On the Power Allocation for MIMO-NOMA Systems With Layered Transmissions

Jinho Choi, *Senior Member, IEEE*

***Abstract*—In this paper, we study optimal power allocation for multiple-input multiple-output (MIMO) nonorthogonal multiple access (NOMA) systems when a layered transmission scheme is employed. An approach to maximize the sum rate of MIMO-NOMA with layered transmissions is proposed once we show that the sum rate is concave in allocated powers to multiple layers of users. We also derive a closed-form expression for the average sum rate when statistical channel state information (CSI) is available at a transmitter, which allows us to allocate powers to multiple layers for the maximization of the average sum rate. We find lower- and upper-bounds on the average sum rate, from which it is shown that the scaling property of MIMO-NOMA with layered transmissions also holds as conventional MIMO does (i.e., the average sum rate grows linearly with the number of antennas).**

***Index Terms*—Nonorthogonal multiple access, power allocation, successive interference cancellation.**

I. Introduction

In order to improve the spectral efficiency in a cellular system, nonorthogonal multiple access (NOMA) is recently studied in [1]–[4]. In NOMA, the power domain is to be exploited to improve the spectral efficiency when users have different channel gains. For example, if there are one user of high channel gain (for convenience, this user is referred to as user 1) and the other user of low channel gain (this user is referred to as user 2), a base station (BS) can transmit two signals to both the two users simultaneously in the same frequency band and time slot. The BS usually allocates a high transmission power to user 2 and a low transmission power to user 1. From this, at user 1, the signal to user 2 might be decodable. Thus, user 1 decodes the signal to user 2 first and then decodes his/her signal after subtracting the decoded signal to user 2. At user 2, the signal to user 2 can be decoded without significant interference from the signal to user 1 as it might be weak. Clearly, in NOMA, by making use of superposition coding (SC) [5] together with successive interference cancellation (SIC) [6], it is possible to achieve a higher spectral efficiency than that in orthogonal multiple access (OMA).

NOMA has been extended to various systems and with multiple input multiple output (MIMO) systems. In [7], NOMA is studied for downlink coordinated two-point systems. When the BS is equipped with multiple antennas, beamforming can be used for NOMA downlink transmissions as in [2], [8]. A performance analysis is presented in [9] and a power allocation problem for NOMA is studied in [10]. In [11], the ergodic capacity of MIMO-NOMA with two users is derived and compared with that of OMA (time division multiple access (TDMA) is used for OMA), which shows that NOMA can provide a higher sum rate than OMA and the performance gap grows with the channel gain difference between two users. In addition, the power allocation between two users is carried out for open-loop MIMO downlink transmissions (i.e., the BS allocates the power to two users' signals based on statistical CSI).

While SC and SIC are employed in NOMA to deal with inter-user interference, they are also used in single-user MIMO systems [12]–[15]. Layered transmissions in MIMO based on SC and SIC (often called horizontal Bell Labs layered space-time (H-BLAST) schemes) are studied in [12], [14]. In [16], the optimal power allocation for layered transmissions is investigated. A salient feature of layered transmissions is that the complexity to decode signals at a receiver is low as sequence-by-sequence decoding with SIC is used. In particular, MIMO detection, which can require a complexity growing exponentially with the number of transmit antennas for an optimal performance [17], [18], is not necessary since a signal sequence in each layer is to be independently detected/decoded. As a result, in layered transmissions, the decoding complexity grows linearly with the number of layers or transmit antennas.

We note that intra-user SC and SIC are employed for layered transmissions in MIMO, while inter-user SC and SIC are used in NOMA. Thus, in this paper, we consider layered transmissions for MIMO-NOMA with intra-user and inter-user SC and SIC in downlink transmissions. The main advantage of (downlink) MIMO-NOMA with layered transmissions is that the complexity of detection/decoding at users grows linearly (not exponentially) with the number of transmit antennas or layers, which would be crucial as the computing power and energy of users' devices are usually limited. In this paper, we mainly focus on the power allocation for MIMO-NOMA with layered transmissions when known instantaneous channel state information (CSI) or statistical CSI is available at a BS. The optimal power allocation is to maximize the sum rate by optimally allocate powers to multiple layers under a maximum transmission power constraint for each user. An approach for the sum rate maximization is derived once we show that the sum rate is concave in allocated powers to multiple layers of users. Based on a derived closed-form expression for the average sum rate, an optimal power allocation is also considered when statistical

Manuscript received September 16, 2015; revised November 23, 2015 and January 4, 2016; accepted January 5, 2016. Date of publication January 14, 2016; date of current version May 6, 2016. The associate editor coordinating the review of this paper and approving it for publication was E. A. Jorswieck.

The author is with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju 500 712, South Korea (e-mail: jchoi0114@gist.ac.kr).

Digital Object Identifier 10.1109/TWC.2016.2518182

CSI is available at the BS. Throughout the paper, we confine ourselves to the case of two users in the same frequency band and time slot in MIMO-NOMA for tractable analysis although MIMO-NOMA with layered transmissions is applicable to the case of more than two users.

The main contributions of the paper are as follows: *i)* an approach to perform the optimal power allocation for MIMO-NOMA with layered transmissions with maximum transmission power constraints is proposed; *ii)* a closed-form expression for the average sum rate is derived for the optimal power allocation with statistical CSI; *iii)* the scaling property of MIMO-NOMA with layered transmissions is analytically shown with upper- and lower-bounds on the average sum rate.

The rest of the paper is organized as follows. The system model for MIMO-NOMA with layered transmissions is presented in Section II. An approach for the optimal power allocation to maximize the sum rate with known instantaneous CSI is discussed in Section III. In Section IV, a closed-form expression for the average sum rate is derived with statistical CSI at the BS and the optimal power allocation to maximize the average sum rate is studied. Simulation results are presented in Section V. The paper is concluded with some remarks in Section VI.

Notation: Matrices and vectors are denoted by upper- and lower-case boldface letters, respectively. The superscripts T and H denote the transpose and complex conjugate, respectively. For a given vector, $\mathbf{x}$, $[\mathbf{x}]_m$ denotes the mth element. For a matrix $\mathbf{X}$, $[\mathbf{X}]_{m,n}$ represents the (m, n)th element. The determinant of a square matrix $\mathbf{A}$ is denoted by $\det(\mathbf{A})$. $\mathbb{E}[\cdot]$ and $\text{Var}(\cdot)$ denote the statistical expectation and variance, respectively. $\mathcal{CN}(\mathbf{a}, \mathbf{R})$ represents the distribution of circularly symmetric complex Gaussian (CSCG) random vector with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{R}$.

II. System Model

In this section, we present a system model for MIMO-NOMA [11] with a layered transmission scheme that is studied in [13], [14].

A. MIMO-NOMA System

Suppose that a BS is to support two users in the same frequency band and time slot using NOMA for downlink transmissions. The BS is equipped with N antennas and each user is equipped with M antennas. For convenience, we assume that user 1 is sufficiently close to the BS, while user 2 is not close to the BS (e.g., a cell-edge user). Denote by $\mathbf{H}_k$ the channel matrix from the BS to user k. The signal vector of size $M \times 1$ to user k is denoted by $\mathbf{x}_k$. It is assumed that $\mathbf{x}_k$ is a signal vector of a coded sequence. Thus, we may denote a coded sequence by $\{\mathbf{x}_k(l)\}$, where $\mathbf{x}_k(l)$ represents the lth signal vector. For convenience, we omit the time index l.

At user k, the received signal vector is given by

$$\mathbf{y}_k = \mathbf{H}_k(\mathbf{x}_1 + \mathbf{x}_2) + \mathbf{n}_k,\ k \in \{1, 2\}, \tag{1}$$

where $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I})$ is the background noise vector at user k. Throughout the paper, we assume that $\mathbb{E}[\mathbf{x}_k] = \mathbf{0}$. In general, we expect that the magnitudes of the elements of $\mathbf{H}_1$ are larger than those of $\mathbf{H}_2$ due to large-scale fading. Thus, the transmission power allocated to user 1 can be lower than that to user 2 for a similar received power level and the power difference is to be exploited in NOMA transmissions for a higher spectral efficiency.

At user 1, since $\mathbf{x}_2$ can have a higher power, it can be decoded. Once $\mathbf{x}_2$ is obtained, using SIC, the desired signal can be decoded from

$$\begin{aligned}\bar{\mathbf{y}}_1 &= \mathbf{y}_1 - \mathbf{H}_1\mathbf{x}_2 \\ &= \mathbf{H}_1\mathbf{x}_1 + \mathbf{n}_1.\end{aligned} \tag{2}$$

At user 2, since the power of $\mathbf{x}_1$ might be lower than that of $\mathbf{x}_2$, $\mathbf{x}_2$ can be decoded in the presence of $\mathbf{x}_1$ as interference.

In general, the BS would have more antennas than a user has, i.e., $N > M$. In the case that $N > M$, we may assume precoding at the BS with a precoding matrix of size $N \times M$, denoted by $\mathbf{W}$. Then, the effective channel matrix becomes $\mathbf{H}_k\mathbf{W}$, which is an $M \times M$ matrix. Thus, throughout the paper, we assume $N = M$ and $\mathbf{H}_k$ is a composite channel matrix. Note that since we only focus on the power allocation in this paper, we do not study precoding [4].

Although we focus on the case of two users in this paper, it might be possible to consider the case of multiple users as in [19] where a minorization-maximization algorithm (MMA) is employed to approximate a non-convex sum rate maximization problem together with precoding to maximize data rates. Alternatively, a group of users can be divided into multiple subgroups of users under a certain criterion (based on the spatial correlation of channels and power differences between users) that allows precoding to suppress or minimize the interference between subgroups and exploits the gain from NOMA within each subgroup of users. This issue will be studied as a further research topic in the future.

B. Layered Transmissions

Suppose that the BS is to transmit M coded signals simultaneously to each user as in H-BLAST [12]. With the layered transmission scheme, each user can carry out sequence-by-sequence decoding with SIC for spatial multiplexing [12]–[14]. Thus, the decoding complexity at users can be much lower than that with non-layered transmission schemes (in particular, no joint MIMO detection is required with the layered transmission scheme). However, unfortunately, this scheme can reduce the achievable rate. To minimize the decrease of the achievable rate by the layered transmission, we will consider the optimal power allocation for MIMO-NOMA with layered transmissions in Section III.

Denote by $x_{m;k}$ the mth element of $\mathbf{x}_k$. It is assumed that $\{x_{m;k}(l)\}$ is a coded sequence from the mth channel encoder. That is, in each layer, the symbol sequence is assumed to be independently encoded. Consider the QR factorization of $\mathbf{H}_k$ as

$$\mathbf{H}_k = \mathbf{Q}_k\mathbf{R}_k,$$

where $\mathbf{Q}_k$ is an $M \times M$ orthogonal matrix and $\mathbf{R}_k$ is an $M \times M$ upper triangular matrix. At user k, we can have

$$\begin{aligned}\mathbf{z}_k &= \mathbf{Q}_k^{\mathrm{H}}\mathbf{y}_k \\ &= \mathbf{R}_k(\mathbf{x}_1+\mathbf{x}_2)+\tilde{\mathbf{n}}_k,\end{aligned} \tag{3}$$

where $\tilde{\mathbf{n}}_k = \mathbf{Q}_k^{\mathrm{H}}\mathbf{n}_k$. Note that since $\mathbf{Q}_k$ is unitary, $\tilde{\mathbf{n}}_k \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I})$. At layer m, the corresponding element of $\mathbf{z}_k$ is given by

$$z_{m;k} = \sum_{q=m}^{M} r_{m,q;k}(x_{m;1}+x_{m;2}) + \tilde{n}_{m;k}, \tag{4}$$

where $r_{m,q;k}$ represents the (m,q)th element of $\mathbf{R}_k$ and $\tilde{n}_{m;k}$ represents the mth element of $\tilde{\mathbf{n}}_k$. In this layered transmission scheme, user 1 is to perform SIC with respect to the user 2's signals as well as his/her signals in the other layers.

At layer M of user 1, we have

$$z_{M;1} = r_{M,M;1}(x_{M;1}+x_{M;2}) + \tilde{n}_{M;1}. \tag{5}$$

Suppose that user 1 can decode $\{x_{M;2}\}$ and perform SIC to decode $\{x_{M;1}\}$. At layer $M-1$, we have

$$\begin{aligned}z_{M-1;1} &= r_{M-1,M-1;1}(x_{M-1;1}+x_{M-1;2}) \\ &\quad + r_{M-1,M;1}(x_{M;1}+x_{M;2}) + \tilde{n}_{M-1;1}.\end{aligned}$$

Since $\{x_{M;1}\}$ and $\{x_{M;2}\}$ are available, they can be canceled, which results in the following signal:

$$\begin{aligned}&z_{M-1;1} - r_{M-1,M;1}(x_{M;1}+x_{M;2}) \\ &= r_{M-1,M-1;1}(x_{M-1;1}+x_{M-1;2}) + \tilde{n}_{M-1;1}.\end{aligned} \tag{6}$$

In this layer, user 1 is to decode $\{x_{M-1;2}\}$, and perform SIC to decode $\{x_{M-1;1}\}$.

At user 2, the SIC can be performed across layers assuming that the signals from user 1 are interference.

III. Optimal Power Allocation for Layered Transmissions With Instantaneous CSI

In this section, we consider an optimal power allocation for MIMO-NOMA with layered transmissions with known instantaneous CSI. Throughout this section, we assume block fading channels where a coded sequence is transmitted within a coherence time. While we assume that perfect CSI is available at the BS in this section, we will consider the case that statistical properties of the channels are available at the BS and study the power allocation to maximize the average sum rate in Section IV.

A. Optimal Power Allocation

For convenience, let

$$\alpha_m = |r_{m,m;1}|^2 \text{ and } \beta_m = |r_{m,m;2}|^2. \tag{7}$$

The signal powers are denoted by $P_m = \mathbb{E}[|x_{m;1}|^2]$ and $Q_m = \mathbb{E}[|x_{m;2}|^2]$. In addition, we assume that $\mathbb{E}[|\tilde{n}_{m;k}|^2] = N_0 = 1$ for normalization purposes.

Denote by $\eta_{m;k}$ the code rate of layer m for user k. Then, from (5), the code rate of layer M for user 2, $\eta_{M;2}$, has to satisfy the following inequality:

$$\eta_{M;2} \le \log_2\left(1+\frac{\alpha_M Q_M}{\alpha_M P_M + 1}\right). \tag{8}$$

Furthermore, at user 2, the following inequality is required:

$$\eta_{M;2} \le \log_2\left(1+\frac{\beta_M Q_M}{\beta_M P_M + 1}\right). \tag{9}$$

This implies that

$$\begin{aligned}\eta_{M;2} &\le \min\left\{\begin{array}{l}\log_2\left(1+\frac{\alpha_M Q_M}{\alpha_M P_M+1}\right), \\ \log_2\left(1+\frac{\beta_M Q_M}{\beta_M P_M+1}\right)\end{array}\right\} \\ &= \log_2\left(1+\frac{c_M Q_M}{c_M P_M+1}\right),\end{aligned} \tag{10}$$

where $c_M = \min\{\alpha_M, \beta_M\}$. At user 1, the code rate of the Mth layer signal of his/her signal needs to satisfy the following inequality:

$$\eta_{M;1} \le \log_2(1+\alpha_M P_M). \tag{11}$$

Thus, the (achievable) sum rate can be expressed as

$$\begin{aligned}&R(P_1, Q_1, \ldots, P_M, Q_M) \\ &= \sum_{m=1}^{M} \log_2\left(1+\frac{c_m Q_m}{c_m P_m+1}\right) + \log_2(1+\alpha_m P_m).\end{aligned} \tag{12}$$

Theorem 1: $R(P_1, Q_1, \ldots, P_M, Q_M)$ is concave in P_m and Q_m.

Proof: See Appendix A. ■

From Theorem 1, since the sum rate is concave, we may consider the sum rate maximization problem as follows:

$$\begin{aligned}&\max_{\{P_m\},\{Q_m\}} R(P_1, Q_1, \ldots, P_M, Q_M) \\ &\text{subject to} \sum_{m=1}^{M} P_m + Q_m \le P_{\mathrm{T}},\end{aligned} \tag{13}$$

where P_{T} is the total transmission power. Unfortunately, this sum rate maximization may lead to an undesirable solution in most cases. To see this, we can consider a special case of $M=1$ with $\alpha=\alpha_1$, $\beta=\beta_1$, and $c=\min\{\alpha_1,\beta_1\}$. In this case, the sum rate maximization problem becomes

$$\begin{aligned}&\max_{P,Q} \log_2\left(1+\frac{cQ}{cP+1}\right) + \log_2(1+\alpha P) \\ &\text{subject to } P+Q \le P_{\mathrm{T}}.\end{aligned} \tag{14}$$

Since user 1 is closer to the BS than user 2, we usually have $\alpha > \beta$ or $c = \beta$. Thus, the sum rate becomes

$$R(P,Q) = \log_2(1+\beta(P+Q)) + \log_2(1+\alpha P) - \log_2(1+\beta P).$$

To maximize the sum rate, we need to allocate the maximum power to $P+Q$, i.e., $P_{\mathrm{T}} = P+Q$, which results in

$$R(P,Q) = \log_2(1+\beta P_{\mathrm{T}}) + \log_2(1+\alpha P) - \log_2(1+\beta P).$$

Since $\log_2(1+\alpha P) - \log_2(1+\beta P)$ is an increasing function of P for given $\alpha > \beta$, we can see that the optimal power for user 1 is $P^* = P_{\mathrm{T}}$, while the optimal power for user 2 is $Q^* = 0$. This shows that the sum rate maximization does not lead to a fair power allocation between user 1 and user 2.

In order to be fair to user 2, it is necessary to impose the power constraints separately. To this end, we may have the total transmission power constraints to user 1 and user 2 as follows:

$$\sum_{m=1}^{M} P_m \le \bar{P} \text{ and } \sum_{m=1}^{M} Q_m \le \bar{Q}, \tag{15}$$

where $\bar{P}$ and $\bar{Q}$ are the maximum transmission powers to user 1 and user 2, respectively. Then, the sum rate maximization problem can be formulated as

$$\begin{aligned} &\max_{\{P_m\},\{Q_m\}} R(P_1, Q_1, \ldots, P_M, Q_M) \\ &\text{subject to 15.} \end{aligned} \tag{16}$$

B. Alternating Maximization

In this subsection, we consider an approach to find the solution of the sum rate maximization problem in (16).

For convenience, let

$$\begin{aligned} \mathcal{P} &= \{\mathbf{p} = [P_1 \ \ldots \ P_M]^{\mathrm{T}}] \,|\, P_m \ge 0, \ \sum_m P_m \le \bar{P}\} \\ \mathcal{Q} &= \{\mathbf{q} = [Q_1 \ \ldots \ Q_M]^{\mathrm{T}}] \,|\, Q_m \ge 0, \ \sum_m Q_m \le \bar{Q}\}. \end{aligned} \tag{17}$$

In addition, let $R(\mathbf{p}, \mathbf{q}) = R(P_1, Q_1, \ldots, P_M, Q_M)$. Then, to solve (16), we can consider the alternating maximization (AM) algorithm as follows:

$$\mathbf{q}^{(t+1)} = \operatorname*{argmax}_{\mathbf{q} \in \mathcal{Q}} R(\mathbf{p}^{(t)}, \mathbf{q}) \tag{18}$$

$$\mathbf{p}^{(t+1)} = \operatorname*{argmax}_{\mathbf{p} \in \mathcal{P}} R(\mathbf{p}, \mathbf{q}^{(t+1)}). \tag{19}$$

where $\mathbf{p}^{(t)}$ and $\mathbf{q}^{(t)}$ are the power vectors after iteration t for user 1 and user 2, respectively.

Since $\mathcal{P}$ and $\mathcal{Q}$ are convex sets and $R(\mathbf{p}, \mathbf{q})$ is concave, this alternating maximization, which can be seen as a two-block Gauss-Seidel method, can converge to the optimal solution for any initial vectors $\mathbf{p}^{(0)} \in \mathcal{P}$ and $\mathbf{q}^{(0)} \in \mathcal{Q}$ [20].

We now focus on the solution to the maximizations in (18) and (19). The maximization in (18) can be carried out using the water-filling theorem [5]. To see this, we need to consider the terms of $\{Q_m\}$ in $R(\mathbf{p}, \mathbf{q})$ for given $\mathbf{p}$ as follows:

$$\begin{aligned} A(\mathbf{q}; \mathbf{p}) &= \sum_m \log_2(1 + c_m P_m + c_m Q_m) \\ &= \sum_m \log_2((1 + c_m P_m)(1 + \bar{c}_m Q_m)) \\ &= \sum_m \log_2(1 + \bar{c}_m Q_m) + \text{Constant}, \end{aligned} \tag{20}$$

where $\bar{c}_m = \frac{c_m}{1 + c_m P_m}$. Thus, for given $\mathbf{p}$, we have

$$Q_m^*(\mathbf{p}) = \left(\lambda - \frac{1}{\bar{c}_m}\right)^+ = \left(\lambda - \frac{1 + c_m P_m}{c_m}\right)^+, \tag{21}$$

where $(x)^+ = \max\{0, x\}$ and λ is a Lagrange multiplier that can be decided to meet

$$\bar{Q} = \sum_m \left(\lambda - \frac{1 + c_m P_m}{c_m}\right)^+.$$

In (19), for given $\mathbf{q}$, we need to maximize $R(\mathbf{p}, \mathbf{q})$. Using the method of Lagrange Multipliers, we have the following unconstrained problem:

$$\max_{\mathbf{p}} R(\mathbf{p}, \mathbf{q}) - \nu \sum_m \mathbf{p}, \tag{22}$$

where ν is a Lagrange multiplier. Taking the first order derivative with respect to P_m, we have

$$B(P_m) = \nu, \tag{23}$$

where

$$B(P_m) = \frac{\alpha_m}{1 + \alpha_m P_m} + \frac{c_m}{1 + c_m Q_m + c_m P_m} - \frac{c_m}{1 + c_m P_m}.$$

It is easy to show that $B(P_m)$ is a decreasing function. Thus, for a given ν, we can find a unique P_m that satisfies (23), which is denoted by P_m^*. Note that $B(0) = \alpha_m - c_m + \frac{c_m}{1 + c_m Q_m}$. Thus, for $\nu > B_m(0)$, $P_m^* = 0$. The value of ν can be decided to hold $\sum_m P_m^* = \bar{P}$ using the bisection method based on the fact that $\sum_m P_m$ increases as ν decreases.

The computational complexity of the AM algorithm depends on the number of iterations and the complexity of the bisection method that is used to solve (18) and (19) (note that in the water-filling theorem to solve (18), we also need to use the bisection method to decide λ in (21)). The computational complexity of the bisection method is $O(-\log_2 \epsilon)$, where ϵ is the tolerance. From simulations, we observe that a few iterations (less than 5) are required for the AM algorithm to converge. For a fixed torelance and number of iterations, the complexity becomes linear in M, because for given λ and ν, the complexity to decide powers are proportional to M. Thus, the overall complexity is not high.

It is noteworthy that another advantage of layered transmissions over non-layered transmissions is a lower amount of CSI feedback for the power allocation. As shown above, for the power allocation, the BS needs to know the squared magnitudes of the diagonal elements of $\{\mathbf{R}_k\}$, i.e., $\{\alpha_m\}$ and $\{\beta_m\}$, not the whole channel matrices, $\{\mathbf{H}_k\}$. Thus, the number of the parameters for CSI feedback is M from each user, not M^2.

IV. Optimal Power Allocation to Maximize Average Sum Rate

In this section, we consider the power allocation to maximize the average sum rate with maximum transmission power constraints for MIMO-NOMA with layered transmissions once we derive a closed-form expression for the average sum rate. We also derive bounds on the average sum rate.

A. Derivation of Average Sum Rate and Power Optimization

From (12), the average sum rate for given $\mathbf{p}$ and $\mathbf{q}$ is given by

$$\begin{aligned}\bar{R}(\mathbf{p},\mathbf{q}) &= \mathbb{E}[(\mathbf{p},\mathbf{q})] \\ &= \sum_{m=1}^{M} \mathbb{E}\left[\log_2\left(1+\frac{c_m Q_m}{c_m P_m + 1}\right)\right] \\ &\quad + \mathbb{E}\left[\log_2(1+\alpha_m P_m)\right]. \end{aligned} \tag{24}$$

Throughout this section, we consider the following assumption to find the average sum rate.

[**A)**] We assume that the elements of $\mathbf{H}_k$ are iid and $[\mathbf{H}_k]_{n,m} \sim \mathcal{CN}(0,\sigma_k^2)$.

For convenience, define the normalized signal-to-noise ratio (SNR) for user 1 as $Z = \frac{\alpha_m}{\sigma_1^2}$ or $Z = \frac{\beta_m}{\sigma_2^2}$ for user 2. Then, under **A)** , we have the following probability density function (pdf) of Z [21], [22]:

$$f_m(z) = \frac{1}{(M-m)!} z^{M-m} e^{-z}, \ z \geq 0. \tag{25}$$

Let

$$C_{m;1}(P_m) = \mathbb{E}\left[\log_2(1+\alpha_m P_m)\right], \tag{26}$$

which is the ergodic capacity of layer m of user 1 for given P_m. Under **A)** , from [23], we have

$$\begin{aligned} C_{m;1}(P_m) &= \int_0^\infty \log_2(1+\sigma_1^2 P_m z) f_m(z) dz \\ &= \frac{e^{\frac{1}{\sigma_1^2 P_m}}}{\ln 2} \sum_{q=0}^{M-m} E_{q+1}\left(\frac{1}{\sigma_1^2 P_m}\right), \end{aligned} \tag{27}$$

where $E_q(x) = \int_1^\infty e^{-xy} y^{-q} dy$. For convenience, we define

$$\tau_n(x) = \frac{e^{\frac{1}{x}}}{\ln 2} \sum_{q=0}^{n-1} E_{q+1}\left(\frac{1}{x}\right) = \frac{e^{\frac{1}{x}}}{\ln 2} \sum_{q=1}^{n} E_q\left(\frac{1}{x}\right). \tag{28}$$

Thus, the average sum rate at user 1 with layered transmissions becomes

$$C_1(\mathbf{p}) = \sum_{m=1}^{M} C_{m;1}(P_m) = \sum_{m=1}^{M} \tau_{M-m+1}(\sigma_1^2 P_m). \tag{29}$$

This has been derived in [16].

The average sum rate of user 2 is a bit involved. Let

$$C_2(\mathbf{p},\mathbf{q}) = \sum_{m=1}^{M} C_{m;2}(P_m, Q_m), \tag{30}$$

where $C_{2,m}(P_m, Q_m)$ is the ergodic capacity of layer m of user 2 for given P_m and Q_m, which is given by

$$\begin{aligned} C_{2,m}(P_m,Q_m) &= \mathbb{E}\left[\log_2\left(1+\frac{c_m Q_m}{c_m P_m+1}\right)\right] \\ &= \kappa_m(P_m+Q_m) - \kappa_m(P_m). \end{aligned} \tag{31}$$

Here,

$$\kappa_m(x) = \mathbb{E}\left[\log_2(1+c_m x)\right]. \tag{32}$$

Theorem 2: Under **A)** , we have

$$\kappa_m(x) = \sum_{p=0}^{M-m} g_p\left(\frac{\sigma_1^2}{\sigma_2^2}, M-m+1\right) \tau_{M-m+p+1}(\phi x), \tag{33}$$

where $\phi = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ and

$$g_p(t;n) = \binom{p+n-1}{p} \frac{t^n + t^p}{(1+t)^{n+p}}. \tag{34}$$

Proof: See Appendix B. ■

Finally, from (29), (30), and (31), we can have a closed-form expression for the average sum rate as follows:

$$\begin{aligned} \bar{R}(\mathbf{p},\mathbf{q}) &= C_1(\mathbf{p}) + C_2(\mathbf{p},\mathbf{q}) \\ &= \sum_{m=1}^{M} \Big(\tau_{M-m+1}(\sigma_1^2 P_m) \\ &\quad + \kappa_m(P_m+Q_m) - \kappa_m(P_m)\Big) \end{aligned} \tag{35}$$

Using the closed-form expression for the average sum rate in (35), we can consider the following maximization of the average sum rate:

$$\max_{\mathbf{p}\in\mathcal{P},\mathbf{q}\in\mathcal{Q}} \bar{R}(\mathbf{p},\mathbf{q}). \tag{36}$$

According to Theorem 1, $R(\mathbf{p},\mathbf{q})$ is concave in $\mathbf{p}$ and $\mathbf{q}$. Since $\bar{R}(\mathbf{p},\mathbf{q})$ is the average of $R(\mathbf{p},\mathbf{q})$, it is also concave. Thus, we can use the AM algorithm to find the solution of (36). That is,

$$\mathbf{q}^{(t+1)} = \operatorname*{argmax}_{\mathbf{q}\in\mathcal{Q}} \bar{R}(\mathbf{p}^{(t)},\mathbf{q}) \tag{37}$$

$$\mathbf{p}^{(t+1)} = \operatorname*{argmax}_{\mathbf{p}\in\mathcal{P}} \bar{R}(\mathbf{p},\mathbf{q}^{(t+1)}). \tag{38}$$

Unlike the case of the maximization with known CSI, we do not have a closed-form solution of each maximization. Thus, we could resort to any gradient ascent algorithm [24] that requires the first order derivatives of a given objective function. In the following result, we derive the first order derivatives of $\bar{R}(\mathbf{p},\mathbf{q})$.

Theorem 3: Let $g_{p,m} = g_p\left(\frac{\sigma_1^2}{\sigma_2^2}, M-m+1\right)$ for notational convenience. Then, we have

$$\begin{aligned} \frac{d\bar{R}(\mathbf{p},\mathbf{q})}{dQ_m} &= \frac{d\kappa_m(P_m,Q_m)}{dQ_m} \\ &= \nu_m(P_m+Q_m), \end{aligned} \tag{39}$$

where

$$\begin{aligned} \nu_m(z) &= \frac{e^{\frac{1}{\phi z}}}{\phi z^2 \ln 2} \\ &\quad \times \sum_{p=0}^{M-m} g_{p,m} \sum_{q=0}^{M-m+p} \bar{E}_q\left(\frac{1}{\phi z}\right). \end{aligned} \tag{40}$$

Here,

$$\bar{E}_q(x) = E_q(x) - E_{q+1}(x).$$

In addition, we have

$$\frac{d\bar{R}(\mathbf{p},\mathbf{q})}{dP_m} = \sum_{q=0}^{M-m} \left(\frac{e^{\frac{1}{\sigma_1^2 P_m}}}{\sigma_1^2 P_m^2 \ln 2} \bar{E}_q \left(\frac{1}{\sigma_1^2 P_m} \right) + \nu_m(P_m + Q_m) - \nu_m(P_m) \right). \quad (41)$$

Proof: See Appendix C. ■

B. Bounds

In [11], the power allocation for MIMO-NOMA with non-layered transmissions is studied with the ergodic capacity. Although it is not clearly mentioned in [11], the derived ergodic capacity is an upper-bound. In particular, since

$$\mathbb{E}[\min\{X, Y\}] \le \min\{\mathbb{E}[X], \mathbb{E}[Y]\}, \quad (42)$$

the result in [11, Eq. (8)] becomes an upper-bound. Using (42), we can also find an upper-bound on the average sum rate of MIMO-NOMA with non-layered transmissions. In particular, an upper-bound on $C_2(\mathbf{p}, \mathbf{q})$ can be found as follows:

$$\begin{aligned} C_2(\mathbf{p},\mathbf{q}) &= \sum_{m=1}^{M} \mathbb{E}\left[\log_2\left(1 + \frac{c_m Q_m}{c_m P_m + 1}\right)\right] \\ &= \sum_{m=1}^{M} \mathbb{E}\left[\log_2\left(1 + \min\left\{ \frac{\alpha_m Q_m}{\alpha_m P_m + 1}, \frac{\beta_m Q_m}{\beta_m P_m + 1} \right\}\right)\right] \\ &\le \sum_{m=1}^{M} \mathbb{E}[\log_2(1 + \beta_m(P_m + Q_m))] \\ &\quad - \mathbb{E}[\log_2(1 + \beta_m P_m))] \\ &= \sum_{m=1}^{M} \tau_{M-m+1}\left(\sigma_2^2(P_m + Q_m)\right) \\ &\quad - \tau_{M-m+1}\left(\sigma_2^2 P_m\right), \end{aligned} \quad (43)$$

where the inequality is due to (42) and the assumption that $\sigma_1^2 > \sigma_2^2$. Thus, an upper-bound on $\bar{R}(\mathbf{p}, \mathbf{q})$ is given by

$$\bar{R}(\mathbf{p},\mathbf{q}) \le \sum_{m=1}^{M} \Big(\tau_{M-m+1}\left(\sigma_1^2 P_m\right) + \tau_{M-m+1}\left(\sigma_2^2(P_m + Q_m)\right) - \tau_{M-m+1}\left(\sigma_2^2 P_m\right) \Big). \quad (44)$$

We can have a lower-bound on $C_{2,m}(P_m, Q_m)$ as follows.

Theorem 4: Under **A)** , we have

$$C_{2,m}(P_m, Q_m) \ge \tilde{\tau}_{M-m+1}(\phi P_m, \phi Q_m), \quad (45)$$

where

$$\tilde{\tau}_n(x, y) = \tau_n(x + y) - \tau_n(x). \quad (46)$$

Proof: See Appendix D. ■

From (46), a lower-bound on $\bar{R}(\mathbf{p}, \mathbf{q})$ becomes

$$\bar{R}(\mathbf{p},\mathbf{q}) \ge \sum_{m=1}^{M} \Big(\tau_{M-m+1}(\sigma_1^2 P_m) + \tau_{M-m+1}(\phi(P_m + Q_m)) - \tau_{M-m+1}(\phi P_m) \Big). \quad (47)$$

From (44) and (47), we can see that the bounds can be tight if $\phi = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\sigma_2^2 \approx \sigma_2^2$. This is the case that $\sigma_1^2 \gg \sigma_2^2$.

Using the upper- and lower-bounds on $\bar{R}(\mathbf{p}, \mathbf{q})$, we can show that the scaling property of MIMO-NOMA is identical to that of MIMO [25], [26]. For convenience, we consider the equal power allocation, i.e., $P_m = \frac{\bar{P}}{M}$ and $Q_m = \frac{\bar{Q}}{M}$. For $C_1(\mathbf{p})$, we can consider the following approximation [17]:

$$e^x E_q(x) \approx \frac{1}{x + q}.$$

Using this, from (29), we have

$$\begin{aligned} C_1 &\approx \frac{1}{\ln 2} \sum_{m=1}^{M} \sum_{q=1}^{M-m+1} \frac{1}{\frac{M}{\sigma_1^2 \bar{P}} + q} \\ &\approx \frac{1}{\ln 2} \sum_{m=1}^{M} \ln\left(1 + (M - m + 1)\frac{\sigma_1^2 \bar{P}}{M}\right) \\ &= \frac{1}{\ln 2} \sum_{m=1}^{M} \ln\left(1 + m\frac{\sigma_1^2 \bar{P}}{M}\right) \\ &\approx \frac{1}{\ln 2} \int_0^M \ln\left(1 + x\frac{\sigma_1^2 \bar{P}}{M}\right) dx \\ &= M\left(\left(1 + \frac{1}{\sigma_1^2 \bar{P}}\right)\log_2(1 + \sigma_1^2 \bar{P}) - \frac{1}{\ln 2}\right). \end{aligned} \quad (48)$$

This shows that $C_1 = O(M)$, which is the scaling property of MIMO with *layered transmissions* and equal power allocation. Similarly, we can show that $C_2 = O(M)$ with both the upper- and lower-bounds in (44) and (47), respectively. Thus, we have $\bar{R}(\mathbf{p}, \mathbf{q}) = O(M)$ with equal power allocation, which shows that MIMO-NOMA with layered transmissions has the average sum rate that grows linearly with M.

V. Simulation Results

In this section, we present simulation results. We assume the channel model in Assumption **A)** with $\sigma_1^2 = 1$ and $\sigma_2^2 = \bar{\omega}$. In general, we have $\bar{\omega} \le 1$ as the distance between the BS and user 2 is longer than that between the BS and user 1.

A. Power Allocation With Known CSI

In this subsection, we consider the optimal power allocation for MIMO-NOMA with layered transmissions when the instantaneous CSI (i.e., $\{\alpha_m\}$ and $\{\beta_m\}$) is available at the BS. For comparison purposes, we consider the achievable rate of non-layered transmissions, which is given by

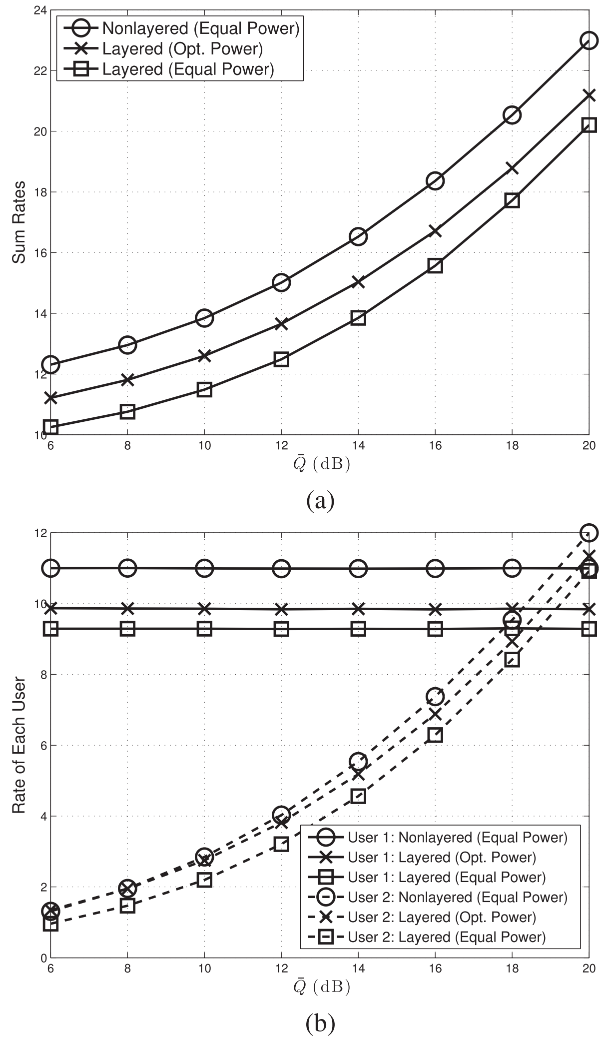

Fig. 1. Achievable rates for different total transmission powers of user 2, $\bar{Q}$ when $M = 6$, $\bar{P} = 6$ dB, and $\bar{\omega} = -6$ dB: (a) sum rates; (b) achievable rates of each user.

$$
\begin{aligned}
&R(\mathbf{C}_1, \mathbf{C}_2) \\
&= \min \left\{ \begin{array}{l} \log_2 \det(\mathbf{I} + (\mathbf{I} + \mathbf{H}_1 \mathbf{C}_1 \mathbf{H}_1^{\mathrm{H}})^{-1} \mathbf{H}_1 \mathbf{C}_2 \mathbf{H}_1^{\mathrm{H}}) \\ \log_2 \det(\mathbf{I} + (\mathbf{I} + \mathbf{H}_2 \mathbf{C}_1 \mathbf{H}_2^{\mathrm{H}})^{-1} \mathbf{H}_2 \mathbf{C}_2 \mathbf{H}_2^{\mathrm{H}}) \end{array} \right\} \\
&\quad + \log_2 \det(\mathbf{I} + \mathbf{H}_1 \mathbf{C}_1 \mathbf{H}_1^{\mathrm{H}}), \qquad (49)
\end{aligned}
$$

where $\mathbf{C}_k = \mathbb{E}[\mathbf{x}_k \mathbf{x}_k^{\mathrm{H}}]$. For non-layered transmissions, we only consider the case that $\mathbf{C}_1 = \frac{\bar{P}}{M}\mathbf{I}$ and $\mathbf{C}_2 = \frac{\bar{Q}}{M}\mathbf{I}$, i.e., equal power allocation. As mentioned earlier, if non-layered transmissions are used, users need to perform joint detection[1] for M symbols, which might be computationally expensive.

In Fig. 1, we show the performances of MIMO-NOMA with non-layered and layered transmissions with a fixed $\bar{P} = 6$ dB and $M = 6$ for different total transmission powers of user 2. For the case of non-layered transmissions, the equal power allocation is considered as mentioned earlier. For the case of layered transmissions, the optimal and equal power allocations are considered. While the non-layered transmission scheme can have a higher sum rate than the layered transmission scheme as shown in Fig. 1 (a), the optimal power allocation can provide a higher sum rate than the equal power allocation in the layered transmission scheme. In Fig. 1 (b), it is interesting to see that the rates of user 2 in the non-layered and layered transmission schemes are similar when $\bar{Q}$ is low, while the gap increases with $\bar{Q}$. This shows that the performance loss due to nulling at the users by the QR factorization for layered transmission becomes significant when the total allocated power, $\bar{Q}$, increases and it is not compensated by the optimal power allocation. To avoid this, we need to consider joint detection/decoding when $\bar{Q}$ is high at the expense of high computational complexity for joint detection.

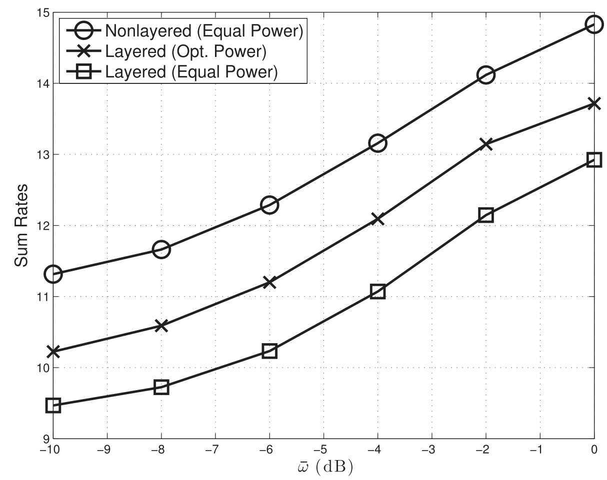

Fig. 2. Sum rates for different channel gains of user 2 when $M = 6$, $\bar{P} = 6$ dB, and $\bar{Q} = 6$ dB.

Fig. 2 shows the sum rates for different channel gains of user 2, $\bar{\omega}$, when $M = 6$, $\bar{P} = 6$ dB, and $\bar{Q} = 6$ dB. The optimal power allocation has about 1 bps/Hz more sum rate than the equal power allocation in the layered transmission scheme for all values of $\bar{\omega} = \sigma_2^2$.

In Fig. 3, assuming that the total transmission power, $\bar{P} + \bar{Q}$, is fixed and set to 6 dB, we obtain the sum rate for different fractions of $\bar{P}$. As $\bar{P}$ approaches the total transmission power or $\frac{\bar{P}}{\bar{P}+\bar{Q}}$ approaches 1, the sum rate increases as expected. Interestingly, we can see that the difference between the sum rates of the layered and non-layered transmission schemes is smaller as $\frac{\bar{P}}{\bar{P}+\bar{Q}}$ decreases. That is, if the total transmission power to user 2 is sufficiently larger than that to user 1, the sum rate of the layered transmission scheme can be close to that of the non-layered transmission scheme. Note that although the sum rate increases with $\bar{P}$, we may need to scarify the sum rate to pursue the fairness by increasing $\bar{Q}$ (i.e., allocating more power to user 2). In this case, the layered transmission scheme could be a good choice as it reduces the complexity of decoding at users and it provides a sum rate close to that with non-layered transmissions.

Fig. 4 shows the sum rates for different numbers of antennas, M, when $\bar{P} = \bar{Q} = 6$ dB and $\bar{\omega} = -6$ dB. An interesting

[1]For joint detection, we can use the maximum likelihood (ML) detector. Unfortunately, the computational complexity of the ML detection grows exponentially with M [18]. On the other hand, the complexity for signal detection based on the QR factorization is linear in M. Since the complexity of the QR factorization is $O(M^3)$, the proposed approach has a much lower computational complexity in the signal detection than the non-layered scheme for a large M.

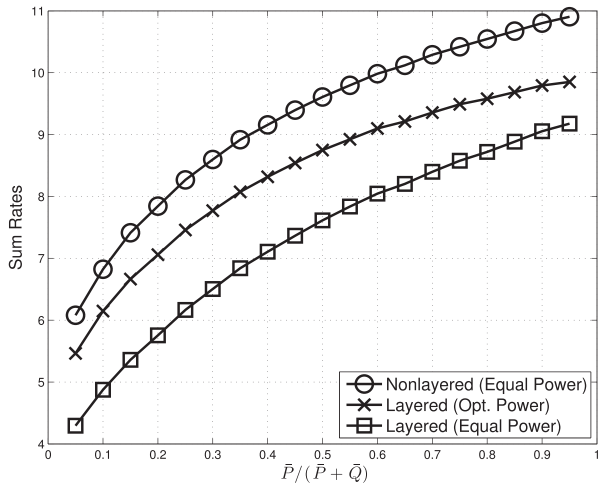

Fig. 3. Sum rates for different fractions of $\bar{P}$ when $M = 6$, $\bar{P} + \bar{Q} = 6$ dB, and $\bar{\omega} = -3$ dB.

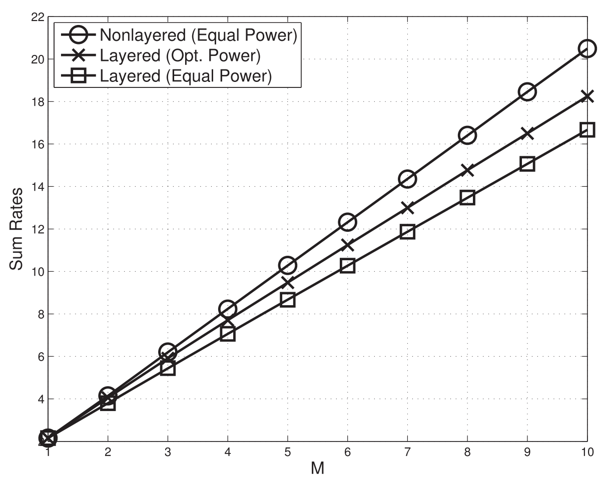

Fig. 4. Sum rates for different numbers of antennas, M, when $\bar{P} = \bar{Q} = 6$ dB and $\bar{\omega} = -6$ dB.

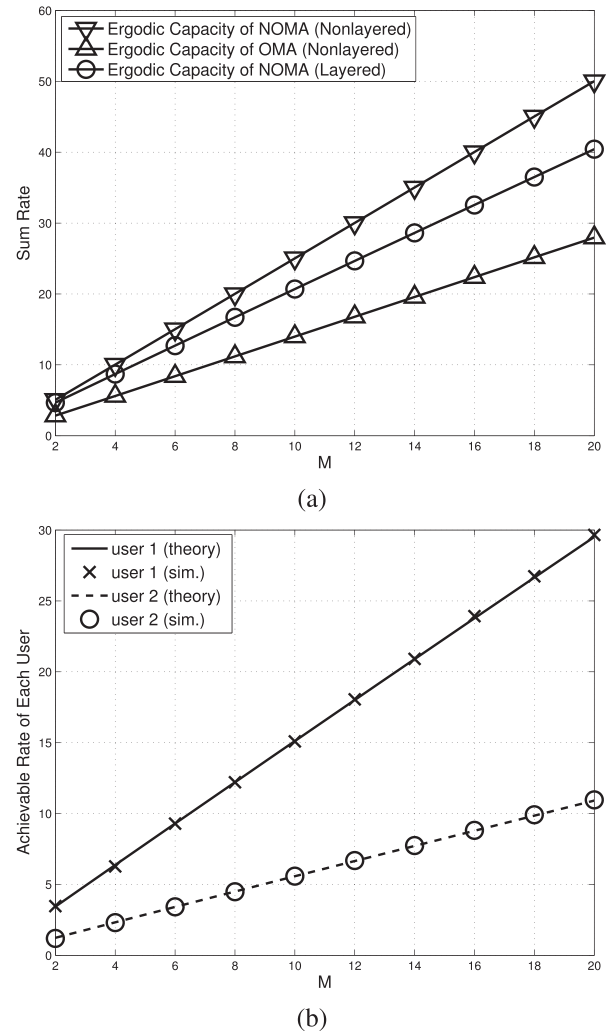

Fig. 5. Average sum rates of different MIMO transmission schemes for various numbers of antennas, M, when $\bar{P} = 6$, $\bar{Q} = 10$ dB, and $\bar{\omega} = -9$ dB: (a) sum rates; (b) achievable rates of each user (MIMO-NOMA with layered transmissions).

observation is that the scaling property in MIMO is also valid in MIMO-NOMA (with both layered and non-layered transmission schemes) with instantaneous CSI. Note that the gap between the sum rates of nonlayered and layered transmissions grows with M as shown in Fig. 4. In layered transmissions, due to nulling, the performance is degraded, although a linear complexity of detection/decoding in M is achieved. We can also claim that this performance degradation grows linearly with M due to the scaling property in both MIMO and MIMO-NOMA.

B. Power Allocation With Statistical CSI

In this subsection, we consider the power allocation to maximize the average sum rate for MIMO-NOMA with layered transmissions.

In Fig. 5 (a), we present the average sum rates (ergodic capacities) of MIMO-NOMA with layered and non-layered transmissions and MIMO-OMA with non-layered transmissions under equal power allocation when $\bar{P} = 6$, $\bar{Q} = 10$ dB, and $\bar{\omega} = -9$ dB. For MIMO-OMA, we consider TDMA where each user has half of a unit time slot. While MIMO-NOMA with layered transmissions has a lower sum rate than MIMO-NOMA with non-layered transmissions, it can have a higher sum rate than MIMO-OMA with non-layered transmissions. This shows that MIMO-NOMA with layered transmissions is a good choice that can enjoy a trade-off between the complexity (in terms of the complexity of the user's receiver) and performance (in comparison with the sum rate of MIMO-OMA).

Fig. 5 (b) shows the analytical results and simulation results of the average achievable rate of each user. It is shown that the derived analytical results agree with simulation results.

In Fig. 6, we show the average sum rates of MIMO-NOMA with layered transmissions under optimal and equal power allocations for different total transmission powers of user 2. We note that if $\bar{Q}$ is sufficiently large, the performance of equal power allocation approaches that of optimal power allocation. This behavior is similar to that of the water filling theorem [5]. In the water filling theorem, the optimal power allocation to maximize the sum rate approaches the equal power allocation as the total transmission power increases. Thus, we can see that

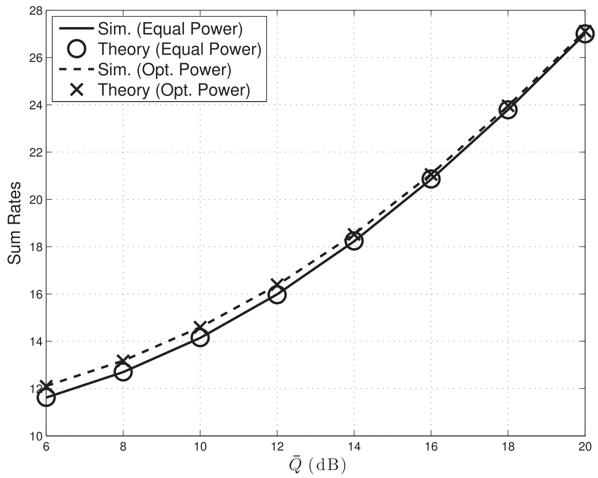

Fig. 6. Average sum rates for different total transmission powers of user 2, $\bar{Q}$ when $M = 6$, $\bar{P} = 6$ dB, and $\bar{\omega} = -6$ dB.

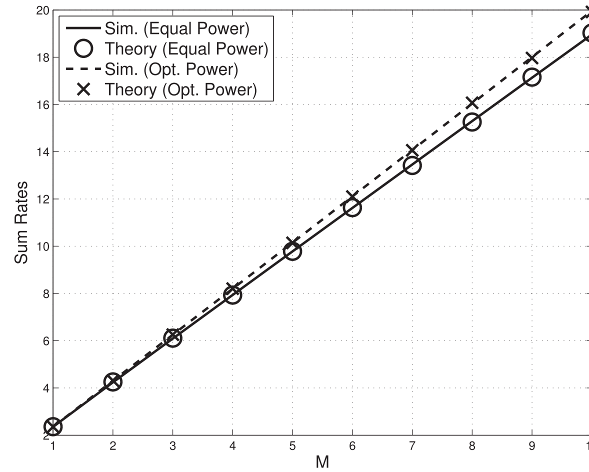

Fig. 8. Average sum rates for different numbers of antennas, M, when $\bar{P} = \bar{Q} = 6$ dB and $\bar{\omega} = -6$ dB.

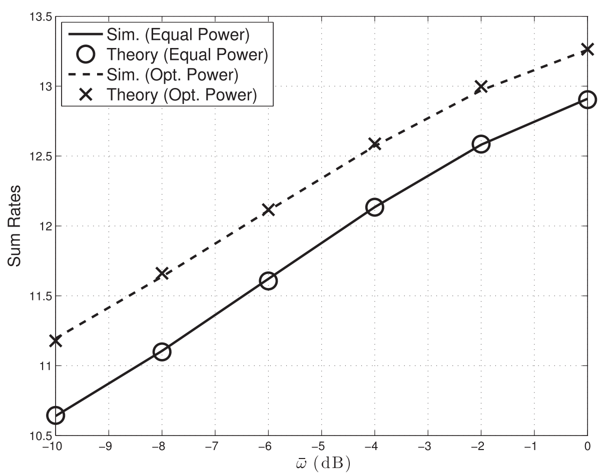

Fig. 7. Average sum rates for different channel gains of user 2 when $M = 6$, $\bar{P} = 6$ dB, and $\bar{Q} = 6$ dB.

the equal power allocation can be used in MIMO-NOMA with layered transmissions for sufficiently high total transmission powers.

In Fig. 7, the impact of $\bar{\omega}$ on the average sum rate is considered when $M = 6$, $\bar{P} = 6$ dB, and $\bar{Q} = 6$ dB. There is about 1 bps/Hz gap in terms of average sum rate between the performances of the equal and optimal power allocations. This gap decreases with $\bar{\omega}$. From this, we can see that the optimal power allocation would be more important as $\bar{\omega}$ is smaller or the distance between the BS and user 2 increases (provided that the distance between the BS and user 1 is fixed).

Fig. 8 shows the average sum rate for different numbers of antennas, M, when $\bar{P} = \bar{Q} = 6$ dB and $\bar{\omega} = -6$ dB. As discussed in Subsection IV-B, MIMO-NOMA has the scaling property that the average sum rate grows linearly with M. In Fig. 8, we can confirm this scaling property. We also note that the gap between the average sum rates of the equal and optimal power allocations grows with M. This shows that as the number of layers, M, increases, the optimal power allocation is more important.

VI. Conclusions

We studied the optimal power allocation to maximize the sum rate of MIMO-NOMA with layered transmissions when each user has a total transmission power constraint. Once we showed that the sum rate is concave in allocated powers to multiple layers of users, we proposed an approach for the optimal power allocation based on the AM algorithm for the cases of known instantaneous CSI and statistical CSI at the BS. To perform the optimal power allocation with statistical CSI, we derived a closed-form expression for the average sum rate as well as upper- and lower-bounds. From the upper- and lower-bounds on the average sum rate, we demonstrated that the scaling property holds (i.e., it was shown that the average sum rate grows linearly with the number of antennas).

While we have focused on the power allocation to maximize the sum rate with two users in this paper, there are also other issues in MIMO-NOMA. An extension to more than two users is an important issue and a generalization with beamforming would be practically interesting. Those issues will be studied as further research topics in the future.

Appendix A
Proof of Theorem 1

From (12), it can be shown that

$$R(P_1, Q_1, \ldots, P_M, Q_M) = \frac{1}{\ln 2} \sum_m \psi_m(P_m, Q_m), \tag{50}$$

where

$$\begin{aligned} \psi_m(P_m, Q_m) &= \ln(1 + c_m(P_m + Q_m)) \\ &\quad + \ln(1 + \alpha_m P_m) - \ln(1 + c_m P_m). \end{aligned} \tag{51}$$

To prove Theorem 1, it is sufficient to show that $\psi_m(P_m, Q_m)$ is concave. To this end, we need to find the Hessian matrix of $\psi_m(P_m, Q_m)$. It can be shown that

$$\frac{\partial^2 \psi_m}{\partial P_m^2} = \frac{\partial^2 \psi_m}{\partial Q_m^2} = \frac{\partial^2 \psi_m}{\partial P_m \partial Q_m} = -\frac{c_m^2}{(1 + c_m(P_m + Q_m))^2}.$$

Thus, the Hessian matrix is negative semi-definite, because for any vector $[x_1\ x_2]^{\mathrm{T}}$, we have

$$[x_1\ x_2] \begin{bmatrix} \frac{\partial^2 \psi_m}{\partial P_m^2} & \frac{\partial^2 \psi_m}{\partial P_m \partial Q_m} \\ \frac{\partial^2 \psi_m}{\partial P_m \partial Q_m} & \frac{\partial^2 \psi_m}{\partial Q_m^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -\frac{c_m^2}{(1 + c_m(P_m + Q_m))^2}(x_1 + x_2)^2 \le 0, \quad (52)$$

which implies that $\psi_m(P_m, Q_m)$ is concave in P_m and Q_m.

Appendix B
Proof of Theorem 2

Let $F_{\alpha_m}(x)$ and $F_{\beta_m}(x)$ denote the cumulative distribution functions (cdfs) of α_m and β_m, respectively. Since $c_m = \min\{\alpha_m, \beta_m\}$, the cdf of c becomes [27]

$$\begin{aligned} F_{c_m}(x) &= 1 - \Pr(\alpha_m > x)\Pr(\beta_m > x) \\ &= 1 - (1 - F_m(x/\sigma_1^2))(1 - F_m(x/\sigma_2^2)) \\ &= F_m(x/\sigma_1^2) + F_m(x/\sigma_2^2) \\ &\quad - F_m(x/\sigma_1^2)F_m(x/\sigma_2^2), \end{aligned} \quad (53)$$

where $F_m(x)$ is the cdf of Z and given by

$$F_m(x) = \frac{\gamma(M - m + 1, x)}{(M - m)!}. \quad (54)$$

Here, $\Gamma(x)$ is the Gamma function and $\gamma(n, x)$ is the lower incomplete Gamma function, which is given by $\gamma(n, x) = \int_0^x t^{n-1} e^{-t} dt$.

Thus, we have

$$f_{c_m}(x) = \frac{1}{\sigma_1^2}\left(1 - F_m\left(\frac{x}{\sigma_2^2}\right)\right) f_m\left(\frac{x}{\sigma_1^2}\right) + \frac{1}{\sigma_2^2}\left(1 - F_m\left(\frac{x}{\sigma_1^2}\right)\right) f_m\left(\frac{x}{\sigma_2^2}\right). \quad (55)$$

Then, it can be shown that

$$\begin{aligned} \kappa_m(x) &= \underbrace{\int_0^\infty \log_2(1 + \sigma_1^2 xt) f_m(t) dt}_{=\text{a)}} \\ &\quad - \underbrace{\int_0^\infty \log_@(1 + \sigma_1^2 xt) F_m\left(\frac{\sigma_1^2}{\sigma_2^2} t\right) f_m(t) dt}_{=\text{b)}} \\ &\quad + \underbrace{\int_0^\infty \log_2(1 + \sigma_2^2 xt) f_m(t) dt}_{=\text{c)}} \\ &\quad - \underbrace{\int_0^\infty \log_2(1 + \sigma_2^2 xt) F_m\left(\frac{\sigma_2^2}{\sigma_1^2} t\right) f_m(t) dt}_{=\text{d)}}. \end{aligned} \quad (56)$$

The first and third terms are

$$\text{a)} = \tau_{M-m+1}\left(\sigma_1^2 x\right) \text{ and c)} = \tau_{\text{M - m + 1}}\left(\sigma_2^2 \text{x}\right). \quad (57)$$

Let $\omega = \frac{\sigma_1^2}{\sigma_2^2}$ and $\bar{\omega} = \frac{\sigma_2^2}{\sigma_1^2}$. If n is a nonnegative integer, the lower incomplete Gamma function becomes

$$\gamma(n, t) = (n - 1)!\left(1 - e^{-t} \sum_{p=0}^{n-1} \frac{t^p}{p!}\right). \quad (58)$$

Using this, after some manipulations, we can find the second and forth terms as follows:

$$\begin{aligned} \text{b)} &= \tau_{M-m+1}(\sigma_1^2 x) \\ &\quad - \sum_{p=0}^{M-m} \tilde{g}_p(\omega; M - m + 1)\tau_{p+M-m+1}\left(\frac{\sigma_1^2 x}{1 + \omega}\right) \end{aligned} \quad (59)$$

$$\begin{aligned} \text{d)} &= \tau_{M-m+1}(\sigma_2^2 x) \\ &\quad - \sum_{p=0}^{M-m} \tilde{g}_p(\bar{\omega}; M - m + 1)\tau_{p+M-m+1}\left(\frac{\sigma_2^2 x}{1 + \bar{\omega}}\right), \end{aligned} \quad (60)$$

where

$$\tilde{g}_p(t; n) = \binom{p + n - 1}{p} \frac{t^p}{(1 + t)^{n+p}}. \quad (61)$$

Since $\bar{\omega} = \frac{1}{\omega}$, we can show that $\frac{\sigma_1^2}{1+\omega} = \frac{\sigma_2^2}{1+\bar{\omega}} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} = \phi$. Thus, we have

$$\begin{aligned} \kappa_m(x) &= \text{a)} - \text{b)} + \text{c)} - \text{d)} \\ &= \sum_{p=0}^{M-m} \left(\tilde{g}_p(\omega; M - m + 1) + \tilde{g}_p(\bar{\omega}; M - m + 1)\right) \\ &\quad \times \tau_{p+M-m+1}\left(\frac{\sigma_1^2\sigma_2^2 x}{\sigma_1^2 + \sigma_2^2}\right). \end{aligned} \quad (62)$$

Finally, noting that $\tilde{g}_p(\omega; n) + \tilde{g}_p(\bar{\omega}; n)$ becomes $g_p(\omega; n)$ in (34) and from (62), we have (33). This completes the proof.

Appendix C
Proof of Theorem 3

The first order derivative of $\tau_n(x)$ is obtained in [16] as

$$\begin{aligned} \frac{d\tau_n(x)}{dx} &= \frac{e^{\frac{1}{x}}}{x^2 \ln 2} \sum_{q=0}^{n-1}\left(E_q\left(\frac{1}{x}\right) - E_{q+1}\left(\frac{1}{x}\right)\right) \\ &= \frac{e^{\frac{1}{x}}}{x^2 \ln 2} \sum_{q=0}^{n-1} \bar{E}_q\left(\frac{1}{x}\right). \end{aligned} \quad (63)$$

The first order derivatives in (39) and (41) can be obtained by using (63). Since the derivations are straightforward and tedious, we omit them.

Appendix D
Proof of Theorem 4

To find a lower-bound on $C_{2,m}(P_m, Q_m)$, we rewrite (31) as

$$C_{2,m}(P_m, Q_m) = \sum_{p=0}^{M-m} g_p(\omega, M-m+1) \times \tilde{\tau}_{M-m+1+p}(\phi P_m, \phi Q_m). \quad (64)$$

From (26)–(28), and (46), we can show that

$$\tilde{\tau}_n(x, y) = \tau_n(x+y) - \tau_n(x) = \mathbb{E}\left[\log_2\left(1 + \frac{\alpha_{M+1-n} y}{1 + \alpha_{M+1-n} x}\right)\right]. \quad (65)$$

Noting that $\log_2\left(1 + \frac{ay}{1+ax}\right)$ is an increasing function of a for given $x, y > 0$, according to [28], it follows

$$\mathbb{E}\left[\log_2\left(1 + \frac{\alpha_m y}{1 + \alpha_m x}\right)\right] \geq \mathbb{E}\left[\log_2\left(1 + \frac{\alpha_{m+1} y}{1 + \alpha_{m+1} x}\right)\right], \quad (66)$$

because α_m is stochastically larger than α_{m+1} due to

$$Pr(\alpha_{m+1} \leq x) = F_{m+1}\left(\frac{x}{\sigma_1^2}\right) > \Pr(\alpha_m \leq x) = F_m\left(\frac{x}{\sigma_1^2}\right),$$

where $F_m(x)$ is defined in (54). Consequently, we have

$$\tilde{\tau}_n(x, y) \leq \tilde{\tau}_{n+1}(x, y). \quad (67)$$

Applying (67) to (64), we have

$$C_{2,m}(P_m, Q_m) \geq \sum_{p=0}^{M-m} g_p(\omega, M-m+1) \times \tilde{\tau}_{M-m+1}(\phi P_m, \phi Q_m). \quad (68)$$

In (68), we can show that

$$\begin{aligned} &\sum_{p=0}^{M-m} g_p(\omega, M-m+1) \\ &= \sum_{p=0}^{M-m} \binom{M-m+p}{p} \frac{\omega^{M-m+1} + \omega^p}{(1+\omega)^{M-m+1+p}} \\ &= \sum_{p=0}^{M-m} \binom{M-m+p}{p} \\ &\quad \times \left[z^{M-m+1}(1-z)^p + (1-z)^{M-m+1} z^p\right] \\ &= 1, \end{aligned} \quad (69)$$

where $z = \frac{\omega}{1+\omega}$ and the last equality is due to [29]. Finally, substituting (69) into (68), we have (45), which completes the proof.

References

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.

[2] B. Kim *et al.*, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE MILCOM*, Nov. 2013, pp. 1278–1283.

[3] A. G. Perotti and B. M. Popovic, "Non-orthogonal multiple access for degraded broadcast channels: RA-CEMA," ArXiv e-prints, Oct. 2014.

[4] 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, Jan. 2015.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[6] R. Gallager, "An inequality on the capacity region of multiple access multipath channels," in *Communications and Cryptography: Two Sides and One Tapestry*, R. E. Blahut, D. J. Costello, U. Maurer, and T. Mittelholzer, Eds. Norwell, MA, USA: Kluwer, 1994, pp. 129–139.

[7] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313–316, Feb. 2014.

[8] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. E98.B, no. 3, pp. 403–414, 2015.

[9] 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.

[10] 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.

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

[12] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[13] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. IEEE Int. Symp. Signals Syst. Electron. (ISSSE'98)*, Sep. 1998, pp. 295–300.

[14] G. Foschini, D. Chizhik, M. Gans, C. Papadias, and R. Valenzuela, "Analysis and performance of some basic space-time architectures," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 3, pp. 303–320, Apr. 2003.

[15] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[16] J. Choi, "Capacity/throughput optimization for H-BLAST with SC receiver over MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 1016–1024, Mar. 2008.

[17] E. Biglieri, *Coding for Wireless Channels*. New York, NY, USA: Springer, 2005.

[18] J. Choi, *Optimal Combining and Detection*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[19] M. Hanif, Z. Ding, T. Ratnarajah, and G. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2015.

[20] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, 2000.

[21] A. Edelman, "Eigenvalues and condition numbers of random matrices," Ph.D. dissertation, MIT, Dept. of Math., Massachusetts Inst. of Tech., Cambridge, MA, USA, 1989.

[22] X. Zhang and S.-Y. Kung, "Capacity analysis for parallel and sequential MIMO equalizers," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2989–3002, Nov. 2003.

[23] H. Shin and J. H. Lee, "Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2636–2647, Oct. 2003.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[25] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311–335, Mar. 1998.

[26] I. Telatar, "Capacity of multiple-antenna gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–595, Nov. 1999.

[27] H. A. David, *Order Statistics*. New York, NY, USA: Wiley, 1980.

[28] M. Shaked and G. Shanthikumar, *Stochastic Orders*. New York, NY, USA: Springer, 2007.

[29] M. Beeler, R. W. Gosper, and R. H. Schroeppel, "Item 42," Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA, Tech. Memo AIM-239, Feb. 1972

Jinho Choi (SM'02) was born in Seoul, South Korea. He received B.E. degree (*magna cum laude*) in electronics engineering from Sogang University, Seoul, South Korea, in 1989 and the M.S.E. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1991 and 1994, respectively. He is a Professor with Gwangju Institute of Science and Technology (GIST). Prior to joining GIST in 2013, he was a Professor/Chair in Wireless with the College of Engineering, Swansea University, Swansea, U.K. His research interests include wireless communications and array/statistical signal processing. He authored two books (Cambridge University Press, 2006 and 2010). Currently, he has been an Associate Editor of the IEEE COMMUNICATIONS LETTERS, and an Editor of *Journal of Communications and Networks* (JCN) since 2005, and served as an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2005 to 2007 and *ETRI* journal. He was the recipient of the 1999 Best Paper Award for Signal Processing from EURASIP and the 2009 Best Paper Award from WPMC (Conference).