (43).
      Solve $\mathbf{P3}_{lk}$ to obtain $\mathbf{p}_{lk}^{(t+1)}$, for fixed optimization variables from blocks other than $\mathbf{p}_{lk}$.
    end
    For feasible point $(\mathbf{q}_l^t, \mathbf{Q}_{-l}^{(t+1,t)}, \overline{\mathbf{P}}^{(t+1)})$, construct surrogate $\tilde{g}_{LK+l}(\mathbf{q}_l; \mathbf{q}_l^t, \mathbf{Q}_{-l}^{(t+1,t)}, \overline{\mathbf{P}}^{t+1})$ as in (46).
    Solve problem $\mathbf{P3}_{LK+l}$ to obtain $\mathbf{q}_l^{(t+1)}$, by fixing auxilary and optimization variables other than $\mathbf{q}_l$.
  end
  if $\|\overline{\mathbf{P}}^{(t+1)} - \overline{\mathbf{P}}^{(t)}\| < \epsilon$ & $\|\mathbf{Q}^{(t+1)} - \mathbf{Q}^{(t)}\| < \epsilon$ **then**
    $\overline{\mathbf{P}}^* = \overline{\mathbf{P}}^{(t+1)}$, $\mathbf{Q}^* = \mathbf{Q}^{(t+1)}$ and break.
  else
    Set $(\overline{\mathbf{P}}^{t+1}, \mathbf{Q}^{(t+1)}) \rightarrow (\overline{\mathbf{P}}^{(t)}, \mathbf{Q}^{(t)})$ and repeat steps 2 - 8.
  end
end
return $\overline{\mathbf{P}}^*$ *and* $\mathbf{Q}^*$.

Computational Complexity of Algorithm 1 for GEE: Problem $\mathbf{P3}$, solved in each iteration of Algorithm 1, for which the complexity with n variables is $\mathcal{O}(n^4)$ [32]. The per-iteration cost of Algorithm 1 is thus $\mathcal{O}\Big(\sum_{l=1}^{L}\sum_{k=1}^{K}\mathcal{U}_{lk}^4 + LK^4\Big)$. *This is lesser than the joint optimization approach in [14] which has a per-iteration complexity of* $\mathcal{O}\Big(\Big(\sum_{l=1}^{L}\sum_{k=1}^{K}\mathcal{U}_{lk} + LK\Big)^4\Big)$.

V. Simulation results

We now numerically corroborate the accuracy of the derived closed-form SE expression and assess the impact of spatial correlation and Rician channels on the SE and GEE of multi-cell multiple-relay-aided mMIMO NOMA system with imperfect SIC. For this study, we consider a multi-relay mMIMO system with $L = 4$ cell, which is deployed in a coverage area of 1200m^2. Each cell consists of an mMIMO BS located at its center, and $K = 5$ relays, deployed uniformly at a distance of $R_r = 200\text{m}$ from the BS. Users are randomly located around the relay within a circular cluster of radius $R_u = 100\text{m}$. Each mMIMO BS is equipped with $N = 100$ antennas, and each relay serves $\mathcal{U}_{lk} = 4$ users. The Rician factors and large-scale fading coefficients of different channels in the system are modelled as follows:

- *Large-scale fading coefficients:* The large scale fading coefficients of channels from jth BS to relay R_{lk} in lth cell and from relay $R_{jk'}$ to nth user in cluster $\mathcal{U}_{lk}$ are modelled as $\beta_{lk}^{j}[\text{dB}] = -30.18 - 26\log_{10}(d_{lk}^{j}) + F_{lk}^{j}$ and $\beta_{lk,n}^{jk'}[\text{dB}] = -30.18 - 26\log_{10}(d_{lk,n}^{jk'}) + F_{lk,n}^{jk'}$. Here d_{lk}^{j} (resp. $d_{lk,n}^{jk'}$) denotes the distance between the relay R_{lk} and the jth BS (resp. nth user in the cluster $\mathcal{U}_{lk}$ and the relay $R_{jk'}$) [3]. The scalars F_{lk}^{j} and $F_{lk,n}^{jk'}$ capture the log-normal shadow fading and are modelled as $\left\{F_{lk}^{j}, F_{lk,n}^{jk'}\right\} \sim \mathcal{CN}(0, \sigma_{\text{sf}}^2)$, with $\sigma_{\text{sf}} = 4$ dB.
- *Rician factors*: The Rician K-factors of the channels $\mathbf{h}_{lk}^{j}$ and $g_{lk,n}^{jk'}$ i.e., K_{lk}^{j} and $K_{lk,n}^{jk'}$, are modelled as $K_{lk}^{j}[\text{dB}] = 13 - 0.03d_{lk}^{j}$ and $K_{lk,n}^{jk'}[\text{dB}] = 13 - 0.03d_{lk,n}^{jk'}$, respectively.
- *Other parameters:* We similar to [26], assume ULA at the BS and model the correlation matrices $\mathbf{R}_{lk}^{j}$ using the Gaussian local scattering model, with angular standard deviation (ASD) 10° [26]. We set the circuit power consumption parameters as $P_{fix} = 5$ W, $P_{bs} = 0.2$ W, $P_{rel} = P_{user} = 0.1$ W and $\zeta = 750$ Gflops/W.

We assume a bandwidth of $B = 20$ MHz, and the relay pilot power of $p_p = 30$ dBm. The coherence interval has $\tau_c = 200$ symbols with a training period of $\tau = K = 5$ symbols. The system parameters remains fixed unless explicitly mentioned. The sum SE and GEE results, to remove the randomness due to user locations, are averaged over 100s of random user deployments.

*1) **Validation of closed-form SE expression**:* We plot in Fig. 2a, the sum SE per cell versus the BS transmit power P_T, and compare the closed-form SE expression in Theorem 1 with its simulated counterpart in (15). For this study, we consider following (relay, user/relay) combinations $(K, \mathcal{U}_{lk}) = \{(2, 10), (4, 5), (5, 4), (10, 2)\}$. We also allocate equal power to all the users and relays i.e., $p_{lk,n} = P_T/(K\mathcal{U}_{lk})$ and $q_{lk} = Q_T \ \forall \ k = 1, \cdots, K, \ l = 1, \cdots, L$ and $n = 1, \cdots, \mathcal{U}_{lk}$. For different K values, the closed-form and simulated curves exactly overlap, which validates the correctness of the former. With increase in K, the sum SE per cell initially increases till $K = 5$, and then decreases for $K = 10$. This is explained as follows:

– For $K = 2$ relays, the relays are far apart and each relay serves $\mathcal{U}_{lk} = 10$ users. The users served by each relay therefore experience stronger intra-relay interference (residual + inherent), which dominates over the inter-relay interference (1st hop + 2nd hop).
– With increase in K from $K = 2$ to $K = 5$ the i) intra-relay interference reduces, as each relay now serves less number of users and; ii) inter-relay interference increases, as the separation distance between the relays decreases. For $K = 5$, both inter-relay and intra-relay interferences are balanced and also minimum, which increases the sum SE.
– For $K \geq 5$, the separation distance between the relays is the least, and the inter-relay

interference dominates over intra-relay, which reduces the sum SE values.

We further note that with increase in the BS transmit power P_T, the sum SE increases and then saturates for $P_T \geq 10$ dBm. This is because the users experience high inter-relay and intra-relay interferences at high BS transmit power. *This study reveals the optimal number of relays and relay-user ratio required to achieve the optimal sum SE. This aspect can be used by the designer to appropriately decide the number of relays and the users per relay.*

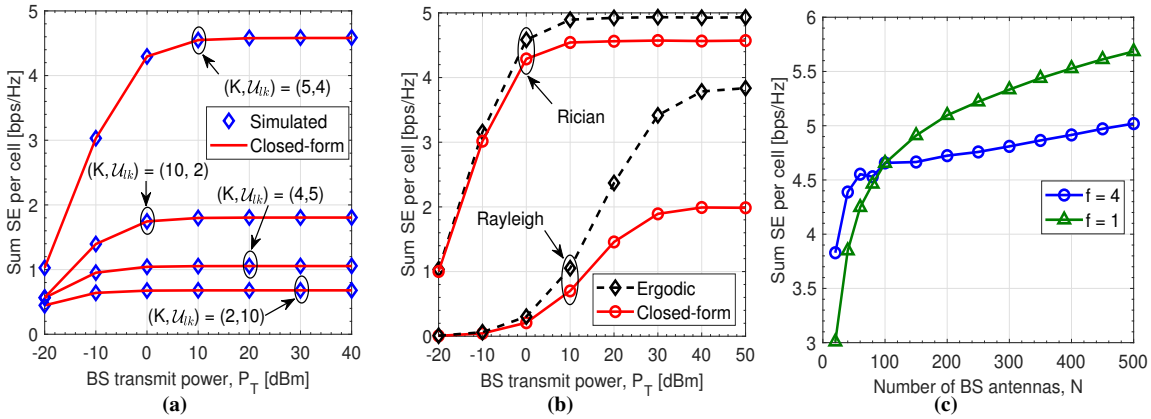

Fig. 2: Sum SE per cell versus BS transmit power P_T a) Lower bound validation; b) tightness of lower bound for Rayleigh and Rician channels and; c) Sum SE per cell vs BS antennas N for different frequency reuse factors.

2) ***Impact of channel hardening****:* The tightness of the closed-form SE expression depends on the variance of channel hardening i.e., $\overline{I}_{lk,n}^{(0)}$ in (16). High spatial correlation reduces the channel hardening, and the current SE lower bound, which exploits channel hardening, should underestimate the SE. The current bound, when evaluated for Rayleigh fading, by setting the LoS components to 0, also shows this degradation. We show this in Fig. 2b by plotting the sum SE per cell obtained using the current closed-form SE lower bound, and its ergodic counterpart for both correlated- Rayleigh and Rician fading channels with ASD $= 10°$. The ergodic sum SE is derived by assuming instantaneous channel knowledge at the users, and therefore represents the real channel rate [26]. We note from Fig. 2b that the current lower bound underestimates the SE for Rayleigh fading. But for Rician channels, it however closely matches with the ergodic SE. This is due to the LoS component in Rician fading, which ensures that the instantaneous combined channel is close to its mean value, which ensures adequate channel hardening. The hardening bound technique, therefore, does not underestimate the SE for a spatially-correlated Rician-faded mMIMO systems, and is commonly used to analyze them [3], [18].

3) ***Impact of pilot contamination and imperfect SIC****:* We now investigate in Fig. 2c the effect of pilot contamination on the sum SE. For this study, we consider $L = 4$ cells and choose

the number of pilot sequences as $\tau = fK$, with f being the pilot reuse factor. We consider two values of f i.e., $\{1, 4\}$. With $f = 1$, the same set of pilots are reused in each cell, while $f = 4$ assigns orthogonal pilots to all the relays in the system. The former f value leads to pilot contamination, while the latter avoids it. We observe the following from Fig. 2c:

- For $N \leq 100$ BS antennas and $f = 4$, the system has a higher sum SE per cell than $f = 1$. This is because the SE depends on the channel estimation overhead and the users SIC capability. The latter, in turn, depends on the hardening of the BS-relay channels $\mathbf{h}_k$, and the quality of relay-user channel estimates $\widehat{g}_{lk,n}^{lk}$ (see term $\overline{I}_{lk,n}^{(2)}$ in Eq. (16)). For $N \leq 100$, the channel does not harden considerably. The SIC performance can thus only be improved by improving the quality of the estimates $\widehat{g}_{lk,n}^{lk}$. With $f = 4$, where system employs orthogonal pilots, the channel estimation quality is substantially improved. It is crucial to note that $f = 4$ will also increase the pilot overhead. *In this regime, however, the SE gain due to better channel estimates, dominates the SE loss due to high estimation overhead.* This yields higher SE for $f = 4$ case.
- For $N \geq 100$ BS antennas, the opposite is true i.e., $f = 1$ has higher SE than $f = 4$. This is because the channel, due to large number of antennas, has sufficiently hardened now. This significantly improves the SIC for $f = 1$ and $f = 4$. The incremental SE gains provided by the improved channel estimation due to $f = 4$ are not enough to compensate the SE degradation due to its higher pilot overhead. This leads to a higher SE for $f = 1$.

This study crucially investigates the joint effect of pilot contamination and the channel hardening on the performance of mMIMO NOMA systems.

4) ***Impact of Spatial correlation and Rician factor****:* We now compare in Fig. 3a the sum SE per cell of a spatially-correlated and uncorrelated (denoted by dotted line) mMIMO NOMA systems by varying the ASD (correlation) and relay radius R_r. We observe the following:

- For a large relay radius of $R_r = 500$m, a correlated system has a higher SE than its uncorrelated counterpart. This is due to the higher inter-relay spatial separation, which reduces the first and second-hop inter-relay interference. The spatially-correlated BS-relay channels lie in non-overlapping eigenspaces, which further reduces the first-hop inter-relay interference. We also note that the increased relay radius also reduces the desired signal strength. The reduction in the first and second-hop inter-relay interference for this case, however, dominates for correlated channels, which increases the SE.

- For $R_r \leq 400$m, an uncorrelated system has a higher SE than correlated one. For these radii, the BS-relay and the inter-relay spatial separations reduce. Further, the channels now have a strong LoS path also, which reduces their randomness. The high correlation further makes these channels more deterministic. These two factors increase the inter-relay interference for the correlated case [26, Section 1.3, Pg. 179]. The increased randomness in the uncorrelated channel, however, weakens the inter-relay interference, which increases their SE.
- Also, irrespective of channel correlation, the SE is highest for $R_r = 200$m. This is because for this cell radius, the desired signal strength and different interferences are optimally balanced.

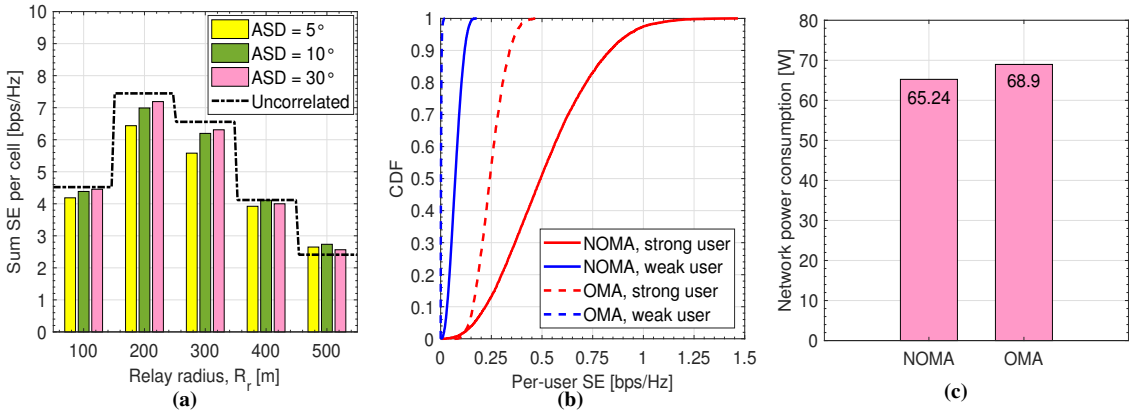

Fig. 3: a) Sum SE per cell versus R_r values. NOMA OMA comparison b) Per-user SE and c) Power consumption.

*5) **NOMA - OMA comparison**:* We next compare the system-level per-user SE and power consumption of a multi-cell NOMA system with its OMA counterpart. In OMA, each relay serves one user on a spectral resource. The coherence interval in OMA is therefore split into $\mathcal{U}_{lk}$ slots. And in each slot, the K relays simultaneously serve K users. For this study, we consider two system settings namely: i) NOMA system with $L = 4$ cells, $K = 5$ relays per cell and $\mathcal{U}_{lk} = 4$ *NOMA users* per relay and; ii) OMA system with $L = 4$ cells, $K = 20$ relays and $\mathcal{U}_{lk} = 1$ user per relay. These two settings are chosen to ensure that the total number of users in the network is constant i.e., $LK\mathcal{U}_{lk} = 80$ users.

We first compare in Fig. 3b, the CDF of per-user SE of the strongest and weakest users in NOMA and OMA systems. The strongest users have the highest SE and vice-versa. We note from Fig. 3b that the 90% likely per-user SE, which is the SE guaranteed for 90% of the users in the system, and occurs at CDF of 0.1, is highest for the NOMA system. This happens for both strong and weak users. This is because an OMA system, due to a higher number of relays, has higher estimation overhead (recall that Sec. II that each relay transmits a pilot signal), but also experiences severe inter-relay interference (1st + 2nd hop). The NOMA system, in contrast, has

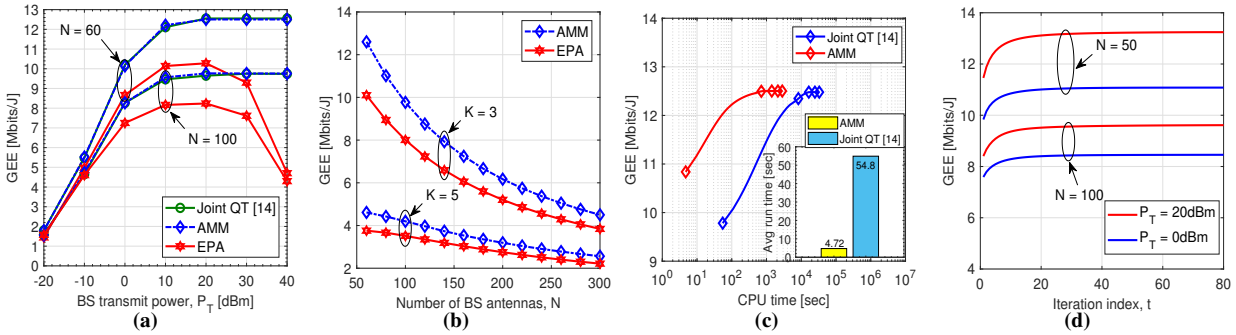

Fig. 4: a) GEE vs BS transmit power P_T: comparison of proposed AMM algorithm with joint quadratic transform in [14] b) GEE vs BS antennas for $K = \{3, 5\}$; c) GEE vs CPU time and; d) GEE vs iteration index t.

four times lesser relays, and consequently requires 4 times lesser number of orthogonal pilots to acquire CSI. Further, each user partially suppress the intra-relay interference using the estimated CSI, which increases the SE. This suggests that, for fixed number of users, NOMA system with $K = 5$ relays has high user fairness than an OMA system with $K = 20$ relays. We next compare the circuit power consumed by the aforementioned NOMA and OMA systems in Fig. 3c. We note that the NOMA system, due to considerably less number of relays, consumes roughly 6% less network power than OMA system, which signifies that NOMA is more energy efficient than OMA.

6) ***GEE optimization****:* We now plot in Fig. 4a, the GEE obtained using the proposed AMM-based algorithm (labelled as *AMM*) and compare it with ii) the joint optimization using QT in [14, Algorithm 1] $-$ which jointly optimizes the relay and BS transmit powers and; ii) equal power allocation (EPA) $-$ which allocates equal power to all the users and full power to relays. For this study, we consider two different number of BS antennas $N = \{60, 100\}$, $L = 4$ cells, $K = 3$ relays, $\mathcal{U}_{lk} = 4$ users per relay and ASD $= 10°$. We fix relay radius $R_r = 150$ m and user radius $R_u = 75$ m. We note that the proposed *AMM* and the joint algorithm in [14] have similar GEE values. The former, however as shown later, has significantly lesser complexity. We also note that GEE of AMM (resp. EPA) increases and saturates (resp. increases and then reduces) with increase in transmit power. This is because the AMM algorithm optimally allocates the transmit power required to maximize the GEE, and does not allocate any extra power after attaining its maximum. The EPA in contrast, keeps allocating all the available power to the BS and relays even after attaining a maximum GEE, and therefore reduces the GEE.

We next plot in Fig. 4b, the GEE by varying the number of BS antennas N. For this study, we fix $P_T = 20$ dBm, $Q_T = 25$ dBm, $\mathcal{U}_{lk} = 4$ users per relay and two different values of the number of relays i.e., $K = \{3, 5\}$. We notice that the GEE reduces with increase in N. This is

because the sum SE increases only logarithmically with N, while the circuit power consumption increases linearly with N. The increase in sum SE is thus not commensurate with the increased circuit power consumption, which reduces the GEE. We further infer that the GEE decreases with increase in K. This is is due to the increased 1st and 2nd hop inter-relay interference.

*7) **Algorithm time complexity and convergence**:* We plot in Fig. 4c the GEE vs CPU time and compare the time complexity and convergence of the proposed AMM algorithm with existing joint quadratic transform (QT) algorithm [14]. For this study, we fix $P_T = 30$ dBm, $N = 60$ BS antennas and run both AMM and joint QT algorithms for $I = 600$ iterations (each marker represents 150 iterations). We see that the AMM (resp. joint QT) algorithm converges to a stationary point within 150 (resp. 300) iterations. The average run time per iteration of the AMM algorithm is lesser than the joint QT algorithm. As a result, the AMM algorithm achieves optimal GEE in around 600 seconds, which is almost 35 times lesser than the time taken by the joint QT. The proposed AMM algorithm therefore, with a lesser run time per iteration and fewer number of iterations, yields the same GEE as that of the joint QT algorithm.

We next show in Fig. 4d, the convergence of the proposed AMM algorithm by plotting the GEE with iteration index t. We perform this study for two different values of transmit power and BS antennas i.e., $P_T = \{0, 20\}$ dBm and $N = \{50, 100\}$, respectively. We see that for both values of P_T and N, the GEE converges to a stationary point that too within 30 iterations.

VI. Conclusion

We derived an SE lower bound for a multi-cell multi-relay spatially-correlated Rician-faded mMIMO NOMA system, with imperfect CSI and SIC. We used this lower bound to numerically calculate the number of relays and the users/relay that the system should deploy to maximize the SE. We designed an AMM algorithm to optimize the non-convex GEE metric by constructing novel surrogate functions that tightly lower bound the GEE objective. We also showed that the proposed algorithm, when compared with the state-of-the-art in [14]: i) provides a similar SE but with a much less complexity; and ii) has a quicker convergence to a stationary point

Appendix A

We now derive the closed-form SE expression, by simplifying each of the expectations in (16). We start with the amplification factor μ_{lk} in (9) and rewrite it as follows:

$$\mu_{lk} = \sqrt{\frac{q_{lk}}{\mathbb{E}\left[\left|\sum_{l''=1}^{L}\sum_{k''=1}^{K}(\mathbf{h}_{lk}^{l''})^T\mathbf{w}_{l''k''}x_{l''k''} + z_{R_{lk}}\right|^2\right]}} \overset{(a)}{=} \sqrt{\frac{q_{lk}}{\mathbb{E}\left[\left|\sum_{l''=1}^{L}\sum_{k''=1}^{K}\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k''}x_{l''k''}\right|^2\right] + \mathbb{E}\left[|z_{R_{lk}}|^2\right]}}.$$

Equality (a) is due to the fact that the NOMA signal $x_{l''k''}$ is independent of the noise $z_{R_{lk}}$ and therefore the cross terms goes to 0. We note that the second term can be calculated as $\mathbb{E}(|z_{R_{lk}}|^2)=1$. We now consider the first term in the denominator and simplify it as follows

$$\mathbb{E}\left[\left|\sum_{l''=1}^{L}\sum_{k''=1}^{K}\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k''}x_{l''k''}\right|^2\right]=\sum_{l''=1}^{L}\sum_{k''=1}^{K}\sum_{l'=1}^{L}\sum_{k'=1}^{K}\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k''}x_{l''k''}x_{l'k'}^{*}\mathbf{w}_{l'k'}^{H}\mathbf{h}_{lk}^{l'*}\right]$$
$$\overset{(a)}{=}\sum_{l''=1}^{L}\sum_{k''=1}^{K}p_{l''k''}\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k''}\mathbf{w}_{l''k''}^{H}\mathbf{h}_{lk}^{l''*}\right] \tag{51}$$

Equality (a) is due to the independence of data signals $\{s_{k,n}\}$, $\forall k,n$. This leads to $\mathbb{E}(x_{l''k''}x_{l'k'}^{*})=0$, for $(l'',k'')\neq(l',k')$, and $\mathbb{E}(|x_{l''k''}|^2)=\sum_{n'=1}^{\mathcal{U}_{l''k''}}p_{l''k'',n'}\triangleq p_{l''k''}$. We now further simplify the expectation in (51) for $(l'',k'')=(l,k)$; $l''\neq l$ and $k''=k$ and; $k''\neq k$ cases as follows:

- *For $k''\neq k$:* The expectation $\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k''}\mathbf{w}_{l''k''}^{H}\mathbf{h}_{lk}^{l''*}\right]$ is simplified as

$$\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k''}\mathbf{w}_{l''k''}^{H}\mathbf{h}_{lk}^{l''*}\right]=\frac{\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\widehat{\mathbf{h}}_{l''k''}^{l''*}\widehat{\mathbf{h}}_{l''k''}^{l''T}\mathbf{h}_{lk}^{l''*}\right]}{\mathbb{E}(\|\widehat{\mathbf{h}}_{l''k''}^{l''}\|^2)}\overset{(a)}{=}\frac{\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{h}_{lk}^{l''*}\mathbf{h}_{lk}^{l''T}\right]\mathbb{E}\left[\widehat{\mathbf{h}}_{l''k''}^{l''*}\widehat{\mathbf{h}}_{l''k''}^{l''T}\right]\right)}{\mathbb{E}(\|\widehat{\mathbf{h}}_{l''k''}^{l''}\|^2)}$$
$$\overset{(b)}{=}\frac{\mathrm{Tr}\left((\bar{\mathbf{h}}_{lk}^{l''}\bar{\mathbf{h}}_{lk}^{l''H}+\mathbf{R}_{lk}^{l''})(\bar{\mathbf{h}}_{l''k''}^{l''}\bar{\mathbf{h}}_{l''k''}^{l''H}+\widehat{\mathbf{R}}_{l''k''}^{l''})\right)}{\delta_{l''k''}}\triangleq\rho_{l''k'',lk}. \tag{52}$$

  Here equality (a) is calculated by rewriting the expectation using $\mathrm{Tr}(\mathbf{AB})=\mathrm{Tr}(\mathbf{BA})$ and using the fact that, for $k''\neq k$, the channel $\mathbf{h}_{lk}^{l''}$ and the estimate $\widehat{\mathbf{h}}_{l''k''}^{l''}$ are mutually independent. Equality (b) is because $\mathbb{E}(\mathbf{h}_{lk}^{l''*}\mathbf{h}_{lk}^{l''T})=\mathbb{E}(\mathbf{h}_{lk}^{l''*}\mathbf{h}_{lk}^{l''H})=[\bar{\mathbf{h}}_{lk}^{l''}\bar{\mathbf{h}}_{lk}^{l''H}+\mathbf{R}_{lk}^{l''}]$, $\mathbb{E}(\widehat{\mathbf{h}}_{l''k''}^{l''*}\widehat{\mathbf{h}}_{l''k''}^{l''T})=[\bar{\mathbf{h}}_{l''k''}^{l''}\bar{\mathbf{h}}_{l''k''}^{l''H}+\widehat{\mathbf{R}}_{l''k''}^{l''}]$ and $\mathbb{E}(\left\|\widehat{\mathbf{h}}_{l''k''}^{l''}\right\|^2)=\mathrm{Tr}(\bar{\mathbf{h}}_{l''k''}^{l''}\bar{\mathbf{h}}_{l''k''}^{l''H}+\widehat{\mathbf{R}}_{l''k''}^{l''})\triangleq\delta_{l''k''}$.

- *For $l''\neq l$ and $k''=k$:* The expectation is simplified as

$$\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\mathbf{w}_{l''k}\mathbf{w}_{l''k}^{H}\mathbf{h}_{lk}^{l''*}\right]=\frac{\mathbb{E}\left[\mathbf{h}_{lk}^{l''T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''T}\mathbf{h}_{lk}^{l''*}\right]}{\mathbb{E}(\|\widehat{\mathbf{h}}_{l''k}^{l''}\|^2)}\overset{(a)}{=}\frac{\mathbb{E}\left[\widehat{\mathbf{h}}_{lk}^{l''T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''T}\widehat{\mathbf{h}}_{lk}^{l''*}\right]}{\delta_{l''k}}+\frac{\mathbb{E}\left[\mathbf{e}_{lk}^{l''T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''T}\mathbf{e}_{lk}^{l''*}\right]}{\delta_{l''k}}.$$

  Equality (a) is obtained by i) substituting $\mathbf{h}_{lk}^{l''}=\widehat{\mathbf{h}}_{lk}^{l''}+\mathbf{e}_{lk}^{l''}$; ii) exploiting the fact that MMSE estimate and estimation error are statistically independent and; iii) using $\mathbb{E}(\left\|\widehat{\mathbf{h}}_{l''k}^{l''}\right\|^2)=\delta_{l''k}$. Due to pilot contamination, the estimates $(\widehat{\mathbf{h}}_{lk}^{l''},\widehat{\mathbf{h}}_{l''k}^{l''})$ are correlated. We now use Lemma 5 in [3] and simplify the expectation $\mathbb{E}(\widehat{\mathbf{h}}_{lk}^{l''T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''T}\widehat{\mathbf{h}}_{lk}^{l''*})$ as follows

$$\mathbb{E}(\widehat{\mathbf{h}}_{lk}^{l''T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''T}\widehat{\mathbf{h}}_{lk}^{l''*})=p_{lk}^{\rho}p_{l''k}^{\rho}\tau_p^2\left|\mathrm{Tr}(\mathbf{R}_{lk}^{l''}\mathbf{\Psi}_{l''k}\mathbf{R}_{l''k}^{l''})\right|^2+\mathrm{Tr}(\widehat{\mathbf{R}}_{lk}^{l''}\widehat{\mathbf{R}}_{l''k}^{l''})+\left|\bar{\mathbf{h}}_{l''k}^{l''H}\bar{\mathbf{h}}_{lk}^{l''}\right|^2$$
$$+2\sqrt{p_{l''k}^{\rho}p_{lk}^{\rho}}\tau_p\mathrm{Re}\left\{\mathrm{Tr}(\mathbf{R}_{lk}^{l''}\mathbf{\Psi}_{l''k}\mathbf{R}_{l''k}^{l''})\bar{\mathbf{h}}_{lk}^{l''H}\bar{\mathbf{h}}_{l''k}^{l''}\right\}+\bar{\mathbf{h}}_{lk}^{l''H}\widehat{\mathbf{R}}_{l''k}^{l''}\bar{\mathbf{h}}_{lk}^{l''}+\bar{\mathbf{h}}_{l''k}^{l''H}\widehat{\mathbf{R}}_{lk}^{l''}\bar{\mathbf{h}}_{l''k}^{l''}. \tag{53}$$

  Equality (a) is obtained by substituting $\widehat{\mathbf{h}}_{lk}^{l''}=\bar{\mathbf{h}}_{lk}^{l''}+\mathbf{R}_{lk}^{l''}\mathbf{\Psi}_{l''k}(\mathbf{y}_{l''k}-\bar{\mathbf{y}}_{l''k})$, $\widehat{\mathbf{h}}_{l''k}^{l''}=\bar{\mathbf{h}}_{l''k}^{l''}+\mathbf{R}_{l''k}^{l''}\mathbf{\Psi}_{l''k}(\mathbf{y}_{l''k}-\bar{\mathbf{y}}_{l''k})$, with $\mathbf{y}_{l''k}\sim\mathcal{CN}(\mathbf{0},\tau_p\mathbf{\Psi}_{l''k}^{-1})$ and then using Lemma 5 from [3]. The

second expectation $\mathbb{E}[\mathbf{e}_{lk}^{l''^T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''^T}\mathbf{e}_{lk}^{l''*}]$ is similarly simplified as

$$\mathbb{E}[\mathbf{e}_{lk}^{l''^T}\widehat{\mathbf{h}}_{l''k}^{l''*}\widehat{\mathbf{h}}_{l''k}^{l''^T}\mathbf{e}_{lk}^{l''*}] = \mathrm{Tr}\left((\mathbf{R}_{lk}^{l''} - \widehat{\mathbf{R}}_{lk}^{l''})(\bar{\mathbf{h}}_{l''k}^{l''}\bar{\mathbf{h}}_{l''k}^{l''H} + \widehat{\mathbf{R}}_{l''k}^{l''})\right). \tag{54}$$

The closed-form expression for $\mathbb{E}[\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k}\mathbf{w}_{l''k}^H\mathbf{h}_{lk}^{l''*}]$ is therefore obtained using (53) and (54) as $\mathbb{E}[\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k}\mathbf{w}_{l''k}^H\mathbf{h}_{lk}^{l''*}] = (\rho_{l''k,lk} + \xi_{l''k,lk})$, where

$\rho_{l''k,lk} = \frac{1}{\delta_{l''k}}\left[\mathrm{Tr}((\mathbf{R}_{lk}^{l''} + \bar{\mathbf{h}}_{lk}^{l''}\bar{\mathbf{h}}_{lk}^{l''H})(\widehat{\mathbf{R}}_{l''k}^{l''} + \bar{\mathbf{h}}_{l''k}^{l''}\bar{\mathbf{h}}_{l''k}^{l''H}))\right]$ and

$\xi_{l''k,lk} = \frac{1}{\delta_{l''k}}\left[p_{lk}^{\rho}p_{l''k}^{\rho}\tau_p^2\left|\mathrm{Tr}(\mathbf{R}_{lk}^{l''}\mathbf{\Psi}_{l''k}\mathbf{R}_{l''k}^{l''})\right|^2 + 2\sqrt{p_{l''k}^{\rho}p_{lk}^{\rho}}\tau_p\mathrm{Re}\left\{\mathrm{Tr}(\mathbf{R}_{lk}^{l''}\mathbf{\Psi}_{l''k}\mathbf{R}_{l''k}^{l''})\bar{\mathbf{h}}_{lk}^{l''H}\bar{\mathbf{h}}_{l''k}^{l''}\right\}\right]$.

- *For $(l'', k'') = (l, k)$:* The closed-form expression, obtained on lines similar to previous case, is given as $\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\mathbf{w}_{lk}^H\mathbf{h}_{lk}^{l^*}\right] = \dfrac{\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\widehat{\mathbf{h}}_{lk}^{l^*}\widehat{\mathbf{h}}_{lk}^{l^T}\mathbf{h}_{lk}^{l^*}\right]}{\mathbb{E}(\|\widehat{\mathbf{h}}_{lk}^{l}\|^2)} = (\rho_{lk,lk} + \xi_{lk,lk})$.

The closed-form expression for the first expectation in the denominator of μ_{lk} is simplified as

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{l''=1}^{L}\sum_{k''=1}^{K}\mathbf{h}_{lk}^{l''^T}\mathbf{w}_{l''k''}x_{l''k''}\right|^2\right] &= \sum_{l''=1}^{L}\sum_{k''\neq k}^{K}\rho_{l''k'',lk}p_{l''k''} + \sum_{l''=1}^{L}(\rho_{l''k,lk} + \xi_{l''k,lk})p_{l''k} \\ &\triangleq \sum_{l''=1}^{L}\sum_{k''=1}^{K}\rho_{l''k'',lk}p_{l''k''} + \sum_{l''=1}^{L}\xi_{l''k,lk}p_{l''k}. \end{aligned}\tag{55}$$

The closed-form expression for the amplification factor μ_k is therefore given as

$$\tilde{\mu}_{lk} = \sqrt{\frac{q_{lk}}{\sum_{l''=1}^{L}\sum_{k''=1}^{K}\rho_{l''k'',lk}p_{l''k''} + \sum_{l''=1}^{L}\xi_{l''k,lk}p_{l''k} + 1}}. \tag{56}$$

We now consider the term $\overline{\Delta}_{lk,n}$ in (16) and it can be simplified as

$$\overline{\Delta}_{lk,n} = \left|\mu_{lk}\mathbb{E}\left[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right|^2 p_{lk,n} = \mu_{lk}^2 p_{lk,n}\left|\mathbb{E}\left[\frac{\hat{g}_{lk,n}^{lk}}{|\hat{g}_{lk,n}^{lk}|}g_{lk,n}^{lk}\right]\right|^2\left|\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\frac{\widehat{\mathbf{h}}_{lk}^{l^*}}{\sqrt{\mathbb{E}(\|\widehat{\mathbf{h}}_{lk}^{l}\|^2)}}\right]\right|^2$$

$$\overset{(a)}{=} \mu_{lk}^2 p_{lk,n}|\mathbb{E}(|\hat{g}_{lk,n}^{lk}|)|^2\mathbb{E}(\|\widehat{\mathbf{h}}_{lk}^{l}\|^2) \overset{(b)}{=} \mu_{lk}^2 p_{lk,n}\delta_{lk}\frac{\pi}{4}\upsilon_{lk,n}^{lk}\left[L_{1/2}\left(\frac{-\left|\overline{g}_{lk,n}^{lk}\right|^2}{\upsilon_{lk,n}^{lk}}\right)\right]^2 \overset{(c)}{=} A_{lk,n}p_{lk,n}\tilde{\mu}_{lk}^2. \tag{57}$$

Here $A_{lk,n} = \frac{\pi}{4}\delta_{lk}\upsilon_{lk,n}^{lk}\left[L_{1/2}\left(\frac{-\left|\overline{g}_{lk,n}^{lk}\right|^2}{\upsilon_{lk,n}^{lk}}\right)\right]^2$. Equality (a) is obtained by substituting $\mathbf{h}_{lk}^{l} = \widehat{\mathbf{h}}_{lk}^{l} + \mathbf{e}_{lk}^{l}$, $g_{lk,n}^{lk} = \widehat{g}_{lk,n}^{lk} + e_{lk,n}^{lk}$ and by using the fact that MMSE estimate and estimation error are statistically independent [28]. Equality (b) is because $\mathbb{E}(\|\widehat{\mathbf{h}}_{lk}^{l}\|^2) = \delta_{lk}$ and the expectation $\mathbb{E}(|\widehat{g}_{lk,n}^{lk}|)$ is simplified by noting that the MMSE estimate $\widehat{g}_{lk,n}^{lk}$ follows a complex normal distribution with mean $\overline{g}_{lk,n}^{lk}$ and variance $\upsilon_{lk,n}^{lk}$. The magnitude $|\widehat{g}_{lk,n}^{lk}|$ is Rice distributed as $\mathrm{Rice}(|\overline{g}_{lk,n}^{lk}|, \sqrt{\frac{\upsilon_{lk,n}^{lk}}{2}})$, whose first moment is given by $\mathbb{E}(|\widehat{g}_{lk,n}^{lk}|) = \sqrt{\frac{\pi}{2}}\sqrt{\frac{\upsilon_{lk,n}^{lk}}{2}}L_{\frac{1}{2}}(-|\overline{g}_{lk,n}^{lk}|^2/\upsilon_{lk,n}^{lk})$, where $L_{\frac{1}{2}}(\cdot)$ denotes the Laguerre polynomial function [30, 13.6.9]. Equality (c) is obtained by replacing μ_{lk} with its closed-form expression $\tilde{\mu}_{lk}$ as in (56). We now consider the term $\overline{I}_{lk,n}^{(0)}$ in

(16) and simplify it as follows

$$\begin{aligned}\overline{I}_{lk,n}^{(0)} &= \mu_{lk}^2 p_{lk,n}\Bigg\{\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right|^2\right] - \left|\mathbb{E}\left[f_{lk,n}g_{lk,n}^{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right|^2\Bigg\} \\ &= \mu_{lk}^2 p_{lk,n}\Bigg\{\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{lk}\right|^2\right]\mathbb{E}\left[\left|\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right|^2\right] - \left|\mathbb{E}\left[f_{lk,n}g_{lk,n}^{lk}\right]\mathbb{E}\left[\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right]\right|^2\Bigg\} \\ &\overset{(a)}{=} \tilde{\mu}_{lk}^2 p_{lk,n}\Bigg\{\underbrace{\left(\left|\overline{g}_{lk,n}^{lk}\right|^2+\gamma_{lk,n}^{lk}\right)\left(\rho_{lk,lk}+\xi_{lk,lk}\right)-A_{lk,n}}_{C_{lk,n}^{(0)}}\Bigg\} \overset{(b)}{=} C_{lk,n}^{(0)}p_{lk,n}\tilde{\mu}_{lk}^2. \end{aligned} \tag{58}$$

Equality (a) is obtained by substituting $\mathbb{E}[\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}] = \sqrt{\delta_{lk}}$, $\mathbb{E}(|\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}|^2) = (\rho_{lk,lk}+\xi_{lk,lk})$, $\mathbb{E}(f_{lk,n}g_{lk,n}^{lk}) = \frac{\sqrt{\pi}}{2}\upsilon_{lk,n}^{lk}L_{\frac{1}{2}}\left(\frac{-|\bar{g}_{lk,n}^{lk}|^2}{\upsilon_{lk,n}^{lk}}\right)$ and $\mathbb{E}(|f_{lk,n}g_{lk,n}^{lk}|^2) = \mathbb{E}(\frac{|\widehat{g}_{lk,n}^{lk}|^2|g_{lk,n}^{lk}|^2}{|\widehat{g}_{lk,n}^{lk}|^2}) = \mathbb{E}(|g_{lk,n}^{lk}|^2) = |\bar{g}_{lk,n}^{lk}|^2+\gamma_{lk,n}^{lk}$. We now consider the term $\overline{I}_{lk,n}^{(1)}$ in (16) and obtain its closed-form as follows

$$\begin{aligned}\overline{I}_{lk,n}^{(1)} = \sum_{n'=1}^{n-1}p_{lk,n'}\mathbb{E}\left[|f_{lk,n}g_{lk,n}^{lk}\mu_{lk}\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}|^2\right] &= \sum_{n'=1}^{n-1}p_{lk,n'}\mu_{lk}^2\mathbb{E}\left[\left|f_{lk,n}g_{lk,n}^{lk}\right|^2\right]\mathbb{E}\left[\left|\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}\right|^2\right] \\ &\overset{(a)}{=} C_{lk,n}^{(1)}\sum_{n'=1}^{n-1}p_{lk,n'}\tilde{\mu}_{lk}^2.\end{aligned}$$

Here $C_{k,n}^{(1)} = \chi_{lk,n}^{lk}\left(\rho_{lk,lk}+\xi_{lk,lk}\right)$. Equality (a) is obtained by using $\mathbb{E}(|f_{lk,n}g_{lk,n}^{lk}|^2) = \left(\left|\overline{g}_{lk,n}^{lk}\right|^2+\gamma_{lk,n}^{lk}\right) \triangleq \chi_{lk,n}^{lk}$, $\mathbb{E}(|\mathbf{h}_{lk}^{l^T}\mathbf{w}_{lk}|^2) = \rho_{lk,lk}+\xi_{lk,lk}$ and by replacing μ_{lk} with $\tilde{\mu}_{lk}$ as in (56). Similarly, we can calculate the rest of the interference terms as

$$\overline{I}_{lk,n}^{(2)} = \sum_{n'=n+1}^{\mathcal{U}_{lk}}C_{lk,n}^{(2)}p_{lk,n'}\tilde{\mu}_{lk}^2, \quad \overline{I}_{k,n}^{(7)} = \sum_{l'=1}^{L}\sum_{k'=1}^{K}C_{l'k',lk,n}^{(5)}\tilde{\mu}_{l'k'}^2$$

$$\sum_{m=3}^{6}\overline{I}_{lk,n}^{(m)} = \sum_{(l',k')\neq(l,k)}\sum_{n'=1}^{\mathcal{U}_{l'k'}}C_{l'k',lk,n}^{(3)}p_{l'k',n'}\mu_{l'k'}^2 + \sum_{(l',k')}\sum_{(l'',k'')\neq(l',k')}\sum_{n'=1}^{\mathcal{U}_{l''k''}}\underset{lk,n}{C_{l''k'',l'k'}^{(4)}}p_{l''k'',n'}\tilde{\mu}_{l'k'}^2.$$

Here $C_{lk,n}^{(2)} = \left\{\chi_{lk,n}^{lk}\left(\rho_{lk,lk}+\xi_{lk,lk}\right) - \delta_{lk}\left(\left|\overline{g}_{lk,n}^{lk}\right|^2+\upsilon_{lk,n}^{lk}\right)\right\}$, $C_{l'k',lk,n}^{(3)} = \chi_{lk,n}^{l'k'}(\rho_{l'k',l'k'}+\xi_{l'k',l'k'})$, $\underset{lk,n}{C_{l''k'',l'k'}^{(4)}} = \chi_{lk,n}^{l'k'}\rho_{l''k'',l'k'}$, $C_{k',k,n}^{(5)} = \chi_{lk,n}^{l'k'}$ with $\chi_{lk,n}^{l'k'} = \left(\left|\overline{g}_{lk,n}^{l'k'}\right|^2+\gamma_{lk,n}^{l'k'}\right)$.

APPENDIX B

We first discuss the design of the function $h_{lb}(\mathbf{x}, y_k)$ and later prove that it satisfies the conditions C1 and C2. We now consider the fractional term $A_k(\mathbf{x})/B_k(\mathbf{x})$ and rewrite it by introducing variable y_k and expressing the numerator in quadratic over affine form as

$$\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} = \frac{y_k^2\sqrt{A_k(\mathbf{x})}^2}{y_k^2B_k(\mathbf{x})} \triangleq \frac{N_k(\mathbf{x},y_k)^2}{D_k(\mathbf{x},y_k)}. \tag{59}$$

Here $N_k(\mathbf{x},y_k) = y_k\sqrt{A_k(\mathbf{x})}$ and $D_k(\mathbf{x},y_k) = y_k^2B_k(\mathbf{x})$. The scalar y_k is a non-negative auxilary variable, which is a function of the feasible point $\mathbf{x}^t$ and is designed later to satisfy the conditions C1 and C2. Using arithmetic mean-harmonic mean inequality over the functions $N_k(\mathbf{x},y_k)$,

$D_k(\mathbf{x}, y_k)$, we have

$$\frac{N_k(\mathbf{x}, y_k) + D_k(\mathbf{x}, y_k)}{2} \geq \frac{2}{\frac{1}{N_k(\mathbf{x}, y_k)} + \frac{1}{D_k(\mathbf{x}, y_k)}} \tag{60}$$

$$\overset{(a)}{\Longrightarrow} \frac{N_k(\mathbf{x}, y_k)^2}{D_k(\mathbf{x}, y_k)} \geq 2N_k(\mathbf{x}, y_k) - D_k(\mathbf{x}, y_k) \overset{(b)}{\Longrightarrow} \frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} \geq 2y_k\sqrt{A_k(\mathbf{x})} - y_k^2 B_k(\mathbf{x}) \triangleq h_k^{lb}(\mathbf{x}, y_k).$$

Implication (a) is obtained by re-arranging the terms $N_k(\mathbf{x}, y_k)$ and $D_k(\mathbf{x}, y_k)$, and implication (b) is obtained by substituting $N_k(\mathbf{x}, y_k) = y_k\sqrt{A_k(\mathbf{x})}$ and $D_k(\mathbf{x}, y_k) = y_k^2 B_k(\mathbf{x})$. The auxilary variable y_k is designed to ensure that the function $h_k^{lb}(\mathbf{x}, y_k)$ satisfies the properties C1 and C2. For a given $\mathbf{x} = \mathbf{x}^t$, using property C1, the function $h_k^{lb}(\mathbf{x}, y_k)$ and the original function $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$ should have same function value. The value of variable y_k is therefore obtained by evaluating $h_k^{lb}(\mathbf{x}, y_k)$ at $\mathbf{x} = \mathbf{x}^t$ and equating it with $\frac{A_k(\mathbf{x}^t)}{B_k(\mathbf{x}^t)}$ as

$$\begin{aligned} h_k^{lb}(\mathbf{x}, y_k)\big|_{\mathbf{x}=\mathbf{x}^t} = \frac{A_k(\mathbf{x}^t)}{B_k(\mathbf{x}^t)} &\implies 2y_k\sqrt{A_k(\mathbf{x}^t)} - y_k^2 B_k(\mathbf{x}^t) = \frac{A_k(\mathbf{x}^t)}{B_k(\mathbf{x}^t)} \\ &\implies y_k^2\big(B_k(\mathbf{x}^t)\big)^2 - 2y_k\sqrt{A_k(\mathbf{x}^t)}B_k(\mathbf{x}^t) + A_k(\mathbf{x}^t) = 0 \\ &\implies (y_k B_k(\mathbf{x}^t) - \sqrt{A_k(\mathbf{x}^t)})^2 = 0 \implies y_k = \frac{\sqrt{A_k(\mathbf{x}^t)}}{B_k(\mathbf{x}^t)}. \end{aligned} \tag{61}$$

We now show that the function $h_k^{lb}(\mathbf{x}, y_k)$ satisfies property C2 by evaluating the gradients of the fractional function $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$ at $\mathbf{x} = \mathbf{x}^t$ as

$$\nabla_{\mathbf{x}}\Big[\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}\Big]\Big|_{\mathbf{x}=\mathbf{x}^t} = \frac{B_k(\mathbf{x}^t)\left[\nabla_{\mathbf{x}}A_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t}\right] - A_k(\mathbf{x}^t)\left[\nabla_{\mathbf{x}}B_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t}\right]}{B_k(\mathbf{x}^t)^2} \tag{62}$$

Similarly the gradient of the function $h_k^{lb}(\mathbf{x}, y_k)$ at $\mathbf{x} = \mathbf{x}^t$ is given as

$$\begin{aligned} \nabla_{\mathbf{x}}h_k^{lb}(\mathbf{x}, y_k)\Big|_{\mathbf{x}=\mathbf{x}^t} &= 2\frac{\sqrt{A_k(\mathbf{x}^t)}}{B_k(\mathbf{x}^t)}\frac{1}{2\sqrt{A_k(\mathbf{x}^t)}}\nabla_{\mathbf{x}}A_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t} - \frac{A_k(\mathbf{x}^t)}{B_k(\mathbf{x}^t)^2}\nabla_{\mathbf{x}}B_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t} \\ &= \frac{B_k(\mathbf{x}^t)\left[\nabla_{\mathbf{x}}A_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t}\right] - A_k(\mathbf{x}^t)\left[\nabla_{\mathbf{x}}B_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t}\right]}{B_k(\mathbf{x}^t)^2} = \nabla_{\mathbf{x}}\Big[\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}\Big]\Big|_{\mathbf{x}=\mathbf{x}^t}. \end{aligned} \tag{63}$$

We note that the function $h_{lb}(\mathbf{x}, y_k)$ and the original fraction both has same gradient at $\mathbf{x} = \mathbf{x}^t$ and therefore satisfies C2. The function $h_k(\mathbf{x}, y_k)$ is therefore a valid surrogate function.

APPENDIX C

We now discuss the construction of the function $h_k^{ub}(\mathbf{x}, z)$. We consider the fraction $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$, introduce the variable z_k and rewrite it as product of functions as follows

$$\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} = \frac{\sqrt{z_k}A(\mathbf{x})}{\sqrt{z_k}B(\mathbf{x})} \triangleq \overline{N}_k(\mathbf{x}, z_k)\overline{D}_k(\mathbf{x}, z_k), \tag{64}$$

where $\overline{N}_k(\mathbf{x}, z_k) = \sqrt{z_k}A_k(\mathbf{x})$ and $\overline{D}_k(\mathbf{x}, z_k) = 1/(\sqrt{z_k}B_k(\mathbf{x}))$. We now use the arithmetic mean-geometric mean inequality and upper bound the product of functions in (64) as

$$\overline{N}_k(\mathbf{x}, z_k)\overline{D}_k(\mathbf{x}, z_k) \leq \frac{1}{2}(\overline{N}_k(\mathbf{x}, z_k))^2 + \frac{1}{2}(\overline{D}_k(\mathbf{x}, z_k))^2$$

$$\implies \frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} \leq \frac{z_k}{2}\left[A_k(\mathbf{x})\right]^2 + \frac{1}{2z_k}\left[\frac{1}{B_k(\mathbf{x})}\right]^2 \triangleq h_k^{ub}(\mathbf{x}, z_k). \tag{65}$$

The scalar z_k is calculated using property C1 i.e., by substituting the feasible point $\mathbf{x} = \mathbf{x}^t$ in the original fraction $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$ and $h_{ub}(\mathbf{x}, z)$ and equating them, as follows

$$\frac{z_k}{2}\left[A_k(\mathbf{x}^t)\right]^2 + \frac{1}{2z_k}\left[\frac{1}{B_k(\mathbf{x}^t)}\right]^2 = \frac{A_k(\mathbf{x}^t)}{B_k(\mathbf{x}^t)} \implies z_k^2\left[A_k(\mathbf{x}^t)B_k(\mathbf{x}^t)\right]^2 - 2z_kA_k(\mathbf{x}^t)B_k(\mathbf{x}^t) + 1 = 0$$

$$\implies \left[z_kA_k(\mathbf{x}^t)B_k(\mathbf{x}^t) - 1\right]^2 = 0 \implies z_k = \frac{1}{A_k(\mathbf{x}^t)B_k(\mathbf{x}^t)}. \tag{66}$$

We now prove that the proposed upper bound satisfies property C2. We now evaluate the gradient of function $h_k^{ub}(\mathbf{x}, z_k)$ at feasible point $\mathbf{x} = \mathbf{x}^t$ as

$$\nabla_{\mathbf{x}} h_k^{ub}(\mathbf{x}, z_k)\Big|_{\mathbf{x}=\mathbf{x}^t} = \frac{1}{A_k(\mathbf{x}^t)B_k(\mathbf{x}^t)}A_k(\mathbf{x}^t)\nabla_{\mathbf{x}}A_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t} - \frac{A_k(\mathbf{x}^t)B_k(\mathbf{x}^t)}{\left[B_k(\mathbf{x}^t)\right]^3}\nabla_{\mathbf{x}}B_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^t}. \tag{67}$$

By rearranging the terms in the above expression, we note that the gradient of the function $h_k^{ub}(\mathbf{x}, z_k)$ at $\mathbf{x} = \mathbf{x}^t$ exactly matches with the gradient of original function $\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})}$ in (62) and therefore satisfies property C2. The function $h_k^{ub}(\mathbf{x}, z_k)$ is therefore a valid surrogate function.

REFERENCES

[1] E. Bjornson, L. Van der Perre, S. Buzzi, and E. G. Larsson, "Massive MIMO in sub-6 Ghz and mmWave: Physical, practical, and use-case differences," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 100–108, 2019.

[2] L. Sanguinetti, E. Björnson, and J. Hoydis, "Toward massive MIMO 2.0: Understanding spatial correlation, interference suppression, and pilot contamination," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 232–257, Jan. 2020.

[3] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Massive MIMO with spatially correlated Rician fading channels," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3234–3250, 2019.

[4] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32–38, 2018.

[5] A. S. de Sena, F. R. M. Lima, D. B. da Costa, Z. Ding, P. H. J. Nardelli, U. S. Dias, and C. B. Papadias, "Massive MIMO-NOMA networks with imperfect SIC: design and fairness enhancement," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6100–6115, 2020.

[6] L. Liu, Y. Chi, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving MIMO-NOMA: Iterative LMMSE detection," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1758–1773, 2019.

[7] V. Mandawaria, E. Sharma, and R. Budhiraja, "WSEE maximization of mmwave NOMA systems," *IEEE Commun. Lett.*, vol. 23, no. 8, pp. 1413 – 1417, 2019.

[8] D. Kudathanthirige and G. A. A. Baduge, "NOMA-aided multicell downlink massive MIMO," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 612–627, 2019.

[9] Y. Li and G. A. A. Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, 2018.

[10] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the performance of cell-free massive MIMO relying on adaptive NOMA/OMA mode-switching," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 792–810, 2020.

[11] M. Vaezi, G. A. Aruma Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 900–919, 2019.

[12] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan, and T. Sato, "Performance analysis of non-regenerative massive-MIMO-NOMA relay systems for 5G," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4777–4790, 2017.

[13] X. Chen, R. Jia, and D. W. K. Ng, "The application of relay to massive non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5168–5180, 2018.

[14] Mandawaria, E. Sharma, and R. Budhiraja, "Energy-efficient massive MIMO multi-relay NOMA systems with CSI errors," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7410–7428, 2020.

[15] Y. Li and G. Amarasuriya, "Multiple relay-aided massive MIMO NOMA," in *2019 IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.

[16] Y. Li and G. A. A. Baduge, "Relay-aided downlink massive MIMO NOMA with estimated CSI," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2258–2271, 2021.

[17] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan, and T. Sato, "Performance analysis of non-regenerative massive-MIMO-NOMA relay systems for 5G," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4777–4790, 2017.

[18] J. Zhang, J. Fan, B. Ai, and D. W. K. Ng, "NOMA-based cell-free massive MIMO over spatially correlated rician fading channels," in *2020 IEEE Int. Conf. on Commun., ICC 2020, Dublin, Ireland, June 7-11, 2020*. IEEE, 2020, pp. 1–6.

[19] M. R. Zamani, M. Eslami, M. Khorramizadeh, and Z. Ding, "Energy-efficient power allocation for NOMA with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 1009–1013, 2019.

[20] V. Khodamoradi, A. Sali, O. Messadi, A. Khalili, and B. M. Ali, "Energy-efficient massive MIMO SWIPT-enabled systems," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2022.

[21] J. Shi, W. Yu, Q. Ni, W. Liang, Z. Li, and P. Xiao, "Energy efficient resource allocation in hybrid non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3496–3511, 2019.

[22] A. Zappone and E. A. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Found. and Trends in Commun. and Inf. Theory*, vol. 11, no. 3-4, p. 185–396, 2015.

[23] K. Shen, W. Yu, L. Zhao, and D. P. Palomar, "Optimization of MIMO device-to-device networks via matrix fractional programming: A minorization–maximization approach," *IEEE/ACM Trans. Netw.*, vol. 27, no. 5, pp. 2164–2177, Oct 2019.

[24] A. Arora, C. G. Tsinos, M. R. B. Shankar, S. Chatzinotas, and B. Ottersten, "Efficient algorithms for constant-modulus analog beamforming," *IEEE Trans. Signal Process.*, vol. 70, pp. 756–771, 2022.

[25] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, 2017.

[26] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, energy, and hardware efficiency," *Found. and Trends® in Signal Process.*, vol. 11, no. 3-4, pp. 154–655, 2017.

[27] X. Chen, R. Jia, and D. W. K. Ng, "On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2539–2551, 2019.

[28] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1997.

[29] X. Liu and L. Hanzo, "A unified exact BER performance analysis of asynchronous DS-CDMA systems using BPSK modulation over fading channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3504–3509, 2007.

[30] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth Dover printing ed. New York: Dover, 1964.

[31] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[32] A. Ben-Tal and A. S. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications, vol. 2. Philadelphia, PA, USA: SIAM*, 2001.