NOMA-Aided Multi-Way Massive MIMO Relaying

Shashindra Silva, *Student Member, IEEE*, Gayan Amarasuriya Aruma Baduge, *Member, IEEE*, Masoud Ardakani, *Senior Member, IEEE*, and Chintha Tellambura, *Fellow, IEEE*

***Abstract*—For a multi-way relay network (MWRN) with K users, K time slots are needed for full data exchange. Thus, the overall spectral efficiency, due to the $1/K$ pre-log factor, declines as number of users grows. It has recently been improved to roughly $K/2$ time slots, but even this improvement does not arrest the decline. Herein, we reduce this task to just two time slots regardless of K. To do this, we exploit the performance gains of non-orthogonal multiple-access (NOMA) and a massive multiple-input multiple-output (MIMO) relay. First, the users transmit their signals to the relay, which uses maximal ratio combining reception. Next, the relay transmits a superposition-coded signal for all users by using maximal ratio transmission. Each user then performs successive interference cancellation (SIC) decoding of data symbols of the other $K-1$ user nodes. We use the so-called worst-case Gaussian approximation to derive the overall sum rate and demonstrate significant spectral-efficiency gains and energy-efficiency gains over the existing MWRN counterparts. We also design the relay power allocation matrix to maximize the minimum among the user rates, thus maximizing the user fairness. Furthermore, the effects of imperfect SIC and imperfect channel state information (CSI) on the sum rate are analyzed.**

***Index Terms*—Massive MIMO, Multi-way relay, MWRN, NOMA, Imperfect SIC**

I. Introduction

A. Background

Multi-way relay networks (MWRNs) are used to allow full, partial, or clustered data exchange among a set of users (more than two) through an intermediate relay node [1]–[4]. MWRN channel generalizes two-way relay channel, which allows two users to fully exchange their information via a relay. And achievable information rates of the MWRN channel were first developed in [2]. In this network, multiple users first transmit their messages to the relay and then the relay will transmit signals over multiple time slots to enable message exchange among users. In full data exchange, each users has a common message to all the others. There are no direct links among the users due to propagation impairments such as large-scale fading and/or shadowing effects. Applications of MWRNs include simultaneous multi-directional communications via a base station, wireless sensor networks, and satellite communications [2] and will increase with the emergence of Internet-of-Things (IoT) for the next-generation wireless systems [5]. Thus, beamforming methods [6], [7], relay selection schemes [8], and coding schemes [9] have been developed recently for MWRNs.

Shashindra Silva, Masoud Ardakani and Chintha Tellambura are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6G 2V4. Email: {jayamuni, ardakani, chintha}@ece.ualberta.ca. Gayan Amarasuriya Aruma Baduge is with the Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA 62901. Email: gayan.baduge@siu.edu. This work in part has been presented at IEEE International Conference on Communications (ICC), May 2019, Shanghai, China [1].

Massive multiple-input multiple-output (MIMO) is another key enabler of 5G wireless [10]. Here, it exploits an unprecedented amount of spatial degrees-of-freedom to serve many user nodes simultaneously in the same time-frequency resource via spatial multiplexing [11], which thus provides higher spectral efficiencies and energy efficiencies. Thus, integration of massive MIMO with MWRN is an important topic [12]–[14]. For instance, the MWRN scheme in [4] is used on a multi-cell massive MIMO system to quantify the achievable rates in the presence of imperfect channel state information (CSI) [12]. For the same system of [12], an efficient transmit power allocation scheme has been proposed In [13].

On the other hand, non-orthogonal multiple access (NOMA) [1], [15], [16] has also emerged as a main enabler of next-generation (5G and beyond) wireless systems. NOMA offers improved spectral efficiencies, reduced latency, and massive connectivity [17]–[21]. In NOMA, multiple users simultaneously access a single spectrum band and via signal multiplexing based on power domain or code domain [16], [22]–[28]. In power-domain NOMA, different power levels are used to differentiate transmit signals of multiple users. Successive interference cancellation (SIC) then enables the decoding of data. SIC for two signals on the downlink operates as follows [15], [29]. Instead of sending two signals separately, the transmitter combines the two signals and transmit the superimposed signal. The receiver decodes the stronger signal first and subtract it from the combined signal. Next, the receiver decodes the difference as the weaker signal. For the general case with more than two signals, multiple signals are combined and sent and the receiver iteratively decode the strongest signal each time and subtract it from the remaining signal [29].

Imperfect SIC, heavily degrades the performance of NOMA systems [24] and thus, modelling this effect has been the focus of several recent papers. Imperfect SIC has been modelled as a portion of the decoded signal in [30] and as the estimation error of the decoded signal in [31], [32]. In this paper, we use the model of [31], [32] to analyze the effect of SIC on the data rate of the proposed NOMA MWRN protocol.

Critically important for NOMA, the power allocation allows different user signals be assigned with different power levels in order to achieve desired performance targets. Power allocation is extensively studied for NOMA systems both in uplink [33] and downlink [34], [35]. Thus, it can be designed to achieve max-min fairness [33], [35], sum rate maximization [34], and energy efficiency maximization [34], [36]. In this paper, we derive it to maximize the minimum achievable sum rate of the MWRN.

B. Motivation

In this paper, we consider the fundamental question of improving the spectral efficiency of a K-user MWRN with full data exchange. In the basic configuration, with an intermediate MIMO relay, this is possible in exactly K orthogonal time slots [4]. Similarly, the MWRN protocols in [12] and [13] require K time slots for K user nodes. Thus, the overall spectral efficiency is low due to the $1/K$ pre-log factor, appearing in achievable rate expressions, and hence it diminishes as the number of users grows. The work done on [37] analyses a relay system which utilizes $K+1$ time slots for one way data transfer from K source nodes to K destination nodes via the support of an intermediate relay. However, it assumes the availability of the direct channel between users and will require $2(K+1)$ time slot for the full data exchange. To attack this fundamental issue, [14] develops a new MWRN transmit protocol to do this in $\lceil (K-1)/2 \rceil + 1$ time slots, where $\lceil x \rceil$ is the ceiling function. This number is roughly $K/2$. Thus, this state-of-the-art protocol [14] halves the number of time slots, equivalently doubling the spectral efficiency. This improvement stems from the adoption of linear processing, self-interference cancellation, and SIC decoding. However, the achievable sum rate of [14] still suffers the $O(1/K)$ decline of spectral efficiency with the number of users.

C. Contributions

Herein, we propose a novel MWRN protocol, which reduces the number of time slots to just two, regardless of the number of users. Thus, it can provide significant spectral efficiency gains compared to those of [12]–[14]. To realize these remarkable gains, we integrate the concepts of power-domain NOMA and massive MIMO with MWRNs.

It is designed as follows. In the first time-slot, all user nodes transmit simultaneously to the relay, which in turn uses maximal ratio combining (MRC) for reception. The relay then generates a superposition-coded signal, applies an amplification factor, and transmits back to the user nodes by using a maximal ratio transmission (MRT) precoder. Then each user node performs SIC to decode the symbols belonging to the remaining $K-1$ users' messages.

More specifically, the contributions of this paper are summarized as follows.

1) We propose a NOMA based massive MIMO MWRN transmit protocol to enable full-mutual data exchange among $K > 2$ users. The key feature is that it uses just two time slots. Thus, it potentially achieves a sum rate gain of $K/2$ (approximately) over the current state-of-the-art counterpart in [14], and the gain scales up with the number of user nodes (K).
2) We obtain closed-form results for the sum rate and the energy efficiency of the proposed scheme by using the so called additive white Gaussian noise (AWGN) approximation [38]. Furthermore, effects of imperfect CSI and SIC is also considered during this analysis.
3) We also obtain the asymptotic sum rate value when the number of antennas at the relay grows unbounded.
4) We propose a power allocation matrix at the relay based upon the asymptotic sum rate values. We are able to get not only the closed-form solution for the matrix based on the asymptotic sum rates of the system, but also ensure max-min fairness even when the number of relay antennas is low.

D. Significance

Proposed protocol achieves a full data exchange among K spatially-distributed user nodes in exactly two time-slots amounting to time-slot reduction of $(1-4/K)\times 100\%$ over the current state-of-the-art MWRN protocol [14]. For example with $K=8$ users, the time-slot saving is 50%. This reduction directly translates into a significant spectral efficiency gain over all existing MWRN counterparts [12]–[14]. This gain is a result of power-domain NOMA and massive MIMO via superposition coding, SIC, and linear detection/precoding.

As number of time slots required in our scheme is constant regardless of number of users, it may be a helpful step in the context of massive connectivity envisaged in 5G and beyond, where a massive number of new IoT devices will connect to a next-generation wireless network. Industry estimates that total number of IoT connected cellular devices will be between 1.6B and 4.6B by 2020. They will generate both data and connection traffic. Clearly, some of these devices could be supported in the MWRN configuration, and the spectral efficiency gains of the proposed protocol may help to mitigate the overall traffic growth.

This present work goes beyond our related conference paper [1], which does not consider imperfectly estimated CSI, imperfect SIC and optimization of relay power allocation matrix.

Notation: $\mathbf{Z}^H$, $\mathbf{Z}^T$, $[\mathbf{Z}]_k$, and $[\mathbf{Z}]_{k,l}$ denote the conjugate transpose, transpose, the k-th row, and the (k,l)-th element of the matrix, $\mathbf{Z}$, respectively. $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote the statistical expectation and variance. A complex Gaussian random variable X with mean μ and standard deviation σ is denoted as $X \sim \mathcal{CN}(\mu, \sigma^2)$. The diagonal matrix $\mathbf{D}$ with kth diagonal element d_k is denoted as $\mathrm{diag}(d_k)$. $\mathbf{I}_M$ and $\mathbf{O}_{M\times N}$ are the $M \times M$ identity matrix and $M \times N$ matrix of all zeros, respectively.

II. System, channel, and signal model

A. System and channel model

We consider a system with K users and denote the k-th user by S_k, $k \in \{1, \cdots, K\}$. They all are single-antenna terminals. The data exchange requirement can be stated as follows. For all $k \in \{1, \cdots, K\}$, S_k must transmit its data to the remaining $K-1$ users and must receive data from all of them too. The relay, R, is equipped with M antennas and is massive MIMO type ($M \gg K$)[1] The direct channels between the user pairs are not available due to unfavorable channel propagation conditions [39], [40] or not utilized in order to minimize mutual interference among the users. Thus, the purpose of R is to accommodate the data transfer among the users.

The wireless channel between S_k and R is represented as $\mathbf{h}_k = \sqrt{\beta_k}\tilde{\mathbf{h}}_k \in \mathcal{C}^M$ where β_k is the large-scale fading

[1]Here, the relay is a specialized wireless node with additional hardware complexity that can accommodate the data exchange among multiple users.

coefficient (may include range-dependent path loss and shadow fading), and $\tilde{\mathbf{h}}_k$ is the small-scale fading vector, which is distributed as

$$\tilde{\mathbf{h}}_k \sim \mathcal{CN}(0, \mathbf{I}_M). \quad (1)$$

The matrix $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \cdots, \tilde{\mathbf{h}}_K] \in \mathcal{C}^{M \times K}$ represents the small-scale channel fading channel coefficients from all the users towards the relay. Thus, this channel matrix incorporates both large-scale and small-scale propagation effects, and it may be expressed as

$$\mathbf{H} = \tilde{\mathbf{H}}\mathbf{D}^{1/2}, \quad (2)$$

where $\mathbf{D} = \mathrm{diag}(\beta_1, \ldots, \beta_K)$.

B. Channel estimation

To design the MRC/MRT detector/precoder ($\mathbf{W}_R$ and $\mathbf{W}_T$), the relay requires CSI between itself and the users. This CSI could be estimated via the conventional estimation techniques. The most common technique is the transmission of sequences of known training symbols (pilots) periodically [11], [41]. Alternatively, blind and semi-blind techniques have also been developed (see [42], [43] and the references therein). Here we consider that the users simultaneously transmit orthogonal pilot sequences to the relay, to avoid mutual interference. The set of orthogonal pilot sequences are given as $\mathbf{\Phi} = \{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_K\}^T$ where $\mathbf{\Phi}_k$ is the $1 \times \tau$ pilot sequence of the k-th user. Here, τ is the length of the pilot sequence used for channel estimation. Because $\mathbf{\Phi}_k$s are mutually orthogonal, the matrix is unitary ($\mathbf{\Phi}\mathbf{\Phi}^H = \mathbf{I}_K$). The received signal at the relay during the pilot transmission period is given as

$$\mathbf{Y}_p = \sqrt{P}\mathbf{H}\mathbf{A}\mathbf{\Phi} + \mathbf{N}_p, \quad (3)$$

where $\mathbf{N}_p$ is the additive white Gaussian noise (AWGN) at the relay with $\mathcal{CN}(0, \mathbf{I}_M)$ distribution and P is the transmit power of the users. Also $\mathbf{A} = \mathrm{diag}(\sqrt{\alpha_{p,1}}, \ldots, \sqrt{\alpha_{p,K}})$ is the power scaling coefficient matrix, where $\alpha_{p,k}$ corresponds to the coefficient used by S_k during the pilot transmission. Relay multiplies the above received signal by $\mathbf{\Phi}^H$ to estimate the channels [41] and obtain

$$\mathbf{y}_k = [\mathbf{Y}_p\mathbf{\Phi}^H]_k = \sqrt{P}\sqrt{\alpha_{p,k}\beta_k}\tilde{\mathbf{h}}_k + \mathbf{n}_{p,k}, \quad (4)$$

where $\mathbf{n}_{p,k} = [\mathbf{N}_p\Phi^H]_k \sim \mathcal{CN}(0, \mathbf{I}_M)$. Based on the above result and by using minimum mean square error (MMSE) criterion [41], the channel estimate for $\mathbf{h}_k$ is given as [17]

$$\hat{\mathbf{h}}_k = \frac{\sqrt{P\alpha_{p,k}\beta_k}}{P\alpha_{p,k}\beta_k + 1}\mathbf{y}_k. \quad (5)$$

For beamforming purposes, we use the estimated channel $\hat{\mathbf{h}}_k$. The true channel $\mathbf{h}_k$ can be written in terms of its estimate by virtue of orthogonality principle of MMSE criterion as [41]

$$\mathbf{h}_k = \hat{\mathbf{h}}_k + \mathbf{e}_k, \quad (6)$$

where $\mathbf{e}_k$ is the error vector, which is independent from $\hat{\mathbf{h}}_k$. The probability distributions of the kth element of estimate ($\hat{\mathbf{h}}_k$) and the error terms ($\mathbf{e}_k$) are $\mathcal{CN}\left(0, \frac{P\alpha_{p,k}\beta_k^2}{P\alpha_{p,k}\beta_k+1}\right)$ and $\mathcal{CN}\left(0, \frac{\beta_k}{P\alpha_{p,k}\beta_k+1}\right)$, respectively.

C. Signal model

Data transmission among the users requires two time slots. In the first time slot, all the users transmit to the relay R, which applies receive beamforming. Thus, for $k = 1, \ldots, K$, user S_k transmits the signal

$$\bar{x}_k = \sqrt{\alpha_k P}x_k, \quad (7)$$

where x_k is the data symbol, P is the allowable maximum transmit power (assumed to be equal for all the users) and $0 < \alpha_k \leq 1$ is the power scaling factor of the k-th user. The received signal at the relay is the sum of all user signals and the additive noise, which is given as

$$\mathbf{y}_r = \sqrt{P}\mathbf{H}\alpha^{1/2}\mathbf{x} + \mathbf{n}_R, \quad (8)$$

where $\mathbf{x} = [x_1, \cdots, x_K]^T$, $\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_K)$, and $\mathbf{n}_R$ is $M \times 1$ additive white Gaussian noise (AWGN) vector at the relay satisfying $\mathbb{E}\left[\mathbf{n}_R^H\mathbf{n}_R\right] = \mathbf{I}_M\sigma_R^2$. Next, the relay apply the receive beamforming by multiplying by the matrix, $\mathbf{W}_R$ and the processed signal can be given as

$$\mathbf{y}_p = \mathbf{W}_R\mathbf{y}_r = \mathbf{W}_R\left(\sqrt{P}\mathbf{H}\alpha^{1/2}\mathbf{x} + \mathbf{n}_R\right). \quad (9)$$

In the second time slot, the relay transmits the following superposition-coded signal to all the users:

$$\mathbf{y}_t = \Psi\mathbf{W}_T\mathbf{\Lambda}\mathbf{y}_p = \Psi\mathbf{W}_T\begin{bmatrix}\lambda_{1,1} & \lambda_{1,2} & \cdots \\ \lambda_{1,2} & \lambda_{2,2} & \cdots \\ \vdots & \ddots & \vdots \\ \lambda_{K,1} & \cdots & \lambda_{K,K}\end{bmatrix}\mathbf{W}_R\mathbf{y}_r, \quad (10)$$

where $\mathbf{W}_T$ is the MRT precoder at the relay, $\mathbf{\Lambda}$ is the $K \times K$ power allocation matrix at the relay, and Ψ is the power control factor. The selection of MRC and MRT at the relay is due to the simplicity of those methods compared to other linear beamformers. However, similar results can be obtained for other beamforming methods such as zero forcing and omitted in this paper due to space limitations. The total power constraint at the relay is given as

$$\begin{aligned} P_R &= \mathrm{Tr}\left(\mathbf{y}_t\mathbf{y}_t^H\right) \\ &= \Psi^2 P\mathrm{Tr}\left(\mathbf{V}\mathbf{H}\alpha\mathbf{H}^H\mathbf{V}^H\right) + \Psi^2\sigma_R^2\mathrm{Tr}\left(\mathbf{V}\mathbf{V}^H\right), \end{aligned} \quad (11)$$

where P_R is the transmit power at the relay and we define $\mathbf{V} = \mathbf{W}_T\mathbf{\Lambda}\mathbf{W}_R$ as the effective/cascaded detector/precoder at the relay, and the relay gain Ψ is computed to constrain the average transmit power as

$$\Psi = \sqrt{\frac{P_R}{P\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}\mathbf{H}\alpha\mathbf{H}^H\mathbf{V}^H\right)\right] + \sigma_R^2\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}\mathbf{V}^H\right)\right]}}. \quad (12)$$

Based on (10), the intended transmit signal for S_k is given as $[\mathbf{y}_t]_k$, which is the kth row of $\mathbf{y}_t$. The received signal at S_k is given as

$$y_k = \Psi\mathbf{h}_k^T\mathbf{V}\sum_{m=1}^{K}\sqrt{P\alpha_m}\mathbf{h}_m x_m + \Psi\mathbf{h}_k^T\mathbf{V}\mathbf{n}_R + n_k, \quad (13)$$

where n_k is an AWGN at S_k with power σ_k^2.

$$\hat{\mathbf{n}}_{\mathbf{k},\mathbf{n}} = \underbrace{\Psi\sqrt{P\alpha_{f_k(n)}}\left(\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)} - \mathbb{E}\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\right]\right)}_{\text{detection uncertainty}} + \underbrace{\sum_{m'=1}^{n-1}\Psi\sqrt{P\alpha_{f_k(m')}}\left(\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(m')}x_{f_k(m')} - \mathbb{E}\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(m')}\right]\hat{x}_{f_k(m'),k}\right)}_{\text{effect of imperfect SIC}}$$
$$+ \underbrace{\Psi\mathbf{h}_k^T\mathbf{V}\sum_{m=n+1}^{K-1}\sqrt{P\alpha_{f_k(m)}}\mathbf{h}_{f_k(m)}x_{f_k(m)}}_{\text{interference from other users}} + \underbrace{\Psi\mathbf{h}_k^T\mathbf{V}\mathbf{n}_R}_{\text{amplified noise}} + \underbrace{n_k}_{\text{AWGN noise at the receiver}}. \tag{16}$$

$$\mathcal{R}_{k,f_k(n)} = \frac{(T_C-\tau)}{2T_C}\log\left(1+\frac{\Psi^2 P\alpha_{f_k(n)}M_{k,n}^2}{\Psi^2\left(P\alpha_{f_k(n)}N_{k,n}+\sum_{m=n+1}^{K-1}P\alpha_{f_k(m)}P_{k,m}+\sum_{m'=1}^{n-1}P\alpha_{f_k(m')}R_{k,m'}+\sigma_R^2Q_k\right)+\sigma_k^2}\right). \tag{18}$$

D. Imperfect SIC decoding at the user nodes

After receiving y_k, S_k will decode the symbols (i.e., all $x_{k'}$ for $k' \neq k$ and $k' \in \{1, \cdots, K\}$) transmitted by all other users. It will use SIC decoding for this, which is an iterative process. In each step, a symbol x_l is decoded, and subtracted (i.e., cancelled) from y_k. Due to K users in total, $K-1$ decoding steps occur at each user. The index of the user that has to be decoded at the n-th iteration at S_k is denoted by the function $f_k(n)$. Thus, $1 \leq n \leq K-1$ and $1 \leq f_k(n) \leq K$ with $f_k(n) \neq k$.

At the nth decoding step, S_k will decode the signal transmitted by user $S_{f_k(n)}$, denoted as $x_{f_k(n)}$. The estimate of $x_{f_k(n)}$ at S_k is represented by $\hat{x}_{k,f_k(n)}$. Both of these are assumed to be jointly Gaussian distributed with a normalized correlation coefficient of $\rho_{k,f_k(n)}$ and may be expressed as [31]

$$\hat{x}_{k,f_k(n)} = \rho_{k,f_k(n)}x_{f_k(n)} + e_{k,f_k(n)}, \tag{14}$$

where the estimate $\hat{x}_{k,f_k(n)} \sim \mathcal{CN}(0,1)$, the estimation error $e_{k,f_k(n)} \sim \mathcal{CN}\left(0, \sigma^2_{e_{f_k(n)}}/\sqrt{1+\sigma^2_{e_{f_k(n)}}}\right)$, and the correlation coefficient $\rho_{k,f_k(n)} = 1/\sqrt{1+\sigma^2_{e_{f_k(n)}}}$. Furthermore, the estimation error and the estimated value are statistically independent. The perfect SIC case is specified by the values $\rho_{k,f_k(n)} = 1$ and $\sigma^2_{e_{f_k(n)}} = 0$. In the next section, we will investigate the effect of imperfect SIC on the system performance.

III. Achievable sum rate analysis

Next, we derive the achievable sum rate between each user pairs by using the worst-case Gaussian technique [38] that can be summarized as follows. The received signal at a user node can be decomposed into a desired signal and effective noise. The desired signal consists of the intended message symbol, while effective noise component contains all interference terms and AWGN. It can easily be shown that the desired signal component and effective noise term are uncorrelated. Moreover, each term within the effective noise is uncorrelated. Thus, this effective noise term can be approximated by using a Gaussian random variable with same variance, which represents the worst-case scenario [38]. Based on this framework, we derive the sum rate as follows. The residual signal at S_k after decoding canceling signals belonging to $n-1$ users can be written as

$$y_{k,n} = \Psi\sqrt{P\alpha_{f_k(n)}}\mathbb{E}\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\right] + \hat{\mathbf{n}}_{\mathbf{k},\mathbf{n}}, \tag{15}$$

where $\hat{\mathbf{n}}_{\mathbf{n},\mathbf{k}}$ is the effective noise and the first term is the desired signal. The function $f_k(n)$ is the index of the user that will be decoded in the n-th step. Moreover, the noise term in (15) can be expressed as (16) at the top of the page. The matrix $\mathbf{V}$ in (15) and (16) for MRC/MRT beamforming is given as

$$\mathbf{V} = \hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H. \tag{17}$$

Based on (15) and (16), and by assuming additive noise as independently distributed Gaussian noise having the same variance [38], a tight approximation for the achievable sum rate can be given as (18) at the top of the page. Here T_C is the coherence time of the channel. The pre-log factor $(T_C-\tau)/T_C$ accounts for the pilot overhead [11]. The two time-slots required for the data transmission between the users results in the pre-log factor of $1/2$. The values of $M_{k,m}$, $N_{k,m}$, $P_{k,m}$, $R_{k,m}$, and Q_k in (18) are given as

$$M_{k,m} = \mathbb{E}\left[\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}\right], \tag{19a}$$
$$N_{k,m} = \mathbb{V}\left[\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}\right], \tag{19b}$$
$$P_{k,m} = \mathbb{E}\left[|\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}|^2\right], \tag{19c}$$
$$R_{k,m} = \mathbb{E}\left[|\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}|^2\right] + (1-2\rho_{k,f_k(m)})\mathbb{E}\left[\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}\right]^2, \tag{19d}$$
$$Q_k = \mathbb{E}\left[\|\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\|^2\right]. \tag{19e}$$

The value of $R_{k,m}$ can further be written as

$$R_{k,m} = P_{k,m} + (1-2\rho_{k,f_k(m)})M_{k,m}^2. \tag{20}$$

The closed-form evaluations of (19) are provided in Appendix A. The value for Ψ is given as

$$\Psi = \sqrt{\frac{P_R}{PL_1+\sigma_R^2L_2}}, \tag{21}$$

$$\mathcal{R}_{k,\hat{n}}^{\infty}=\frac{(T_C-\tau)}{2T_C}\log\left(1+\frac{P\left(\Psi^{\infty}\right)^2\alpha_{\hat{n}}\lambda_{k,\hat{m}}^2\hat{\beta}_k^2\hat{\beta}_{\hat{n}}^2}{P\left(\Psi^{\infty}\right)^2\left(\sum\limits_{m'=1}^{n-1}\left(1-\rho_{k,\hat{m}'}\right)^2\alpha_{\hat{m}'}\lambda_{k,\hat{m}'}^2\hat{\beta}_k^2\hat{\beta}_{\hat{m}'}^2+\sum\limits_{m=n+1}^{K-1}\alpha_{\hat{m}}\lambda_{k,\hat{m}}^2\hat{\beta}_k^2\hat{\beta}_{\hat{m}}^2\right)+\sigma_k^2}\right), \tag{34}$$

where L_1 and L_2 are given as

$$L_1=\mathrm{Tr}\left(\mathbb{E}\left[\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{H}\alpha\mathbf{H}^H\hat{\mathbf{H}}\mathbf{\Lambda}^H\hat{\mathbf{H}}^T\right]\right), \tag{22a}$$

$$L_2=\mathrm{Tr}\left(\mathbb{E}\left[\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\hat{\mathbf{H}}\mathbf{\Lambda}^H\hat{\mathbf{H}}^T\right]\right), \tag{22b}$$

and are derived in Appendix B.

The rate of data transmission for each user is limited by the achievable sum rates among itself and all other users. Thus, the achievable transmission rate of S_m is given as

$$\mathcal{R}_m=\min_{k\in(1,\cdots,K),k\neq m}\left(\mathcal{R}_{k,m}\right). \tag{23}$$

Based on this, the total achievable sum rate of the system is obtained as

$$\mathcal{R}=(K-1)\sum_{m=1}^{K}\mathcal{R}_m. \tag{24}$$

Here in (24), the factor $(K-1)$ represents the transmission of each users data to all the other users.

Remark 1. *The results obtained in (18) and (24) are valid under any power allocation matrix $\mathbf{\Lambda}$ and any decoding order $f_k(n)$. Thus these closed form solutions can be utilized to design power allocation schemes to obtain different user requirements in MWRNs.*

IV. Asymptotic sum rate analysis

Asymptotic refers to the fact that the number of relay antennas M grows unbounded. The significance of this condition is that the relay can then scale down the transmit power inversely proportional to the number of antennas [44], which tends to improve energy efficiency overall. Therefore, it is important to find the sum rate under this condition. The relay transmit power may thus be expressed as

$$P_R=E_R/M, \tag{25}$$

where E_R corresponds to the total transmit power available at the relay node. We first derive an asymptotic limit for Ψ, the transmit power control factor at the relay. To obtain Ψ, we need to find the limits of L_1 and L_2 for extremely large values of M. For this, we use the generalized forms of the law of large numbers. The details are given in Appendix D and using the fact that $\hat{\mathbf{H}}$ and $\mathbf{E}$ are independent of each other, we obtain the following asymptotic result:

$$\frac{L_1}{M^3}\xrightarrow[M\to\infty]{a.s.}\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_j\lambda_{i,j}^2\hat{\beta}_i\hat{\beta}_j^2. \tag{26}$$

Here, $\hat{\beta}_k$ values are defined as (58) in Appendix A. Similarly, an asymptotic limit for L_2 is derived as

$$\frac{L_2}{M^2}\xrightarrow[M\to\infty]{a.s.}\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_{i,j}^2\hat{\beta}_i\hat{\beta}_j. \tag{27}$$

By using (26) and (27), the limit for Ψ can be obtained as

$$M^2\Psi\xrightarrow[M\to\infty]{a.s.}\sqrt{\frac{E_R}{E\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_j\lambda_{i,j}^2\hat{\beta}_i\hat{\beta}_j^2}}=\Psi^{\infty}. \tag{28}$$

By using the above asymptotic results and the value for $\mathbf{V}$, we rewrite (15) as

$$\begin{aligned}y_{k,n}=&M^2\Psi\sqrt{P\alpha_{f_k(n)}}\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{h}_{f_k(n)}}{M}x_{f_k(n)}\\&+M^2\Psi\sum_{m=n+1}^{K-1}\sqrt{P\alpha_{f_k(m)}}\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}}{M}x_{f_k(m)}\\&+M^2\Psi\sum_{m'=1}^{n-1}\sqrt{P\alpha_{f_k(m')}}\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m')}}{M}\left(x_{f_k(m')}-\hat{x}_{k,f_k(m')}\right)\\&+M^2\Psi\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{n}_R}{M}+n_k.\end{aligned} \tag{29}$$

We then derive the asymptotic results for each term in (29) as follows:

$$\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{h}_{f_k(n)}}{M}x_{f_k(n)}\xrightarrow[M\to\infty]{a.s.}\lambda_{k,f_k(n)}\hat{\beta}_k\hat{\beta}_{f_k(n)}, \tag{30}$$

$$\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)}}{M}x_{f_k(m)}\xrightarrow[M\to\infty]{a.s.}\lambda_{k,f_k(m)}\hat{\beta}_k\hat{\beta}_{f_k(m)}, \tag{31}$$

$$\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{n}_R}{M}\xrightarrow[M\to\infty]{a.s.}0. \tag{32}$$

The asymptotic limit of the effect of imperfect SIC can be written as

$$\begin{aligned}&\frac{\mathbf{h}_k^T\hat{\mathbf{H}}^*}{M}\mathbf{\Lambda}\frac{\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m')}}{M}\left(x_{f_k(m')}-\hat{x}_{k,f_k(m')}\right)\\&\qquad\xrightarrow[M\to\infty]{a.s.}\left(1-\rho_{k,f_k(m')}\right)\lambda_{k,f_k(m)}\hat{\beta}_k\hat{\beta}_{f_k(m)}.\end{aligned} \tag{33}$$

By using the aforementioned asymptotic results, we derive the asymptotically achievable sum rate as (34) at the top of the page. Here, in (34), $\hat{m}=f_k(m)$.

Remark 2. *Similar to the non asymptotic case, the result in (34) is valid under any power allocation matrix $\mathbf{\Lambda}$ and any decoding order $f_k(n)$.*

$$\mathcal{R}_{k,\hat{n}}=\frac{(T_C-\tau)}{2T_C}\log\left(1+\frac{\Psi^2\alpha_{\hat{n}}\left(\lambda_{k,\hat{n}}^2A_{k,\hat{n}}+\lambda_{k,\hat{n}}\lambda_{\hat{n},k}B_{k,\hat{n}}+\lambda_{\hat{n},k}^2C_{k,\hat{n}}\right)}{\Psi^2\alpha_{\hat{n}}\sum\limits_{i=1}^{K}\sum\limits_{j=1}^{K}\left(\lambda_{i,\hat{n}}\lambda_{j,\hat{n}}D_{i,j,\hat{n}}+\lambda_{i,k}\lambda_{j,k}D_{i,j,k}+\lambda_{\hat{n},i}\lambda_{\hat{n},j}E_{i,j,\hat{n}}+\lambda_{k,i}\lambda_{k,j}E_{i,j,k}\right)+\mathcal{Z}}\right). \tag{40}$$

$$\mathcal{Z}=\sum_{m=n+1}^{K-1}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\lambda_{i,m}\lambda_{j,m}G_{i,j,m}+\lambda_{i,k}\lambda_{j,k}G_{i,j,k}+\lambda_{m,i}\lambda_{m,j}H_{i,j,m}+\lambda_{k,i}\lambda_{k,j}H_{i,j,k}\right)+\sigma_k^2, \tag{41}$$

V. Energy efficiency of the system

In this section, we analyze the energy efficiency of the proposed MWRN. In recent years, energy efficiency, which is the number of information bits per unit of transmit energy, has been studied extensively [41]. It is defined as

$$\rho=\frac{\mathcal{R}}{P_{Tot}}, \tag{35}$$

where P_{tot} is the total power consumption and $\mathcal{R}$ is the overall sum rate. The value for P_{tot} can be written as follows [45]

$$P_{Tot}=KP+P_R+P_{U,TC}+KP_{R,TC}+P_{C/D}+P_{R,LP}, \tag{36}$$

where $P_{U,TC}$ and $P_{R,TC}$ are the power consumed in transceiver chains in each user and the relay, $P_{C/D}$ is the power consumed for coding and decoding, and $P_{R,LP}$ is the power consumed in the relay to perform for linear processing. The values for these power consumption components are given as follows [45]:

$$P_{U,TC}=2P_{U,C}+2P_{SYN}\qquad P_{R,TC}=2MP_{R,C}+2P_{SYN}, \tag{37}$$

$$P_{C/D}=2\mathcal{R}(P_{COD}+P_{DEC}), \tag{38}$$

$$P_{R,LP}=\frac{2BKM}{L_{BS}}+\frac{B}{U}\frac{3MK}{L_{BS}}, \tag{39}$$

where $P_{U,C}$ and $P_{R,C}$ are the powers required to run the circuit components at the users and the relay, P_{SYN} is the power of the local oscillator, P_{COD} and P_{DEC} is the coding and decoding power consumption, B is the bandwidth, L_{BS} is the computational efficiency (given in flops/W) of the end nodes, and U is the coherence block.

VI. Design of power allocation matrix

In this section, we analyze the design of $\mathbf{\Lambda}$ while satisfying the relay power constraints. By changing the values in $\mathbf{\Lambda}$, different sum rates can be achieved for different users in the system. Several optimization problems can be formulated to design $\mathbf{\Lambda}$ based on different criteria such as maximizing the minimum data rate (max-min fairness), maximizing the total sum rate etc. The objective functions of these problems are constrained by three variables, namely $\mathbf{\Lambda}, \boldsymbol{\alpha}$, and $f_k(n)$. Here, $\mathbf{\Lambda}$ is a $K\times K$ matrix, while $\boldsymbol{\alpha}=\{\alpha_n,\ldots,\alpha_K\}$ is a vector of length K. Furthermore, $f_k(n)$ is a $\mathbb{R}^2\rightarrow\mathbb{R}$ function which can be represented by $K\times(K-1)$ matrix. The achievable rates that are used in the optimizing problems under the worst-case Gaussian technique can be written as (40) at the top of the page. Here, $\mathcal{Z}$ is used to display (40) in a manageable way and is given as (41) at the top of the page. Here, $A_{k,\hat{n}}$, $B_{k,\hat{n}}$, $C_{k,\hat{n}}$, $D_{i,j,\hat{n}}$, $E_{i,j,\hat{n}}$, $G_{i,j,\hat{n}}$, and $H_{i,j,\hat{n}}$ are functions of M.

As evident from (40), the sum rate between each pair is a complex function of $\boldsymbol{\alpha}$, $\mathbf{\Lambda}$, and Ψ. Furthermore, the second degree terms of $\mathbf{\Lambda}$ components appear on the sum rate equation. Due to these reasons, solving optimizing problems involving (40) appear intractable.

To overcome this problem and to formulate solvable optimization problems, we take the following three steps. (1) the above sum rates are replaced by their asymptotic values, (2) the asymptotic rates are simplified by assuming perfect SIC, and (3) the decoding order functions $f_k(n)$ are determined according to the large-scale fading coefficients. The simplified asymptotic sum rate is obtained as follows:

$$\mathcal{R}_{k,f_k(n)}^{\infty}=\frac{1}{2}\frac{(T_C-\tau)}{T_C}\log\left(1+\frac{\lambda_{k,f_k(n)}^2M_{k,n}}{\sum\limits_{m=n+1}^{K-1}\lambda_{k,f_k(m)}^2M_{k,m}+\sigma_k^2}\right), \tag{42}$$

where $M_{k,n}=E\alpha_{f_k(n)}\hat{\beta}_k^2\hat{\beta}_{f_k(n)}^2$. This is obtained by removing the terms corresponding to the imperfect SIC in (34) and replacing Ψ with 1. The decoding order is defined as follows:

$$f_k(n)=\begin{cases}n & n<k\\ n+1 & n\geq k,\end{cases} \tag{43}$$

where the users are ordered according to the descending order of large-scale fading coefficients between them and the relay (i.e., $\beta_1\geq\beta_2\geq\cdots\geq\beta_K$). Based on the above simplifications, we present the max-min fairness power allocation optimization problem as follows:

$$\underset{\lambda}{\text{maximize}}\ \min\left(\frac{(T_C-\tau)}{2T_C}\log\left(1+\frac{\lambda_{k,f_k(n)}^2M_{k,f_k(n)}}{\sum\limits_{m=n+1}^{K-1}\lambda_{k,f_k(m)}^2M_{k,f_k(m)}+\sigma_k^2}\right)\right) \tag{44}$$

$$\text{subject to}\quad E\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_j\lambda_{i,j}^2\hat{\beta}_i\hat{\beta}_j^2\leq E_R. \tag{45}$$

$$\lambda_{i,j}\geq 0 \tag{46}$$

The constraint (45) is the power constraint at the relay. It can be proven that the maximum value for (44) can be obtained when the data rates between all the users are equal to each other and when the inequality (45) becomes an equality (i.e., when the relay uses the maximum available power for the transmission). Also it can be shown that the result obtained

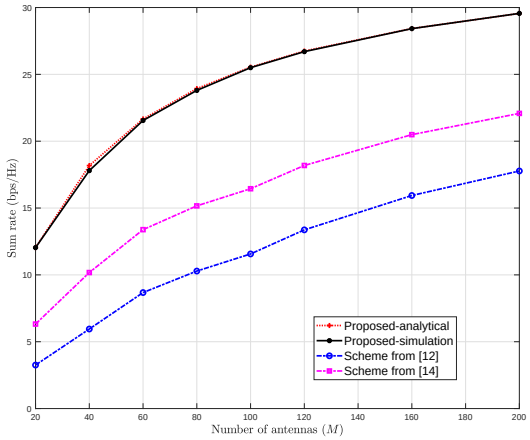

Fig. 1: Total sum rate for $K = 8$ against the number of relay antennas.

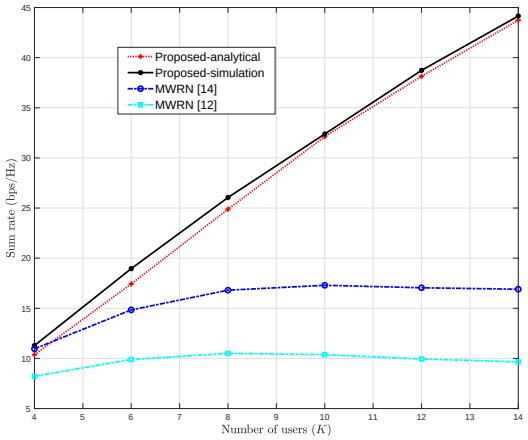

Fig. 2: Average sum rate versus K for $M = 64$.

for this case lies in the feasibility region of (44). Assuming that the SINR between each user reaches a common SINR $\bar{t}$, we can obtain the optimal values for $\lambda_{k,f_k(n)}$'s as follows for $n = K-1, \ldots, 1$:

$$\lambda^2_{k,f_k(n)} = \frac{\bar{t}\left(\sum_{m=n+1}^{K-1} \lambda^2_{k,f_k(m)} M_{k,f_k(m)} + \sigma_k^2\right)}{M_{k,f_k(n)}}. \quad (47)$$

The values for $\lambda_{k,f_k(n)}$ can be found by solving (47) starting from $n = K-1$ to $n = 1$ for each k value. By observing the pattern, a generalized expression for $\lambda^2_{k,f_k(n)}$ can be written as

$$\lambda^2_{k,f_k(n)} = \frac{\bar{t}\sigma_k^2\left(\bar{t}+1\right)^{K-n}}{M_{k,f_k(n)}} = \frac{\bar{t}\sigma_k^2\left(\bar{t}+1\right)^{K-n}}{E\alpha_{f_k(n)}\hat{\beta}_k^2\hat{\beta}^2_{f_k(n)}}. \quad (48)$$

Then by using those obtained values on (45), we derive

$$\sum_{i=1}^{K} \frac{\sigma_i^2}{\hat{\beta}_i} \sum_{j=1}^{K-1} \bar{t}\left(\bar{t}+1\right)^{K-j-1} = E_R. \quad (49)$$

This can be simplified as follows and the value for $\bar{t}$ can be written as

$$\sum_{j=1}^{K-1} \bar{t}\left(\bar{t}+1\right)^{K-j-1} = \left(\bar{t}+1\right)^{K-1} - 1 = \frac{E_R}{\sum_{i=1}^{K} \frac{\sigma_i^2}{\hat{\beta}_i}}. \quad (50)$$

The non-negative real solution for the above polynomial can be derived as

$$\bar{t} = \sqrt[K-1]{1 + \frac{E_R}{\sum_{i=1}^{K} \frac{\sigma_i^2}{\hat{\beta}_i}}} - 1. \quad (51)$$

Remark 3. *It can be seen that the asymptotically achievable sum rate under max-min fairness does not depend on the values of transmit power at the user nodes (i.e., P and α). Furthermore, our simulations show that it is independent of the decoding order at the user (i.e., $f_k(n)$).*

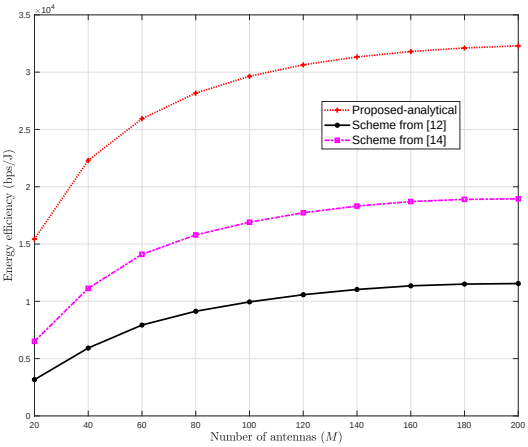

Fig. 3: Average energy efficiency of the system vs M for $K = 12$.

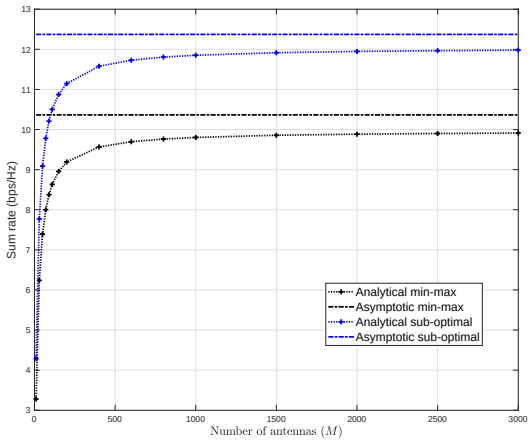

Fig. 4: Average sum rate under different power allocation schemes.

VII. Numerical Results

In this section, we investigate the performance gains of the proposed NOMA scheme via simulations. Apart from the power allocation obtained in Section VI, we use the following

sub-optimal power allocation matrix for comparisons.

$$\mathbf{\Lambda} = \mathbf{D}^{-1}\left(\mathbf{B} + \mathbf{B}^{T}\right), \tag{52}$$

where $\mathbf{B}$ is the $K \times K$ matrix with $[\mathbf{B}]_{i,j} = \sqrt{\frac{j-1}{K}}$ when $i < j$ and zero otherwise. This power allocation is based on the following observations.

- Each row in $\mathbf{\Lambda}$ corresponds to the power allocation factors to each user. Thus, to compensate for the downlink path-losses, we allocate more power to the users which have the highest path-loss by multiplying each row of $\mathbf{\Lambda}$ by the inverse of the path-loss component of each user.
- The ratios between the non-zero coefficients in a single row determines the data rate between the users. Here, we designed the ratios in the form of $\sqrt{1/K}, \sqrt{2/K}, \cdots, \sqrt{K-1/K}$.

A. Comparison with MWRN operations in [12], [14]

Here, we analyze the existing MWRN protocols in [12], [14] for numerical comparison purposes. Firstly, we consider the MWRN in [12], which requires K time slots to transmit the data of all the users to all other users. The use of time-slots in [12] can be listed as follows:

1) **Time-slot 1:** All the users transmit to the relay. This step is similar to the first step in our proposed NOMA-aided protocol.
2) **Time-slot 2 to K:** In these time slots, relay transmits to all the users using beamforming. However, instead of sending a superposition-coded signal of all the other received signals, data of a single user is transmitted to each user. Thus, $K-1$ time slots are required to send the data of all the users to all other users.

The beamforming matrix at the relay for the j-th time slot $(2 \leq j \leq K)$ is given as

$$\mathbf{V}_j = \mathbf{H}^{*}\mathbf{\Lambda}_j\mathbf{H}^{H}. \tag{53}$$

In (53), Λ_j is a permutation matrix in which each row consists of a single one and all zeros. The location of the number one, decides the transmitted signal of the initial set of users. We can obtain the approximation for the end-to-end data rate between each pair of users by using the same steps as the previous case. However, when calculating the achievable data rate, we have to use the pre-log factor $1/K$, as K total time slots are required for the data transmission.

Secondly, we compare the performance of the proposed MWRN with that of [14], which utilizes $\lceil (K-1)/2 \rceil + 1$ time-slots. The above two methods (i.e. [12] and [14]) are used as performance benchmarks.

Fig. 1 plots the achievable sum rate of the proposed MWRN system and those of [12] and [14] against the number of relay antennas. We use the power allocation matrix (52) with eight users ($K = 8$) and path-loss components $\mathbf{D} = \mathrm{diag}(1,\ 0.875,\ 0.75,\ 0.625,\ 0.5,\ 0.375,\ 0.25,\ 0.125)$, and $\alpha_k = 1$ for $1 \leq k \leq K$. The proposed scheme clearly provides a higher achievable sum rate compared to the other two. For instance, with 100 relay antennas, it provides a sum rate of 25.2 bps/Hz while [12] and [14] achieve only 12.2 bps/Hz and 17.1 bps/Hz. This amounts to 106% and 47% gains respectively. Furthermore, we observe that more relay antennas increase the sum rate of the proposed system.

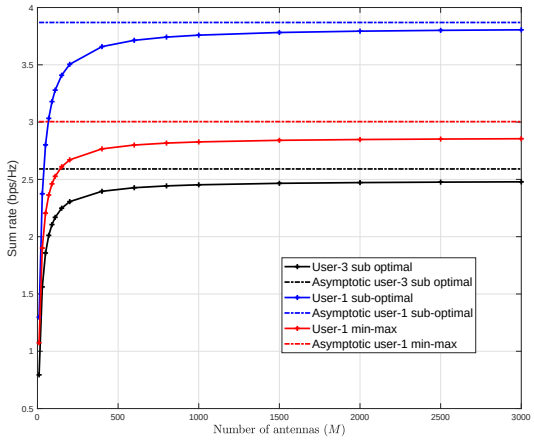

Fig. 5: Average sum rate for individual users with different power allocation schemes.

Fig. 2 plots the achievable sum rates for the proposed system **given by (18)** and the two other MWRN schemes against the number of users (K). While the achievable sum rate reaches a constant value under [12] and [14], our method provides constant sum rate improvements as the number of users are increased. For instance, both our scheme and [14] has the same sum rate performance with 4 users. But with 10 users, our scheme provides a sum rate of 33 bps/Hz, while [14] and [12] only provides 16 bps/Hz and 10 bps/Hz, respectively. Furthermore, the performance gap increases as the number of users are increased. As an example, with 8 users, our system provides 47% and 127% increase of sum rate compared to [14] and [12] while this gain increases to 123% and 280% for 12 users.

In Fig. 3, we plot the energy efficiency in Eqn. (35) against the number of relay antennas M for $K = 12$ users. The values for $P_{U,C}$, $P_{R,C}$, P_{SYN}, P_{COD}, P_{DEC}, B, U, and L_{BS} are adopted from [45]. It can be seen that our proposed protocol provides higher energy efficiency compared to two other methods. As an example, with 100 relay antennas, our energy efficiency is 2.94×10^4 bps/J while the energy efficiency of [12] is only 1×10^4 bps/J. This is almost a 300% increase. This gain is both due to the sum rate increase as well as lower transmit powers due to the reduced number of time slots.

In Fig. 4, we plot the total sum rate of the system under the proposed two power allocation schemes namely the max-min power allocation and the sub-optimal power allocation for four users (i.e., $K = 4$ and $\mathbf{D} = \mathrm{diag}(1, 0.75, 0.5, 0.25)$) with $\tau = 0.3$ and perfect SIC scenario. We obtained the sum rate from (24) and (18) for different M values and the asymptotic value from (34). The figure shows that the sub-optimal power allocation results in slightly higher total sum rate than the min-max fairness power allocation. Also this shows that the asymptotic result (34) is accurate.

In Fig. 5, we plot the individual sum rate for user 1 and user 3 for the same system setup. It can be seen that the

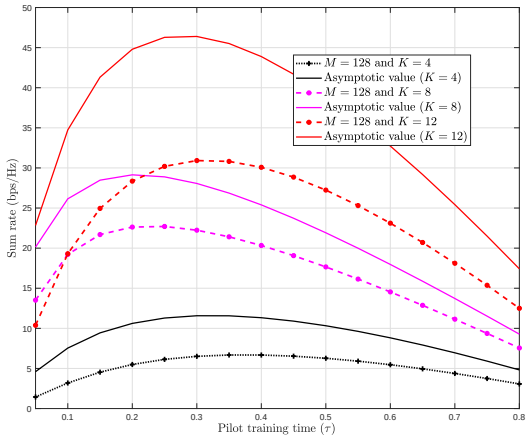

Fig. 6: Average total sum rate for different CSI settings.

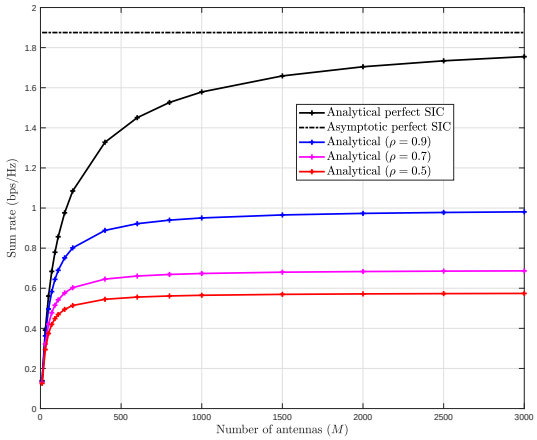

Fig. 7: Sum rate of user-1 under different SIC conditions against the number of antennas.

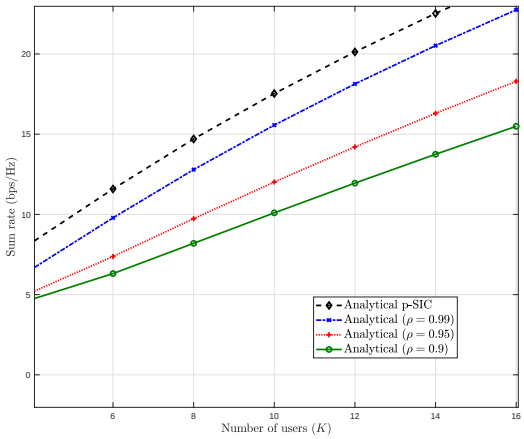

Fig. 8: Average total sum rate for different SIC conditions against the number of users.

max-min power allocation scheme obtains 2.9 bps/Hz sum rate for all the users, while the sum rate obtained from the sub optimal power allocation scheme differs for different users. As an example, in this case user 1 (the user nearest to the relay) obtains 3.9 bps/Hz while the user 3 (user far away from the relay) only obtain 2.5 bps/Hz. Thus, this shows that the proposed power allocation can provide fairness to all users regardless of the distance between the relay and the users. This will be useful when all the users have to be treated equally.

In Fig. 6, we analyze the effect of CSI availability on the performance of the system. Specifically, we plot the achievable sum rate with 128 antennas and the asymptotic sum rate for three different K values (i.e., $K = 4$, $K = 8$, and $K = 12$) against the pilot training time (τ). Note that sum rates increase with τ up to a certain point and then starts to decrease. As an example, for $K = 8$, $\tau = 0.1$ provides 19 bps/Hz and increasing τ to 0.2 provides 22.3 bps/Hz. However after this, the achievable sum rate starts to decrease. A lower value for τ results in poorly estimated channel and lead to lower sum rates, while a higher value for τ will limit the time used for data transmission and also result in lower sum rate for the system. According to Fig. 6, the optimum τ values for $K = 4$, $K = 8$, and $K = 12$, are $\tau = 0.35$, $\tau = 0.2$, and $\tau = 0.3$ respectively for $M = 128$. This shows that the existence of optimum τ value based on the number of antennas, number of users, and other system parameters. However, we do not attempt this optimization of τ, which is left it as a future research topic.

Next, in order to analyze the effect of imperfect SIC, we plot the sum rate of user 1 against the number of relay antennas in Fig. 7. Here, we have assumed 8 users and $\tau = 0.3$. The sum rate is plotted under four scenarios; namely perfect SIC, imperfect SIC with $\rho_{m,k} = 0.9$, $\rho_{m,k} = 0.7$, and $\rho_{m,k} = 0.5$ for all m, k values. As evident from this figure, when SIC is free from error propagation, the system sum rate increases significantly. As an example, with 500 antennas, user 1 obtains 1.4 bps/Hz with perfect SIC, but only 0.9 bps/Hz when $\rho = 0.9$. This shows the significant effect of SIC on NOMA systems.

To analyze this detrimental impact of imperfect SIC further, we plot the total sum rate against the number of users K in Fig. 8. Here we look at three different values for ρ along with the perfect SIC scenario. It can be seen that imperfect SIC reduces the achievable sum rate of a system. As an example with 6 users, the sum rate of the system will degrade to 7.5 bps/Hz from 11.9 bps/Hz when imperfect SIC is present with $\rho = 0.95$. This value further decreases to 6.8 bps/Hz when $\rho = 0.9$. This shows the importance of accurate SIC in NOMA systems.

VIII. Conclusion

We proposed a NOMA-aided massive MIMO MWRN, which enables data exchange between K users within only two time slots. This is a drastic reduction compared to $\lceil (K+1)/2 \rceil + 1$ time slots of the current state-of-the-art in MWRNs. In the first time slot, all user nodes transmit simultaneously to the relay, which in turn applies a linear MRC detector. In the second time slot, for each user, the relay constructs a superposition-coded signal consisting of data symbols belonging to all other users to be transmitted by using linear MRT precoding. Upon receiving this superposition-coded signal, users adopt SIC to decode the data from each

$$\bar{N}_{k,\hat{m}} = M(M+1)\left(\lambda_{k,k}^2\hat{\beta}_k^2\left(\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)\left((M+2)\hat{\beta}_k+\hat{\eta}_k\right)+\lambda_{\hat{m},\hat{m}}^2\hat{\beta}_{\hat{m}}^2\left(\hat{\beta}_k+\hat{\eta}_k\right)\left((M+2)\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)\right)$$
$$+M\hat{\beta}_{\hat{m}}\hat{\beta}_k\left((M+1)\hat{\beta}_{\hat{m}}\hat{\beta}_k\left(\lambda_{k,\hat{m}}^2M(M+1)+2\lambda_{\hat{m},k}^2\right)+(M+1)\left(\hat{\eta}_{\hat{m}}\hat{\beta}_k+\hat{\beta}_{\hat{m}}\hat{\eta}_k\right)\left(\lambda_{k,\hat{m}}^2M+\lambda_{\hat{m},k}^2\right)\right.$$
$$\left.+M\hat{\eta}_{\hat{m}}\hat{\eta}_k\left(\lambda_{k,\hat{m}}^2+\lambda_{\hat{m},k}^2\right)\right)+2\lambda_{k,\hat{m}}\lambda_{\hat{m},k}M\hat{\beta}_{\hat{m}}\hat{\beta}_k\left((M+1)\hat{\beta}_k+\hat{\eta}_k\right)\left((M+1)\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)$$
$$+\sum_{\substack{i=1\\i\neq k,\hat{m}}}^{K}\sum_{\substack{j=1\\j\neq k,\hat{m},i}}^{K}\lambda_{i,j}^2M^2\hat{\beta}_i\hat{\beta}_j\beta_k\beta_{\hat{m}}+\sum_{\substack{i=1\\i\neq k,\hat{m}}}^{K}\lambda_{i,i}^2M(M+1)\hat{\beta}_i^2\beta_k\beta_{\hat{m}}$$
$$+2\sum_{\substack{j=1\\j\neq k,\hat{m}}}^{K}\lambda_{k,j}\lambda_{j,k}M\hat{\beta}_k\hat{\beta}_j\left(\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)\left((M+1)\hat{\beta}_k+\hat{\eta}_k\right)+2\sum_{\substack{j=1\\j\neq k,\hat{m}}}^{K}\lambda_{\hat{m},j}\lambda_{j,\hat{m}}M\hat{\beta}_{\hat{m}}\hat{\beta}_j\left(\hat{\beta}_k+\hat{\eta}_k\right)\left((M+1)\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)$$
$$+\sum_{\substack{j=1\\j\neq k,\hat{m}}}^{K}\lambda_{k,j}^2M^2\hat{\beta}_k\hat{\beta}_j\left(\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)\left((M+1)\hat{\beta}_k+\hat{\eta}_k\right)+\sum_{\substack{j=1\\j\neq k,\hat{m}}}^{K}\lambda_{\hat{m},j}^2M\hat{\beta}_{\hat{m}}\hat{\beta}_j\left(\hat{\beta}_k+\hat{\eta}_k\right)\left((M+1)\hat{\beta}_{\hat{m}}+M\hat{\eta}_{\hat{m}}\right)$$
$$+\sum_{\substack{i=1\\i\neq k,\hat{m}}}^{K}\lambda_{i,k}^2M\hat{\beta}_k\hat{\beta}_i\left(\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right)\left((M+1)\hat{\beta}_k+M\hat{\eta}_k\right)+\sum_{\substack{i=1\\i\neq k,\hat{m}}}^{K}\lambda_{i,\hat{m}}^2M^2\hat{\beta}_{\hat{m}}\hat{\beta}_i\left(\hat{\beta}_k+\hat{\eta}_k\right)\left((M+1)\hat{\beta}_{\hat{m}}+\hat{\eta}_{\hat{m}}\right). \quad (59)$$

$$Q_k = \lambda_{k,k}^2M(M+1)\hat{\beta}_k^2\left((M+2)\hat{\beta}_k+\hat{\eta}_k\right)+\sum_{i=1,i\neq k}^{K}\sum_{j=1,j\neq k}^{K}\lambda_{i,j}M\hat{\beta}_i\beta_k\hat{\beta}_j\left(\lambda_{j,i}+\lambda_{i,j}M\right)$$
$$+\sum_{i=1,i\neq k}^{K}M\hat{\beta}_k\hat{\beta}_i\left(2\lambda_{k,i}\lambda_{i,k}\left((M+1)\hat{\beta}_k+\hat{\eta}_k\right)+\lambda_{k,i}^2M\left((M+1)\hat{\beta}_k+\hat{\eta}_k\right)+\lambda_{i,k}^2\left((M+1)\hat{\beta}_k+M\hat{\eta}_k\right)\right). \quad (62)$$

user. We derived the asymptotic sum rate in closed-form. Our proposed scheme provides a sum rate gain of $(1-4/K)\times 100\%$ over the current state-of-the-art MWRN. The gain is more significant when the number of users (K) is increased. Also, the use of two time slots enable the use of MWRNs in fast fading channels with small coherence times. Also, the proposed scheme provides significant energy efficiency gains due to the improved sum rate and the reduced number of time slots. Furthermore, we proposed a power allocation scheme, which improves user fairness. We also verified the benefits of the proposed scheme from an energy-efficiency perspective.

APPENDIX A
EXPECTED VALUE RESULTS FOR IMPERFECT CSI

A. Derivation of $M_{k,m}$

First, by using (67b), we can rewrite the term inside the expected value as

$$\mathbf{h}_k^T\hat{\mathbf{H}}^*\mathbf{\Lambda}\hat{\mathbf{H}}^H\mathbf{h}_{f_k(m)} = \sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_{i,j}\mathbf{h}_k^T\hat{\mathbf{h}}_i^*\hat{\mathbf{h}}_j^H\mathbf{h}_{f_k(m)}. \quad (54)$$

By substituting the value for $\mathbf{h}_k$, we rewrite (54) as

$$M_{k,m} = \sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_{i,j}\left(\hat{\mathbf{h}}_k^T+\mathbf{e}_k^T\right)\hat{\mathbf{h}}_i^*\hat{\mathbf{h}}_j^H\left(\hat{\mathbf{h}}_{f_k(m)}+\mathbf{e}_{f_k(m)}\right). \quad (55)$$

As $\mathbf{e}_k$'s are independent from $\hat{\mathbf{h}}_k$'s, only the first term of the above summation can have a non-zero expectation. Next, by considering different i and j combinations (i.e., $i=k$, $j=f_k(m)$ and $j=k$, $i=f_k(m)$ and using the fact that $k\neq f_k(m)$) in the double summation in (54), the expected value can be written as

$$\mathbb{E}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_{i,j}\mathbf{h}_k^T\mathbf{h}_i^*\mathbf{h}_j^H\mathbf{h}_{f_k(m)}\right] = \lambda_{k,f_k(m)}\|\mathbf{h}_k\|^2\|\mathbf{h}_{f_k(m)}\|^2 + \lambda_{f_k(m),k}|\mathbf{h}_k\mathbf{h}_{f_k(m)}|^2. \quad (56)$$

Finally, by using the expected value results given in Appendix C, the value for $M_{k,m}$ is obtained as

$$M_{k,m} = M\left(\lambda_{k,f_k(m)}M+\lambda_{f_k(m),k}\right)\hat{\beta}_k\hat{\beta}_m, \quad (57)$$

where $\hat{\beta}_k$ is given as

$$\hat{\beta}_k = \frac{P\alpha_{p,k}\beta_k^2}{(P\alpha_{p,k}\beta_k+1)}. \quad (58)$$

Here, $\hat{\beta}_k$ acts as a correction for β_k values due to the imperfect CSI. When perfect CSI is available, $\hat{\beta}_k$ will become β_k.

B. Derivation of $N_{k,m}$

In order to compute $N_{k,m}$, we first derive the value of $\bar{N}_{k,m} = \mathbb{E}\left[|\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{h}_{f_k(m)}|^2\right]$ by using the same procedure as in $M_{k,m}$ and obtained as (59) at the top of this page. Here in (59), $\hat{m}=f_k(m)$ and $\hat{\eta}_k$ is defined as

$$\hat{\eta}_k = \frac{\beta_k}{(P\alpha_{p,k}\beta_k+1)}. \quad (60)$$

Here, $\hat{\eta}_k$ acts as an error term due to imperfect CSI. When perfect CSI is available, this term vanishes. The value of $N_{k,m}$ is then derived as

$$N_{k,m} = \bar{N}_{k,m} - M_{k,m}^2. \quad (61)$$

C. Derivation of Q_k

Similar to the previous cases, the value of Q_k is written as (62) at the top of this page.

APPENDIX B
DERIVATION OF Ψ FOR IMPERFECT CSI

In this section, we compute the average value of Ψ, which will be used for the SINR calculations.

$$L_1 = \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{m=1}^{K} \lambda_{i,j}\lambda_{k,l}\alpha_{m,m}\left(\mathbb{E}\left[\hat{\mathbf{h}}_i^T\hat{\mathbf{h}}_k^*\hat{\mathbf{h}}_l^H\hat{\mathbf{h}}_m\hat{\mathbf{h}}_m^H\hat{\mathbf{h}}_j\right] + \mathbb{E}\left[\hat{\mathbf{h}}_i^T\hat{\mathbf{h}}_k^*\hat{\mathbf{h}}_l^H\mathbf{e}_m\mathbf{e}_m^H\hat{\mathbf{h}}_j\right]\right). \tag{63}$$

$$\begin{aligned} L_1 = &\sum_{i=1}^{K}\lambda_{i,i}^2 M(M+1)\hat{\beta}_i^2\left((M+2)\hat{\beta}_i+\hat{\eta}_i\right) + \sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\sum_{m=1,m\neq i,j}^{K}\lambda_{i,j}^2 M^2\hat{\beta}_i\hat{\beta}_j\beta_m \\ &+\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K} M\hat{\beta}_i\left(\lambda_{i,i}^2(M+1)\hat{\beta}_i\beta_j + \lambda_{i,j}^2(M+1)\hat{\beta}_j\left(\hat{\beta}_i + M\hat{\beta}_j\right)\right. \\ &\left.+\lambda_{i,j}\lambda_{j,i}(M+1)\hat{\beta}_j\left(\hat{\beta}_i+\hat{\beta}_j\right) + \lambda_{i,j}^2 M\hat{\beta}_j\left(\hat{\eta}_i+\hat{\eta}_j\right) + \lambda_{i,j}\lambda_{j,i}\hat{\beta}_j\left(\hat{\eta}_i+\hat{\eta}_j\right)\right). \end{aligned} \tag{64}$$

A. Computation of L_1

In this section, we derive L_1. By using matrix multiplication identities, L_1 can be simplified as (63) at the top of this page. Then, L_1 can derived in closed-form by looking at all the possibilities for used variables as (64) at the top of this page.

B. Computation of L_2

By using matrix multiplication theories, L_2 can be simplified as follows.

$$L_2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_{i,j}\mathbb{E}\left[\hat{\mathbf{h}}_i^T\hat{\mathbf{H}}^*\mathbf{\Lambda}^H\hat{\mathbf{H}}^H\hat{\mathbf{h}}_j\right]. \tag{65}$$

By using the results in (57), we can write (65) as

$$L_2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\left(\lambda_{i,j}M+\lambda_{j,i}\right)M\lambda_{i,j}\hat{\beta}_i\hat{\beta}_j. \tag{66}$$

Appendix C
Important Results

The following results are used to derive (19a):

$$\mathbf{H}^*\mathbf{H}^H = \mathbf{H}\mathbf{H}^T = \sum_{n=1}^{K}\mathbf{h}_n^*\mathbf{h}_n^H, \tag{67a}$$

$$\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H = \mathbf{H}\mathbf{\Lambda}\mathbf{H}^T = \sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_{i,j}\mathbf{h}_i^*\mathbf{h}_j^H. \tag{67b}$$

Next, we list some of the important expected value results that were used in Appendices A and B.

$$\mathbb{E}\left[\hat{\mathbf{h}}_k^H\hat{\mathbf{h}}_k\right] = \mathbb{E}\left[\|\hat{\mathbf{h}}_k\|^2\right] = M\frac{P\alpha_{p,k}\beta_k^2}{(P\alpha_{p,k}\beta_k+1)} = M\hat{\beta}_k. \tag{68a}$$

$$\mathbb{E}\left[\hat{\mathbf{h}}_k^H\hat{\mathbf{h}}_j\right] = 0 \qquad k\neq j. \tag{68b}$$

$$\mathbb{E}\left[|\hat{\mathbf{h}}_k^H\hat{\mathbf{h}}_j|^2\right] = M\hat{\beta}_k\hat{\beta}_j \qquad k\neq j. \tag{68c}$$

$$\mathbb{E}\left[|\hat{\mathbf{h}}_k^H\hat{\mathbf{h}}_k|^2\right] = (M+1)\hat{\beta}_k^2. \tag{68d}$$

$$\mathbb{E}\left[\mathbf{e}_k^H\mathbf{e}_k\right] = \mathbb{E}\left[\|\mathbf{e}_k\|^2\right] = M\frac{\beta_k}{(P\alpha_{p,k}\beta_k+1)} = M\hat{\eta}_k. \tag{68e}$$

$$\mathbb{E}\left[|\mathbf{e}_k^H\mathbf{e}_j|^2\right] = M\hat{\eta}_k\hat{\eta}_j \qquad k\neq j. \tag{68f}$$

$$\mathbb{E}\left[|\mathbf{e}_k^H\mathbf{e}_k|^2\right] = M(M+1)\hat{\eta}_k^2. \tag{68g}$$

Appendix D
Asymptotic SINR with power scaling

We use the following asymptotic results for the computation of [46]. For two independent Gaussian vectors, $\mathbf{p}\sim\mathcal{CN}_{N\times 1}\left(0,\sigma_p^2\mathbf{I}\right)$ and $\mathbf{q}\sim\mathcal{CN}_{N\times 1}\left(0,\sigma_q^2\mathbf{I}\right)$, the following limits are easy to prove [44], [46]

$$\mathbf{p}^H\mathbf{p}/N \xrightarrow[N\to\infty]{a.s.} \sigma_P^2 \qquad \text{and} \qquad \mathbf{p}^H\mathbf{q}/N \xrightarrow[N\to\infty]{a.s.} 0, \tag{69}$$

$$\mathbf{p}^H\mathbf{q}/\sqrt{N} \xrightarrow[N\to\infty]{d} \mathcal{CN}\left(0,\sigma_p^2\sigma_q^2\mathbf{I}\right), \tag{70}$$

where subscripts *a.s.* and *d* stands for almost sure convergence and the convergence of distributions, respectively. Based on (69) and (70), the following asymptotic limit results can be established [44], [46]:

$$\frac{\mathbf{H}^H\mathbf{H}}{M} = \mathbf{D}^{\frac{1}{2}}\left(\frac{\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}}{M}\right)\mathbf{D}^{\frac{1}{2}} \xrightarrow[M\to\infty]{a.s.} \mathbf{D}, \tag{71a}$$

$$\frac{\mathbf{V}^H\mathbf{H}}{M} \xrightarrow[M\to\infty]{a.s.} \mathbf{0}_K, \tag{71b}$$

$$\frac{\mathbf{H}^H\mathbf{V}}{M} \xrightarrow[M\to\infty]{a.s.} \mathbf{0}_K, \tag{71c}$$

$$\frac{\mathbf{V}^H\mathbf{H}}{M} \xrightarrow[M\to\infty]{a.s.} \mathbf{I}_K, \tag{71d}$$

By using (71a), (71b), (71c) and (71d), the asymptotic limit for $\hat{\mathbf{H}}$ can be derived as follows:

$$\begin{aligned} \frac{\hat{\mathbf{H}}^H\hat{\mathbf{H}}}{M} &= \frac{1}{M}\left[\left(\mathbf{H}+\frac{\mathbf{V}}{\sqrt{P_p}}\right)\tilde{\mathbf{D}}\right]^H\left(\mathbf{H}+\frac{\mathbf{V}}{\sqrt{P_p}}\right)\tilde{\mathbf{D}} \\ &\xrightarrow[M\to\infty]{a.s.} \tilde{\mathbf{D}}^H\left(\mathbf{D}+\mathbf{I}_K\right)\tilde{\mathbf{D}} = \mathrm{diag}\left(\frac{P_P\beta_k^2}{1+P_P\beta_k}\right). \end{aligned} \tag{72}$$

Similarly, when the pilot power P_P is scaled as $P_P = E_P/\sqrt{M}$ the limit results can be written as

$$\frac{\hat{\mathbf{H}}^H\hat{\mathbf{H}}}{\sqrt{M}} \xrightarrow[M\to\infty]{a.s.} \mathrm{diag}\left(E_P\beta_k^2\right). \tag{73}$$

References

[1] S. Silva, G. Amarasuriya, C. Tellambura, and M. Ardakani, "NOMA-aided multi-way massive MIMO relay networks," in *Proc. IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–6.

[2] L. Ong, S. J. Johnson, and C. M. Kellett, "The capacity region of multiway relay channels over finite fields with full data exchange," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 3016–3031, May 2011.

[3] D. Gunduz, A. Yener, A. Goldsmith, and H. V. Poor, "The multiway relay channel," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 51–63, Jan 2013.

[4] G. Amarasuriya, C. Tellambura, and M. Ardakani, "Multi-way MIMO amplify-and-forward relay networks with zero-forcing transmission," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 4847–4863, Dec. 2013.

[5] M. N. Hasan and K. Anwar, "Joint decoding for multiway multirelay networks with coded random access," in *2016 22nd Asia-Pacific Conference on Communications (APCC)*, Aug. 2016, pp. 96–102.

[6] T. Ding, X. Yuan, and S. C. Liew, "Algorithmic beamforming design for MIMO multiway relay channel with clustered full data exchange," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10 081–10 086, Oct. 2018.

[7] S. Rahimian, W. Zhang, M. Noori, Y. Jing, and M. Ardakani, "Partial zero-forcing for multi-way relay networks," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4444–4456, Oct. 2018.

[8] F. L. Duarte and R. C. de Lamare, "Buffer-aided max-link relay selection for multi-way cooperative multi-antenna systems," *IEEE Commun. Lett.*, pp. 1–1, Jul. 2019.

[9] H. Wang and Q. Chen, "LDPC based network coded cooperation design for multi-way relay networks," *IEEE Access*, vol. 7, pp. 62 300–62 311, May 2019.

[10] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[11] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[12] D. P. Kudathanthirige and G. A. A. Baduge, "Multicell multiway massive MIMO relay networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6831–6848, Aug. 2017.

[13] C. D. Ho, H. Q. Ngo, M. Matthaiou, and T. Q. Duong, "On the performance of zero-forcing processing in multi-way massive MIMO relay networks," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 849–852, Apr. 2017.

[14] C. D. Ho, H. Q. Ngo, M. Matthaiou, and L. D. Nguyen, "Power allocation for multi-way massive MIMO relaying," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4457–4472, Oct. 2018.

[15] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, Jun. 2013, pp. 1–5.

[16] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[17] H. V. Cheng, E. Bjrnson, and E. G. Larsson, "Performance analysis of NOMA in training-based multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 372–385, Jan. 2018.

[18] D. Kudathanthirige and G. A. A. Baduge, "NOMA-aided multicell downlink massive MIMO," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 612–627, Jun. 2019.

[19] Z. Ding and H. V. Poor, "Design of Massive-MIMO-NOMA With Limited Feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[20] Z. Ding, L. Dai, R. Schober, and H. V. Poor, "NOMA Meets Finite Resolution Analog Beamforming in Massive MIMO and Millimeter-Wave Networks," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1879–1882, Aug. 2017.

[21] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully Non-Orthogonal Communication for Massive Access," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1717–1731, Apr. 2018.

[22] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[23] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[24] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, Secondquarter 2017.

[25] J. Choi, "On the Power Allocation for a Practical Multiuser Superposition Scheme in NOMA Systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 438–441, Mar. 2016.

[26] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource Management in Non-Orthogonal Multiple Access Networks for 5G and Beyond," *IEEE Network*, vol. 31, no. 4, pp. 8–14, Jul. 2017.

[27] S. Timotheou and I. Krikidis, "Fairness for Non-Orthogonal Multiple Access in 5G Systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[28] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of User Clustering in MIMO Non-Orthogonal Multiple Access Systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Jul. 2016.

[29] N. I. Miridakis and D. D. Vergados, "A survey on the successive interference cancellation performance for single-antenna and multiple-antenna ofdm systems," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 312–335, First 2013.

[30] 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, Mar. 2019.

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

[32] Y. Li and G. Amarasuriya, "NOMA-aided massive MIMO downlink with distributed antenna arrays," in *Proc. IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–7.

[33] F. Fang, Z. Ding, W. Liang, and H. Zhang, "Optimal energy efficient power allocation with user fairness for uplink MC-NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1133–1136, Aug. 2019.

[34] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

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

[36] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[37] Z. Mobini, P. Sadeghi, M. Khabbazian, and S. Zokaei, "Power allocation and group assignment for reducing network coding noise in multi-unicast wireless systems," *IEEE Trans. Veh. Technol.*, vol. 61, no. 8, pp. 3615–3629, Oct. 2012.

[38] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[39] B. Rankov and A. Wittneben, "Spectral efficient protocols for half-duplex fading relay channels," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 2, pp. 379–389, Feb. 2007.

[40] G. Amarasuriya, C. Tellambura, and M. Ardakani, "Performance analysis of zero-forcing for two-way MIMO AF relay networks," *IEEE Wireless Commun. Lett.*, vol. 1, no. 2, pp. 53–56, Apr. 2012.

[41] S. M. Kay, *Fundamentals of statistical signal processing: Estimation theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993, vol. 1.

[42] A. Mezghani and A. L. Swindlehurst, "Blind estimation of sparse broadband massive MIMO channels with ideal and one-bit ADCs," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2972–2983, Jun. 2018.

[43] K. Mawatwal, D. Sen, and R. Roy, "A semi-blind channel estimation algorithm for massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 1, pp. 70–73, Feb. 2017.

[44] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[45] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

[46] H. Cramér, *Random variables and probability distributions*. Cambridge University Press, 1970.

Photo.jpg

Shashindra Silva (S'10) received the B.Sc. degree in engineering with first-class honours from the Department of Electronic and Telecommunication Engineering in the University of Moratuwa, Moratuwa, Sri Lanka, in 2013 and the M.Sc. degree in electrical engineering from the Department of Electrical and Computer Engineering in the University of Alberta, AB, Canada, in 2015. He is currently working towards the Ph.D. degree at the Electrical and Computer Engineering Department, University of Alberta, AB, Canada. His current research interests include massive MIMO, machine learning for wireless systems, and cooperative MIMO relay networks.

Gayan Amarasuriya Aruma Baduge (S'09, M'13, SM'20) received the B.Sc. degree in engineering (first class Hons.) from the Department of Electronics and Telecommunications Engineering, University of Moratuwa, Moratuwa, Sri Lanka, in 2006, and the Ph.D. degree in electrical engineering from the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, in 2013. He was a Postdoctoral Research Fellow with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA from 2014 to 2016. Currently, he is an assistant professor in the Department of Electrical and Computer Engineering in Southern Illinois University, IL, USA. He is an Associate Editor for IEEE Communications Letters, IEEE Wireless Communications and IEEE Open Journal of the Communications Society.

Masoud Ardakani (M'04-SM'09) received the B.Sc. degree from Isfahan University of Technology in 1994, the M.Sc. degree from Tehran University in 1997, and the Ph.D. degree from the University of Toronto, Canada, in 2004, all in electrical engineering. He was a Postdoctoral fellow at the University of Toronto from 2004 to 2005. He is currently a Professor of Electrical and Computer Engineering at the University of Alberta, Canada. His research interests are in the general area of information theory. Dr. Ardakani serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and has served as an Associate Editor for the IEEE WIRELESS COMMUNICATIONS and as a senior editor for the IEEE COMMUNICATION LETTERS.

Chintha Tellambura (F11) received the B.Sc. degree (with first-class honor) from the University of Moratuwa, Sri Lanka, the MSc degree in Electronics from Kings College, University of London, United Kingdom, and the PhD degree in Electrical Engineering from the University of Victoria, Canada. He was with Monash University, Australia, from 1997 to 2002. Presently, he is a Professor with the Department of Electrical and Computer Engineering, University of Alberta. His current research interests include the design, modelling and analysis of cognitive radio, heterogeneous cellular networks and 5G wireless networks. Prof. Tellambura served as an editor for both IEEE Transactions on Communications (1999-2011) and IEEE Transactions on Wireless Communications (2001-2007) and for the latter he was the Area Editor for Wireless Communications Systems and Theory during 2007-2012. He has received best paper awards in the Communication Theory Symposium in 2012 IEEE International Conference on Communications (ICC) in Canada and 2017 ICC in France.

He is the winner of the prestigious McCalla Professorship and the Killam Annual Professorship from the University of Alberta. In 2011, he was elected as an IEEE Fellow for his contributions to physical layer wireless communication theory. In 2017, he was elected as a Fellow of Canadian Academy of Engineering. He has authored or coauthored over 500 journal and conference papers with an H-index of 74 (Google Scholar).