Low Complexity Optimization of the Asymptotic Spectral Efficiency in Massive MIMO NOMA

Lucinda Hadley and Ioannis Chatzigeorgiou

***Abstract*—Massive multiple-input multiple-output (MIMO) technology facilitates huge increases in the capacity of wireless channels, while non-orthogonal multiple access (NOMA) addresses the problem of limited resources in traditional orthogonal multiple access (OMA) techniques, promising enhanced spectral efficiency. This work uses asymptotic capacity computation results to reduce the complexity of a power allocation algorithm for small-scale MIMO-NOMA, so that it may be applied for systems with massive MIMO arrays. The proposed method maximizes the sum-capacity of the considered system, subject to power and performance constraints, and demonstrates greater accuracy than alternative approaches despite remaining low-complexity for arbitrarily large antenna arrays.**

***Index Terms*—Non-orthogonal multiple access (NOMA), multiple-input multiple-output (MIMO), ergodic capacity, power allocation, asymptotic eigenvalue distribution.**

I. Introduction

The demand for fast data links has increased rapidly over the last two decades as a result of an increasing number of users and devices. Moreover, there is a need for adaptable and scalable technologies to meet the diverse requirements of the internet of things (IoT). Fifth generation (5G) and sixth generation (6G) networks must be able to support increased multi-terabyte per second data traffic, while maintaining a high quality of service in terms of security, reliability and delay [1].

A key facilitator of the increased spectral efficiency (SE) seen between third and fourth generation mobile networks was the use of multiple-input multiple-output (MIMO) technology. MIMO enables dramatic increases in SE by exploiting spatial diversity [2] and can be extended by using even more antennas in 'massive MIMO' (MM). In 2018, a line of products with MM capability was approved by the Federal Communications Commission. These included 64-antenna arrays, such as the Ericsson AIR 6468. Similar products, including the Huawei AAU and Nokia Airscale, have also been launched with Huawei quoted as saying at the 2019 Mobile World Congress that "95% of their current commercial shipments has either 32 or 64 antennas" [3]. It is speculated that antenna arrays with dimensions of order 10^3 or even 10^4 could be used in future designs in so called 'supermassive MIMO'. MM is therefore of critical importance in industry and large-scale arrays are a topic of great interest in current research [1].

Rate optimization of a wireless network requires knowledge of the theoretical SE of its channels. In 1999 Telatar's ground-breaking work introduced the use of asymptotic properties of random matrices, in particular the limiting distributions of their eigenvalues, in computing the asymptotic SE of MIMO channels [4]. In 2004, [5] and [6] demonstrated some ways of generalizing the result, but the work was premature with respect to small-scale MIMO, whose capacity is more easily computed using the celebrated 'log-det' result [7]. With the recent introduction of MM, however, the analysis of very large random matrices is required, and the use of asymptotic results has resurfaced. The last several years have seen methods, such as free probability theory, used to compute the asymptotic eigenvalue distributions (AEDs) of a wider class of MIMO channel matrices [8]–[10].

Another method for enhancing SE is to share spectrum more effectively. Non-orthogonal multiple access (NOMA) is an emerging technology that shows promise in this area. Traditional NOMA uses the power domain to discriminate between signals (although a code-domain implementation of NOMA has also been proposed) [11]. Unlike orthogonal multiple access (OMA) methods, such as time and frequency division multiple access (TDMA and FDMA), which split the respective resources (spectrum and time) into 'orthogonal' frequency bands and time slots, NOMA serves multiple users in a single resource block (band or slot), thus enabling massive connectivity. This, along with the mitigating effect of using successive interference cancellation (SIC) to remove unwanted signals and improve the signal-to-interference-plus-noise ratio (SINR), results in increased SE [12]. NOMA is considered fairer than alternative multiple access schemes as it prioritizes the experience of cell-edge users with weaker channel connections. Moreover, it reduces average latency compared to OMA since users do not have to wait for specific slots [13].

Due to early results demonstrating its potential, NOMA already features in the 3GPP-LTE-A standard and was proposed for inclusion in the 5G New Radio (NR) [14]. Ultimately, NOMA was not included in 5G NR as a work-item, but was earmarked for use beyond 5G because the capacity benefits were considered to be outweighed by the implementation complexity [15], [16]. Therefore, it is necessary to increase the capacity benefits in relation to the complexity in order to make NOMA a viable option, and the use of massive antenna arrays is an obvious strategy. For the multi-user case in which the base station is equipped with multi-antenna arrays, while the user devices have a single antenna, [17] compares some user-pairing algorithms and investigates a new method for maximizing throughput, while in [18] the authors demonstrate the superior capacity of MIMO-NOMA over MIMO-OMA for communication between a multi-antenna receiver and clusters

This work has been funded by Lancaster University's EPSRC Doctoral Training Partnership.

L. Hadley and I. Chatzigeorgiou are with InfoLab21, School of Computing and Communications, Lancaster University, LA1 4WA, UK (e-mail: {lucinda.hadley, i.chatzigeorgiou}@lancaster.ac.uk).

of multi-antenna destinations. This is extended to massive-MIMO NOMA (MM-NOMA) in [19], which shows that a non-regenerative relay system where the base station is equipped with up to 500 antennas, outperforms a traditional MIMO-NOMA arrangement.

In this work we consider a low-complexity power allocation algorithm for two-user power-domain NOMA in which MM arrays are employed at all nodes and signals can be separated using superposition coding (SC) at the transmitter and SIC at the receiver. We assume that the transmitter has access to statistical channel state information (CSIT) only and we aim to maximize the ergodic capacity subject to power and rate constraints. This non-convex optimization problem was addressed for the case of small-scale MIMO by implementing a suboptimal algorithm and comparing it to the optimal bisection method in [20]. We extend the work to consider arbitrarily large MM arrays and demonstrate that it is possible to reduce the complexity of the bisection method further, by combining it with Telatar's method of asymptotic capacity computation, without loss of optimality. As far as the authors are aware, this approach has not previously been considered for this scenario.

Notations: $(\cdot)^\dagger$ denotes the conjugate transpose, $\mathrm{Tr}(\cdot)$ represents the matrix trace, $\mathbb{I}_N$ denotes the $N \times N$ identity matrix and $\mathbb{E}(\cdot)$ is the expectation.

II. System Model

Consider the open-loop MIMO system given in Fig. 1, where a source S transmits data to two users simultaneously using N_S antennas and user i receives using N_i antennas, where $i \in \{1, 2\}$. The signal vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are transmitted to user 1 and user 2 and the diagonal power allocation matrices are $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{C}^{N_\mathrm{S} \times N_\mathrm{S}}$ at each user respectively, where $\mathrm{Tr}(\mathbf{Q}_i)$ is the total power allocated to user i. Both signals occupy the same frequency and time slot but their transmit power varies, as is the usual convention for NOMA transmission. User 1 and user 2 are taken to be the 'weak user' and 'strong user', respectively. This could occur, for example, when S is a base station, user 1 is at the cell-edge and user 2 is near the center of the cell. It was determined in [20, Lemma 2] that uniform power allocation across each user's antennas results in optimal performance. Therefore, hereafter we will consider the case where the diagonal entries of $\mathbf{Q}_i$ are all equal and replace each $\mathbf{Q}_i$ with the constant scalar $p_i = \frac{\mathrm{Tr}(\mathbf{Q}_i)}{N_\mathrm{S}}$, which represents the power allocated to the desired signal of user i per antenna at the source.

User 1 and user 2 receive signals $\mathbf{y}_1$ and $\mathbf{y}_2$ respectively, which can be expressed as:

$$\mathbf{y}_1 = \sqrt{p_1}\,\mathbf{H}_1\mathbf{x}_1 + \sqrt{p_2}\,\mathbf{H}_1\mathbf{x}_2 + \mathbf{n}_1, \tag{1}$$

$$\mathbf{y}_2 = \sqrt{p_1}\,\mathbf{H}_2\mathbf{x}_1 + \sqrt{p_2}\,\mathbf{H}_2\mathbf{x}_2 + \mathbf{n}_2, \tag{2}$$

where $\mathbf{x}_i$ is the $N_\mathrm{S} \times 1$ vector of the transmitted signal carrying the message for user i, and $\mathbf{y}_i$ is the $N_i \times 1$ vector of the signal received by user i. Matrices $\mathbf{H}_i \in \mathbb{C}^{N_i \times N_\mathrm{S}}$ have random complex entries distributed as $\mathcal{CN}(0, \sigma^2_{\mathbf{H}_i})$, which model flat Rayleigh fading. Each entry of $\mathbf{H}_i$, denoted by h^i_{jk}, represents the channel gain between the kth transmit antenna of S and the jth receive antenna of user i. We assume that $\sigma^2_{\mathbf{H}_1} < \sigma^2_{\mathbf{H}_2}$ because user 1 is the weak user. Finally, the $N_i \times 1$ vector $\mathbf{n}_i$ models the normalized additive white Gaussian noise across the corresponding channel.

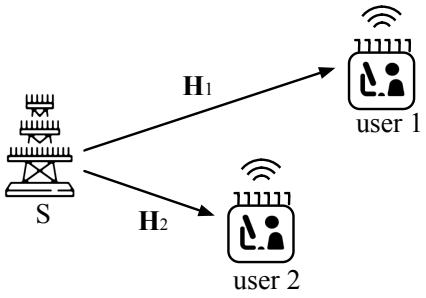

Figure 1: Broadcast MM-NOMA system model using SIC.

Since we are using NOMA, the source simultaneously communicates with the users using the same resource block, and their signals are multiplexed by allocating a different transmission power, p_i, for each user's signal, at each antenna. Because the weaker user is allocated more power, it is able to decode the message by treating the interference from the other user's signal as noise. Define C_1 and C_2 as the SEs of user 1 and user 2 respectively. We will set a minimum rate constraint of $C_1 > R_0$ for the weak user and assume that the SINR of the weak user's signal is always smaller at the weak user than it is at the strong user so that

$$C_1 \leq \log_2 \left| \mathbb{I}_{N_2} + \left(\mathbb{I}_{N_2} + p_2 \mathbf{H}_2 \mathbf{H}_2^\dagger \right)^{-1} p_1 \mathbf{H}_2 \mathbf{H}_2^\dagger \right|, \tag{3}$$

which guarantees successful SIC detection at the strong user. This means that the strong user can decode the weak user's message and subtract it from the overall signal in order to decode its own message [21].

The weak user decodes its own signal, $\mathbf{x}_1$, while interpreting the interference caused by $\mathbf{x}_2$ as noise. The achievable ergodic SEs are therefore given by:

$$\begin{aligned} C_1 =& \mathbb{E}_{\mathbf{H}_1} \left(\log_2 \left| \mathbb{I}_{N_1} + \left(\mathbb{I}_{N_1} + p_2 \mathbf{H}_1 \mathbf{H}_1^\dagger \right)^{-1} p_1 \mathbf{H}_1 \mathbf{H}_1^\dagger \right| \right) \\ =& \mathbb{E}_{\mathbf{H}_1} \left(\log_2 \left| \mathbb{I}_{N_1} + (p_1 + p_2)\, \mathbf{H}_1 \mathbf{H}_1^\dagger \right| \right) \\ & - \mathbb{E}_{\mathbf{H}_1} \left(\log_2 \left| \mathbb{I}_{N_1} + p_2 \mathbf{H}_1 \mathbf{H}_1^\dagger \right| \right), \end{aligned} \tag{4}$$

$$C_2 = \mathbb{E}_{\mathbf{H}_2} \left(\log_2 \left| \mathbb{I}_{N_2} + p_2 \mathbf{H}_2 \mathbf{H}_2^\dagger \right| \right). \tag{5}$$

III. Optimization Problem

The optimization problem of maximizing the combined SE of the two users, subject to power and minimum rate constraints, can be formulated as:

$$\begin{aligned} \max_{p_1, p_2 \geq 0} \quad & C_1(p_1, p_2) + C_2(p_2), \\ \text{s.t.} \quad & C_1(p_1, p_2) \geq R_0 \\ & (p_1 + p_2) N_\mathrm{S} \leq p_{\max}, \end{aligned} \tag{6}$$

where $p_{\max}$ denotes the total available power at the source, R_0 is the minimum SE required for reasonable performance at the

Table I: Optimal bisection algorithm‡

Initialize $p_{2,\min} = 0$, $p_{2,\max} = p_{\max}$
while $p_{2,\max} - p_{2,\min} > \epsilon$ **do**
Set $p_2^\star = (p_{2,\min} + p_{2,\max})/2$, $p_1^\star = p_{\max} - p_2^\star$.
Calculate $C_1(p_1^\star, p_2^\star)$.
If $C_1(p_1^\star, p_2^\star) < R_0$, set $p_{2,\max} = p_2^\star$; Else, set $p_{2,\min} = p_2^\star$.
end while
Output: $p_1 = p_1^\star$, $p_2 = p_2^\star$.

‡ $p_{\max}$ in the algorithm is set equal to $p_{\max}/N_{\mathrm{S}}$ as per (6).

weak user and $C_1(p_1, p_2)$ and $C_2(p_2)$ refer to the SEs defined in (4) and (5) respectively, written in terms of the optimization variables p_1 and p_2.

In [20] the authors develop an optimal and suboptimal method of solving the problem. Since the function $C_1 + C_2$ increases with p_2, the optimal solution is on the boundary of the feasible region. In particular, it occurs when p_1 is as small as possible while ensuring that $C_1 > R_0$. This p_1 can be found using repeated bisection as shown in Table I, where ϵ is reduced for greater precision. The suboptimal method relies on an approximation of C_1 and is successful for MIMO systems with $N_{\mathrm{S}}, N_i \leq 4$. However, the optimality of the results using this method deteriorates as the numbers of antennas at each end of the communication link increase.

In this paper, we demonstrate how to reduce the complexity of the optimal bisection method by computing C_1 using the asymptotic eigenvalue distribution of the channel matrices, thus improving the accuracy of the optimization for MM-NOMA systems.

IV. Theory

Let $\mathbf{G}_\beta \in \mathbb{C}^{N_r \times N_t}$ be a random matrix, where the limit of the ratio $\frac{N_t}{N_r}$ is β as both N_t and N_r tend to infinity, and $\mathbf{X}_\beta = \mathbf{G}_\beta \mathbf{G}_\beta^\dagger \in \mathbb{C}^{N_r \times N_r}$. When the entries of $\mathbf{G}_\beta$ conform to certain distribution rules and α is a scalar, a 'log-det' expression, $\frac{1}{N_r}\log_2|\mathbb{I}_{N_r} + \alpha \mathbf{X}_\beta|$ can be expressed in terms of the AED, $f_{\mathbf{X}_\beta}(x)$, of $\mathbf{X}_\beta$. Using this result, the SE of a channel modeled as $\mathbf{G}_\beta$ can then be written in terms of the AED of $\mathbf{X}_\beta$ as [5]:

$$\begin{aligned}\mathcal{C}_{\alpha\mathbf{X}_\beta}^{Asy} &= N_r \left(\lim_{\substack{N_t, N_r \to \infty \\ \frac{N_t}{N_r} \to \beta}} \frac{1}{N_r} \log_2 |\mathbb{I}_{N_r} + \alpha \mathbf{X}_\beta| \right) \\ &= N_r \left(\lim_{\substack{N_t, N_r \to \infty \\ \frac{N_r}{N_t} \to \beta}} \frac{1}{N_r} \sum_{i=1}^{N_i} \log_2 \left(1 + \alpha \lambda_{\mathbf{X}_\beta}(i)\right) \right) \\ &= N_r \int_0^\infty \log_2 (1 + \alpha x)\, f_{\mathbf{X}_\beta}(x)\, dx, \end{aligned} \tag{7}$$

where $\lambda_{\mathbf{X}_\beta}(i)$ is the ith eigenvalue of $\mathbf{X}_\beta$.

There are many existing works in which the main result has been to compute the AEDs of non-standard channel matrices, usually with the aim of applying (7) to compute their capacity. For example, Pan *et al.* [8] use free probability theory to compute the AED of massive MIMO channel matrices with transmit and receive correlation. Hadley *et al.* [10] derive the AED of the combined channels in the second hop of a multi-relay system, while Diaz and Pérez-Abreu [9] find the AED for more generalized block matrices. Shlyakhtenko [22] shows how to extend existing results to find the AED of band Gaussian matrices used to model independent but non-identically distributed Gaussian channels.

In this paper, the channels are modeled as having entries distributed as $\mathcal{CN}(0, \sigma^2_{\mathbf{H}_i})$, which is the canonical model for single-user narrowband MIMO channels [5], and so we make use of the following result.

Definition 1: The AED of $\mathbf{X}_\beta = \mathbf{G}_\beta \mathbf{G}_\beta^\dagger$ as $N_t, N_r \to \infty$ and $\frac{N_t}{N_r} \to \beta$, where $\mathbf{G}_\beta \in \mathbb{C}^{N_r \times N_t}$ is a standard Gaussian random matrix with entries distributed as $\mathcal{CN}(0,1)$, is given by the Marçenko-Pasteur distribution [5]:

$$f_{\mathbf{X}_\beta}(x) = \frac{\sqrt{(x-a)^+ (b-x)^+}}{2\pi\beta x} + \left(1 - \frac{1}{\beta}\right)^+ \delta(x), \tag{8}$$

where $a = \left(1 - \sqrt{\beta}\right)^2$, $b = \left(1 + \sqrt{\beta}\right)^2$, $(z)^+ = \max(0, z)$ and $\delta(x)$ is the Dirac-delta function.

Our aim is to find C_1 and C_2, as given in (4) and (5). Now our channel matrices $\mathbf{H}_i$ can be written as $\sigma_i \mathbf{G}_{\beta_i}$, where we have substituted $\beta = \beta_i$ into Definition 1 so that $\mathbf{H}_i \mathbf{H}_i^\dagger = \sigma_i^2 \mathbf{X}_{\beta_i}$. Therefore, to find C_1 and C_2 in closed form, we can apply (7) to obtain:

$$\begin{aligned} C_1 &= \log_2 |\mathbb{I}_{N_1} + c_1 \mathbf{X}_{\beta_1}| - \log_2 |\mathbb{I}_{N_2} + c_2 \mathbf{X}_{\beta_1}| \\ &= \mathcal{C}_{c_1 \mathbf{X}_{\beta_1}}^{Asy} - \mathcal{C}_{c_2 \mathbf{X}_{\beta_1}}^{Asy} \\ &= \int_0^\infty \log_2 \left(\frac{1 + c_1 x}{1 + c_2 x} \right)^{f_{\mathbf{X}_{\beta_1}}(x)} dx \\ &= \log_2 \left(\frac{e^{\frac{\mathcal{Q}_{2,1}}{c_2}} (1 + c_1 - \mathcal{Q}_{1,1})^{\beta_1} (1 + c_1\beta_1 - \mathcal{Q}_{1,1})}{e^{\frac{\mathcal{Q}_{1,1}}{c_1}} (1 + c_2 - \mathcal{Q}_{2,1})^{\beta_1} (1 + c_2\beta_1 - \mathcal{Q}_{2,1})} \right) \end{aligned} \tag{9}$$

$$\begin{aligned} C_2 &= \mathcal{C}_{c_3 \mathbf{X}_{\beta_2}}^{Asy} \\ &= \int_0^\infty \log_2 (1 + c_3 x)\, f_{\mathbf{X}_{\beta_2}}(x) \\ &= \log_2 \left(\frac{(1 + c_3 - \mathcal{Q}_{3,2})^{\beta_2} (1 + c_3\beta_2 - \mathcal{Q}_{3,2})}{e^{\frac{\mathcal{Q}_{3,2}}{c_3}}} \right), \end{aligned} \tag{10}$$

where $c_1 = (p_1 + p_2)\sigma_1^2$, $c_2 = p_2\sigma_1^2$, $c_3 = p_2\sigma_2^2$, $f_{\mathbf{X}}(x)$ is given by (8) and, for notational convenience, we have set:

$$\mathcal{Q}_{\rho,q} = \frac{1}{4} \left(\sqrt{c_\rho \left(1 + \sqrt{\beta_q}\right)^2 + 1} - \sqrt{c_\rho \left(1 - \sqrt{\beta_q}\right)^2 + 1} \right)^2.$$

V. Results and Discussion

In this section we compare: (i) the bisection algorithm described in [20], which relies on the traditional method of capacity computation given in (4) and (5) and finds the optimal power allocation, (ii) the suboptimal algorithm also derived in [20] which omits the need for repeated bisections but still relies on computing the expectation over multiple realizations of the determinant of a matrix, and (iii) the bisection method

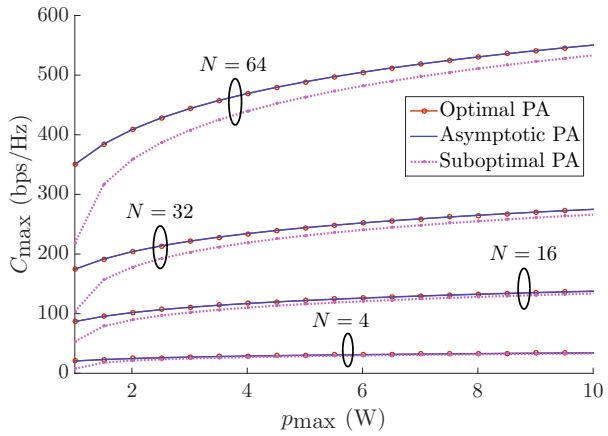

Figure 2: Sum-capacity vs total transmission power

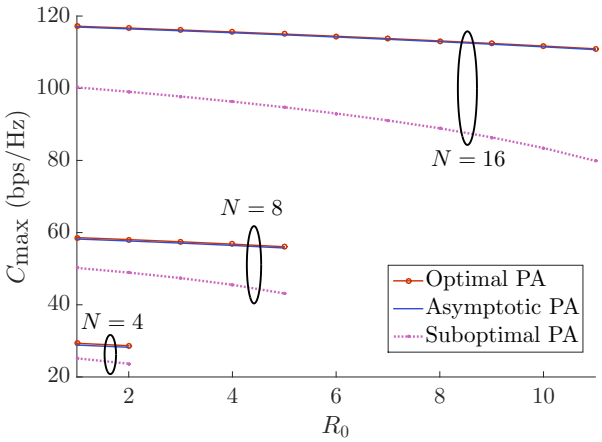

Figure 3: Sum-capacity vs minimum rate of weak user

using our asymptotic capacity equations (9) and (10) in placec of the traditional method. For the sake of simplicity, we have considered the cases where $N_{\mathrm{S}} = N_i = N$ in our results.

Fig. 2 plots the total available power $p_{\max}$ against the maximized sum of the ergodic capacities of the two users obtained using (6), which we shall denote by $C_{\max}$. We fixed $\sigma^2_{\mathbf{H}_1} = 20$ dB, $\sigma^2_{\mathbf{H}_2} = 5$ dB and $R_0 = 2$ bps/Hz. Both the asymptotic and suboptimal methods appear to achieve very close to optimal performance for smaller MIMO arrays of 4×4 antennas, however, as we increase the number of antennas the suboptimal method becomes less efficient. On the other hand, the asymptotic approach is able to match the optimal result perfectly regardless of the array size. The suboptimal result is also shown to be less accurate for systems with low power availability, while the asymptotic approach is unaffected.

Fig. 3 plots the minimum rate requirement of the weak user against $C_{\max}$ with $\sigma^2_{\mathbf{H}_1} = 20$ dB, $\sigma^2_{\mathbf{H}_2} = 1$ dB, $p_{\max} = 4$ W for various antenna array sizes. The range of values of R_0 is restricted by the assumption given in (3), however for larger MIMO arrays this restriction is reduced. We see that the asymptotic approach is optimal for any rate restraint whereas the suboptimal method deteriorates significantly when the rate requirement of the weak user increases and that the degree of the deterioration increases with N.

Fig. 4 plots the channel gain of the weak user against $C_{\max}$, for $\sigma^2_{\mathbf{H}_1} = 20$ dB, $p_{\max} = 4$ W, $R_0 = 2$ bps/Hz and various antenna array sizes. Again, the performance of the suboptimal method suffers for larger antenna arrays, most significantly in the case where the channel gain of the weak user is very small compared to that of the strong user, $\sigma^2_{\mathbf{H}_1} << \sigma^2_{\mathbf{H}_2}$, which would happen when the strong user was very near to the base station while the weak user was very remote. As before, the asymptotic approach remains accurate in all cases.

Next we consider the computational complexity, which depends on the number of antennas (for which we will consider the case where $N_{\mathrm{S}} \neq N_i$), the number of iterations used to compute the expectations involved in the optimal and suboptimal methods K, and the number of bisections M required for the optimal and asymptotic methods.

The optimal bisection method is the most complex. It

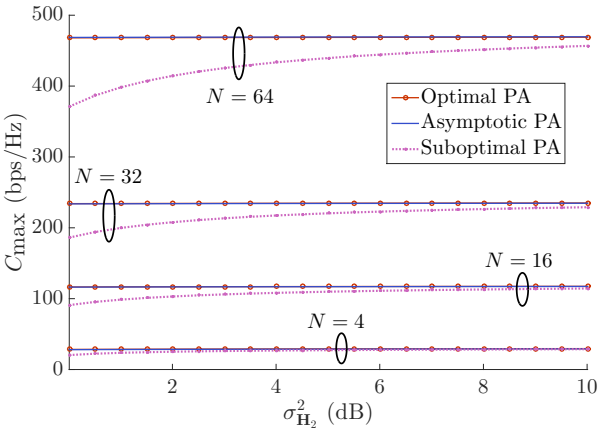

Figure 4: Sum-capacity vs channel gain of weak user

involves looping through the computation M times and computing C_1 K times in each loop to find the expectation. The complexity order of calculating C_1 is $\mathcal{O}(N_1!)$ since the most complex operation is taking the determinant of the $N_1 \times N_1$ matrix $\left[\mathbb{I}_{N_1} + (\mathbb{I}_{N_1} + (p_2\mathbf{H}_1\mathbf{H}_1^\dagger)^{-1})p_1\mathbf{H}_1\mathbf{H}_1^\dagger\right]$ in (4) (recall that $\mathbf{H}_i \in \mathbb{C}^{N_i\times N_{\mathrm{S}}}$). The overall complexity order of this method is $\mathcal{O}(KMN_1!)$, where we note that increasing N_{S} and N_2 does increase the complexity, but the complexity order is dominated by N_1.

In comparison the asymptotic approach also loops over the capacity computation M times but computes the capacity using the closed form in (9), for which the complexity is invariant with respect to N_{S}, N_i, K and M, thus the overall complexity order of this method is $\mathcal{O}(M)$.

Finally, the complexity of the suboptimal approach does not require looping through M bisections, however it still involves computing the expectation over K iterations of a computation involving the determinant of an $N_1 \times N_1$ matrix, thus it has complexity order $\mathcal{O}(KN_1!)$.

We note that the complexity order of the determinant computation can be reduced from $\mathcal{O}(N_1!)$ to as little as $\mathcal{O}(N_1^{2.81})$ using the methods in [23][Theorem 6.6]. However, the implementation of these methods is beyond the scope of

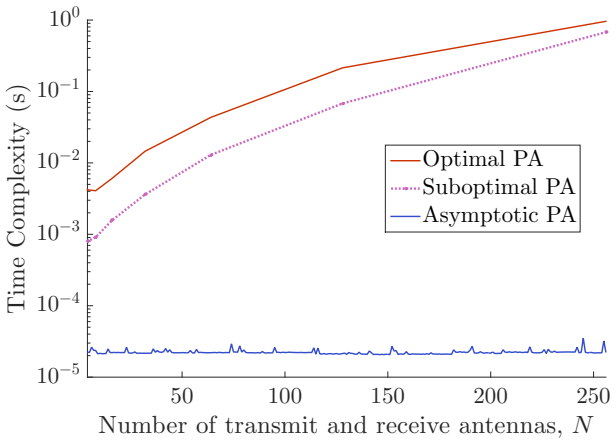

Figure 5: Time complexity of power allocation algorithms

this paper. We have used the Matlab function `det`, which relies on the LU decomposition method for calculating the determinant and has complexity order $\mathcal{O}(N_1^3)$, which gives complexity orders $\mathcal{O}(KMN_1^3)$, $\mathcal{O}(M)$ and $\mathcal{O}(KN_1^3)$ for the respective methods.

We compare the time complexity of the three approaches for increasingly large antenna arrays in Fig. 5. Note that we fixed $K=10$ for the expectation calculations. Experimentation demonstrated that accurate results for the considered range of N are observed if the number of bisections is at least $M = 13$ for $\epsilon = 0.001$ (ϵ is used in Table I). With K and M fixed, the complexity of the optimal and suboptimal methods depends only on the number of antennas, as is corroborated by Fig. 5. In agreement with our calculations, the complexity of the asymptotic approach remains constant regardless of the size of the antenna array.

VI. Conclusions

We have used asymptotic analysis to extend the results of [20] and demonstrated how best to allocate power resources to achieve optimal sum-capacity for an MM-NOMA system. We have demonstrated that the proposed asymptotic approach performs optimally for arbitrarily large antenna arrays while the accuracy of the suboptimal method of [20] decreases significantly with size for arrays larger than 4×4. Moreover, we have shown that the suboptimal method deteriorates in the cases of (i) low total power availability (ii) high minimum rate requirement at the weak user and (iii) significant difference between channel gains of users. The asymptotic method, on the other hand, agrees with the optimal method and is unaffected by these changes. Finally, we have demonstrated that the complexity of the asymptotic algorithm is lower than that of the optimal and suboptimal approaches regardless of array size. We conclude that the proposed power optimization method is superior for MM-NOMA.

References

[1] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Jul. 2019.

[2] J. R. Hampton, *Introduction to MIMO communications*. Cambridge University Press, 2013.

[3] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality - What is next?: Five promising research directions for antenna arrays," *Dig. Sig. Process.*, vol. 94, pp. 3–20, Nov. 2019.

[4] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[5] A. M. Tulino and S. Verdú, *Random matrix theory and wireless communications*. Now Publishers, 2004, vol. 1.

[6] R. R. Müller, "Random matrices, free probability and the replica method," in *Proc. 12th Eur. Sig. Process. Conf.*, Vienna, Austria, Sep. 2004.

[7] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO wireless communications*. Cambridge University Press, 2007.

[8] P. Pan, Y. Zhang, Y. Sun, and L. Yang, "On the asymptotic spectral efficiency of uplink MIMO-CDMA systems over Rayleigh fading channels with arbitrary spatial correlation," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 679–691, Feb. 2013.

[9] M. Diaz and V. Pérez-Abreu, "On the capacity of block multiantenna channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5286–5298, Aug. 2017.

[10] L. Hadley, Z. Ding, and Z. Qin, "Capacity analysis of asymmetric multi-antenna relay systems using free probability theory," in *Proc. IEEE 89th Veh. Tech. Conf. (VTC Spring)*, Kuala Lumpur, Malaysia, Apr. 2019.

[11] M. Vaezi, Z. Ding, and H. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Springer, 2019.

[12] Z. Ding, Z. Yang, P. Fan, and H. 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.

[13] S. Islam, M. Zeng, and O. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus*, vol. 1, no. 2, pp. 1–6, Jun. 2017.

[14] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[15] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. of the Commun. Soc.*, vol. 1, pp. 179–189, Jan. 2020.

[16] "Study on non-orthogonal multiple access (NOMA) for NR," Dec. 2018. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.812

[17] S. Islam, M. Zeng, O. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[18] M. Zeng, A. Yadav, O. Dobre, G.I.Tsiropoulos, and H. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Jul. 2017.

[19] 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, Aug. 2017.

[20] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Apr. 2015.

[21] Z. Ding, F. Adachi, and H. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Sep. 2015.

[22] D. Shlyakhtenko, "Random Gaussian band matrices and freeness with amalgamation," *Int. Math. Research Notices*, vol. 1996, no. 20, pp. 1013–1025, Jan. 1996.

[23] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The design and analysis of computer algorithms*. Addison-Wesley, 1974.