Design of Downlink Hybrid NOMA Transmission

Zhiguo Ding, *Fellow, IEEE*, Robert Schober, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—The aim of this paper is to develop hybrid non-orthogonal multiple access (NOMA) assisted downlink transmission. First, for the single-input single-output (SISO) scenario, i.e., each node is equipped with a single antenna, a novel hybrid NOMA scheme is introduced, where NOMA is implemented as an add-on of a legacy time division multiple access (TDMA) network. Because of the simplicity of the SISO scenario, analytical results can be developed to reveal important properties of downlink hybrid NOMA. For example, in the case that the users' channel gains are ordered and the durations of their time slots are the same, downlink hybrid NOMA is shown to always outperform TDMA, which is different from the existing conclusion for uplink hybrid NOMA. Second, the proposed downlink SISO hybrid NOMA scheme is extended to the multiple-input single-output (MISO) scenario, i.e., the base station has multiple antennas. For the MISO scenario, near-field communication is considered to illustrate how NOMA can be used as an add-on in legacy networks based on space division multiple access and TDMA. Simulation results verify the developed analytical results and demonstrate the superior performance of downlink hybrid NOMA compared to conventional orthogonal multiple access.**

***Index Terms*—Downlink hybrid non-orthogonal multiple access (NOMA), space division multiple access, near-field communication, resolution of near-field beamforming.**

I. Introduction

Multiple access techniques can be viewed as the foundation stone of modern mobile networks, since the design of many crucial components of mobile networks, such as scheduling, resource allocation, channel estimation and signal detection, depends on which multiple access technique is used [1]. In the sixth-generation (6G) era, non-orthogonal multiple access (NOMA) has already received considerable attention due to its superior spectral efficiency, compared to conventional orthogonal multiple access (OMA) [2]–[5].

Unlike most existing works, e.g. [6]–[10], which viewed NOMA and OMA as competing systems, this paper considers NOMA as an add-on of OMA, which yields the following two benefits [11]. One is to shed light on the design of a unified framework for next-generation multiple access, and the other is to develop a vision for how NOMA can be integrated into existing wireless systems, which are based on OMA. We note that there are some existing works that have recognized the importance of allowing NOMA and OMA to co-exist. For example, in [12]–[16], user clustering has been carried out, where NOMA was implemented among the users within the same cluster and OMA was used to avoid inter-cluster interference. Similarly, in [17]–[19], various schemes have been proposed to ensure that a user can intelligently switch between the NOMA and OMA modes. While these existing approaches might realize a sophisticated coexistence between NOMA and OMA, they cannot ensure that NOMA is used as a simple add-on of OMA, which causes major disruptions to OMA based legacy networks, i.e., these NOMA approaches cannot be straightforwardly implemented in the currently deployed OMA networks.

Z. Ding is with Khalifa University, Abu Dhabi, UAE, and University of Manchester, Manchester, M1 9BB, UK. R. Schober is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg (FAU), Germany. H. V. Poor is with Princeton University, Princeton, NJ 08544, USA.

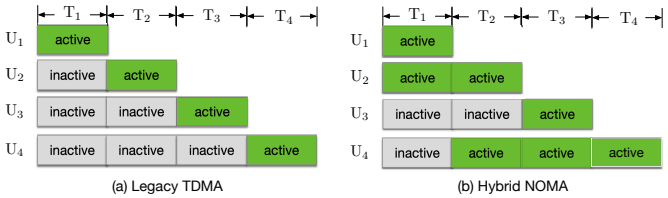


Fig. 1. Illustration of the key idea of hybrid NOMA transmission.

Hybrid NOMA, a concept originally developed in mobile edge computing (MEC) networks [20]–[22], can ensure that NOMA is implemented as an effective add-on of OMA. The key idea of hybrid NOMA can be illustrated by the four-user example shown in Fig. 1. In particular, with conventional time division multiple access (TDMA), the four users, denoted by U_m, $1 \leq m \leq 4$, are served in four time slots individually, as shown in Fig. 1(a). By using hybrid NOMA, the four users are still scheduled to finish their data transmission as in OMA, i.e., U_m finishes its transmission by the end of the m-th time slot, as shown in Fig. 1(b). Unlike TDMA, the use of hybrid NOMA ensures that a user can also use the time slots which are allocated to other users in TDMA, e.g., U_4 can use not only its own TDMA time slot, i.e., the fourth time slot, but also the second and third time slots which belong exclusively to U_2 and U_3 in TDMA. Because the users have more flexibility to transmit, naturally hybrid NOMA can outperform TDMA. Time-slot allocation for hybrid NOMA can be effectively accomplished by applying multi-time-slot power allocation. This is beneficial from the optimization perspective, since the optimal solution of power allocation is less challenging to obtain compared to time-slot allocation which is an integer programming problem. We note that this multi-time-slot optimization feature makes hybrid NOMA different from the existing single-time-slot non-hybrid NOMA schemes considered in [6]–[9], [11].

Because hybrid NOMA was originally developed for MEC, all existing works on hybrid NOMA focused on uplink transmission. In particular, the use of hybrid NOMA ensures that multiple users can cooperate with each other for offloading (or transmiting) their computation tasks to a base station [20]. As shown in [20]–[22], if the energy consumption of MEC offloading is used as the performance metric, uplink hybrid NOMA yields better performance than pure NOMA, and for the two-user special case, hybrid NOMA is shown to outperform OMA if one user's task deadline is less than two times of the other user's task deadline. Uplink hybrid NOMA has also been shown to outperform OMA, if energy efficiency is used as the performance metric [23], [24]. The performance of uplink hybrid NOMA can be further improved by applying intelligent reflecting surfaces and deep reinforcement learning as shown in [25], [26]. The use of uplink hybrid NOMA

has also been shown to be beneficial to improve the secrecy performance of MEC offloading [27].

Unlike the aforementioned uplink hybrid NOMA works, this paper aims to design downlink hybrid NOMA transmission. The contributions of this paper are listed as follows:

- For the single-input single-output (SISO) scenario, i.e., both the base station and the users are equipped with a single antenna, a new hybrid NOMA assisted downlink transmission scheme is proposed. In particular, it is assumed that there exists a TDMA based legacy network, and with hybrid NOMA, the users are encouraged to use the time slots which they would not have access to in TDMA. In addition, a multi-objective energy consumption minimization problem is formulated, and solved by using successive resource allocation [20].
- The properties of the obtained power allocation solutions are analyzed to unveil the important features of hybrid NOMA downlink transmission. For example, for the case that the users' channel gains are ordered and the durations of their time slots are the same, downlink hybrid NOMA is shown to always outperform OMA. This conclusion is different from the one previously reported for the uplink case [22]. In addition, the obtained analytical results show that it is optimal for each user to use the same transmit power across all NOMA time slots, and the users' accumulated transmit powers on different time slots are the same. Furthermore, the solution obtained from successive resource allocation is shown to be a Pareto-optimal solution of the formulated multi-objective energy minimization problem.
- The developed downlink SISO hybrid NOMA scheme is then extended to the multiple-input single-output (MISO) scenario, i.e., the base station has multiple antennas and each user is equipped with a single antenna. Unlike the SISO network, the legacy downlink MISO network is based on both TDMA and space-division multiple access (SDMA). In particular, it is assumed that there exist two groups of users, where the users in each group are served simultaneously via SDMA, and TDMA is used to avoid inter-group interference. In order to ensure the compatibility to the legacy network, downlink hybrid NOMA is used to realizes beam sharing, i.e., one group of users can use the spatial beams preconfigured for the other group of users. In order to demonstrate the feasibility of hybrid NOMA, near-field downlink transmission is considered as an illustrative example, as in general the accurate beamfocusing in near-field communications can make beam sharing difficult.
- For the MISO scenario, an energy consumption minimization problem is first formulated. Compared to the case of SISO downlink hybrid NOMA, the considered energy minimization problem in the MISO case is more challenging. This is due to the fact that beam sharing leads to potential inter-beam interference, which makes the formulated energy minimization problem non-convex. By applying successive convex approximation (SCA), a low-complexity sub-optimal power allocation solution is obtained and shown to outperform the OMA solution for various simulation setups.

The remainder of this paper is organized as follows. In Section II, the SISO scenario is considered, where the design of downlink hybrid NOMA transmission is studied, and the properties of SISO hybrid NOMA power allocation are unveiled. In Sections III, the MISO scenario is focused on, where the combination of SDMA and NOMA is investigated, and an energy consumption minimization problem is formulated and solved. In Section IV, simulation results are presented to demonstrate the performance of hybrid NOMA, and the paper is concluded in Section V.

II. Downlink SISO Hybrid NOMA Transmission

In this section, the application of hybrid NOMA in SISO downlink transmission is studied. Due to the simplicity of the SISO scenario, an insightful understanding of the key features of hybrid NOMA assisted downlink transmission can be obtained as will be shown in the following.

A. Description of SISO Hybrid NOMA Transmission

Consider the scenario with a legacy SISO TDMA network with M users, denoted by U_m, where the base station serves U_m in the m-th time slot, denoted by T_m. The key idea of hybrid NOMA is to encourage spectrum sharing among the users, where a user can have access to multiple time slots, as shown in Fig. 1. For illustrative purposes, it is assumed that in T_i, $(M-i+1)$ users, i.e., U_m, $i \leq m \leq M$, are served simultaneously. For example, in T_1, all M users are served simultaneously, whereas in T_M, only U_M is served [1]. Therefore, in T_i, U_m receives the following signal:

$$y_{m,i} = h_m \sum_{j=i}^{M} \sqrt{P_{j,i}} s_{j,i} + n_{m,i}, \tag{1}$$

where h_m, $P_{m,i}$, $s_{m,i}$, and $n_{m,i}$ denote U_m's channel gain, transmit power, transmit signal and received noise in T_i, respectively. Quasi-static fading is assumed, i.e., each user's channel gain remains constant within one time frame consisting of M time slots. Furthermore, the base station is assumed to have access to the users' channel state information (CSI). In practice, this CSI assumption can be realized by asking each user to first perform channel estimation based on the pilot signals broadcasted by the base station, and then feed the CSI back to the base station via a reliable feedback channel.

In T_i, U_m observes the signals of $(M-i+1)$ users, and the following successive interference cancellation (SIC) is carried out. In particular, in each time slot, U_M's signal is always the first to be decoded, then U_{M-1}'s signal is decoded. In other words, a descending decoding order is used, e.g., U_m's signal is decoded before U_j's, $m > j$. This descending decoding order ensures that in T_m, U_m's signal can be decoded in the last SIC stage without interference, i.e., as if U_m solely occupied T_m via OMA.

[1]How the users are scheduled depends on their quality of service requirements, the priority of the network traffic, and the features of the legacy network. For example, for MEC applications, the users can be ordered according to the urgency of their computation tasks.

Given this SIC decoding order, in T_i, U_k can decode U_m's signal with the following achievable data rate [2]:

$$R_{m,i}^{k}=\log\left(1+\frac{|h_k|^2P_{m,i}}{|h_k|^2\sum_{j=i}^{m-1}P_{j,i}+1}\right), \tag{2}$$

for $i\leq k\leq m$ and $1\leq i\leq m$, where the noise power is assumed to be normalized. As a result, by using hybrid NOMA, the achievable data rate of U_m at T_i is give by

$$R_{m,i}=\min\left\{R_{m,i}^{i},\cdots,R_{m,i}^{m}\right\}. \tag{3}$$

Remark 1: Downlink hybrid NOMA is a general framework, of which conventional OMA is a special case. In particular, U_m can choose $P_{m,i}=0$, $i\leq m$, and $P_{m,m}\neq 0$. This means that U_m uses T_m only and hence pure OMA is adopted since T_m is allocated to U_m in OMA, as illustrated in Fig. 1(a). By adjusting the power allocation coefficients, $P_{m,i}$, the use of downlink hybrid NOMA can ensure that each user fully benefits from the advantages of both NOMA and OMA transmissions.

B. Power Allocation for SISO Hybrid NOMA Transmission

The users' energy consumption will be used as the metric for performance evaluation, as explained in the following. With downlink hybrid NOMA, each user can have access to multiple time slots, but an improper use of these time slots can lead to a surge in energy consumption, e.g., when a user's transmit power was mistakenly chosen to be large in a time slot when its channel condition is poor.

The energy minimization problem considered in this paper can be formulated as follows:

$$\min_{P_{m,i}\geq 0}\quad \mathbf{E}\triangleq\begin{bmatrix}E_1 & \cdots & E_M\end{bmatrix}^T \tag{P1a}$$

$$s.t.\quad \sum_{i=1}^{m}R_{m,i}T\geq N_b, 1\leq m\leq M, \tag{P1b}$$

where it is assumed that each user needs to receive the same amount of data nats, denoted by N_b, $E_m=\sum_{i=1}^{m}P_{m,i}T$, and T denotes the duration of each time slot.

Problem (P1) is a multi-objective optimization problem, and hence challenging to solve. As pointed out in [20], the successive nature of SIC can be used to carry out successive resource allocation, as described in the following. In particular, U_1's power allocation coefficient, $P_{1,1}$, is first optimized, by assuming that the other users' parameters are fixed. Then, U_m's power allocation coefficients, $P_{m,i}$, $1\leq i\leq m$, can be optimized by assuming U_j's coefficients are fixed, $j>i$. The optimality of successive resource allocation will be discussed later after the closed-form expressions of the optimal power allocation coefficients are obtained.

[2]For notational simplicity, it is assumed that $\sum_{j=m}^{m-1}P_{j,i}=0$, and the natural logarithm is used for the data rate expressions.

By using this successive resource allocation approach, U_m's power allocation coefficients can be obtained by solving the following simplified optimization problem:

$$\min_{P_{m,i}\geq 0}\quad \sum_{i=1}^{m}P_{m,i} \tag{P2a}$$

$$s.t.\quad \sum_{i=1}^{m}R_{m,i}\geq R, \tag{P2b}$$

where $R=\frac{N_b}{T}$. It is straightforward to show that problem (P2) is a convex optimization problem, and hence can be solved by using off-shelf optimization solvers. However, it is challenging to obtain a closed-form expression for the optimal solution of problem (P2), mainly due to the dynamic nature of hybrid NOMA power allocation. For example, U_m might choose to transmit in a few non-consecutive time slots, and keep silent in the other time slots.

C. Properties of SISO Hybrid NOMA Power Allocation

In order to obtain an insightful understanding of the properties of downlink hybrid NOMA, two power allocation solutions for two special cases are provided. The first case is the pure OMA solution, $P_{m,m}^{\mathrm{O}}=\frac{e^R-1}{|h_m|^2}$ and $P_{m,i}^{\mathrm{O}}=0$ for $i<m$. We note that the OMA solution is always a feasible solution of problem (P2), but not necessarily the optimal solution. The solution for the second case of interest is presented in the following lemma.

Lemma 1. *Without the constraint of $P_{m,i}^{\mathrm{H}}\geq 0$, an optimal solution for the energy minimization problem shown in* (P2) *is given by*

$$P_{m,i}^{\mathrm{H}}=\left(\frac{e^R}{\prod_{p=1}^{m}\frac{|\bar{h}_m|^2}{|\bar{h}_m|^2\sum_{j=p}^{m-1}P_{j,p}+1}}\right)^{\frac{1}{m}}-\sum_{j=i}^{m-1}P_{j,i}^{\mathrm{H}}-\frac{1}{|\bar{h}_{m,i}|^2}, \tag{4}$$

where $|\bar{h}_{m,i}|^2=\min\left\{|h_i|^2,\cdots,|h_m|^2\right\}$.

Proof. See Appendix A. □

Because Lemma 1 is obtained by omitting the constraint $P_{m,i}\geq 0$, it is possible that $P_{m,i}^{\mathrm{H}}\leq 0$, i.e., Lemma 1 cannot be used for the general case. Instead, off-shelf optimization solvers should be used for the general case to find the optimal solution of problem (P2). The remainder of the section is to show that $P_{m,i}^{\mathrm{H}}>0$ holds in (4) for the special case, when the users' channel gains are ordered. The fact that $P_{m,i}^{\mathrm{H}}>0$ is significant since it means that Lemma 1 yields the optimal solution of problem (P2), and downlink hybrid NOMA outperforms OMA. To facilitate the performance analysis, an important feature of hybrid NOMA power allocation is established first.

Lemma 2. *Consider the special case when the users are ordered according to their channel gains (i.e., $|h_m|^2>$*

$|h_{m+1}|^2$). *If the downlink hybrid NOMA power allocation in* (4) *is adopted by* U_i, $1 \leq i \leq m$, *the following equality holds*

$$\sum_{i=1}^{m} P_{i,1}^{\mathrm{H}} = \cdots = \sum_{i=m-1}^{m} P_{i,m-1}^{\mathrm{H}} = P_{m,m}^{\mathrm{H}}. \tag{5}$$

Proof. See Appendix B. □

Remark 2: The term $\sum_{i=j}^{m} P_{i,j}^{\mathrm{H}}$ can be viewed as the accumulated interference in the j-th time slot. Lemma 2 reveals that the use of downlink hybrid NOMA ensures that the accumulated interference in different time slots is identical. A conclusion similar to Lemma 2 has been previously reported for hybrid NOMA uplink transmission [20]–[22].

With the help of Lemma 2, the optimality of the downlink hybrid NOMA power allocation in Lemma 1 can be established as shown in the following lemma.

Lemma 3. *For the considered special case with ordered channel gains, the solution shown in Lemma 1 is an optimal solution of problem* (P2).

Proof. See Appendix C. □

We note that Lemma 3 is not sufficient to show the superiority of downlink hybrid NOMA over OMA, since the OMA solution could also be an optimal solution of problem (P2).

Lemma 4. *For the considered special case with ordered channel gains, there exists a single optimal solution for problem* (P2).

Proof. See Appendix D. □

Based on Lemmas 3 and 4, the following corollary can be obtained straightforwardly.

Corollary 1. *For the considered special case with ordered channel gains, downlink hybrid NOMA always outperforms OMA.*

Remark 3: Corollary 1 shows that there is a unique difference between uplink and downlink hybrid NOMA. In particular, for the scenario considered in Corollary 1, uplink OMA outperforms uplink hybrid NOMA, as illustrated by the following two-user example. Similar to problem (P1), an energy minimization problem for hybrid NOMA uplink transmission can be formulated as follows:

$$\min \quad P_{2,1} + P_{2,2} \tag{P3a}$$

$$s.t. \quad \log\left(1 + \frac{|h_2|^2 P_{2,1}}{|h_1|^2 P_{1,1} + 1}\right) + \log\left(1 + |h_2|^2 P_{2,2}\right) \geq R. \tag{P3b}$$

By applying the Karush–Kuhn–Tucker (KKT) conditions, the optimal solutions of problem (P3) are give by [28]

$$P_{2,1} = \left(\frac{e^R\left(|h_1|^2 P_{1,1} + 1\right)}{|h_2|^2}\right)^{\frac{1}{2}} - \frac{|h_1|^2 P_{1,1} + 1}{|h_2|^2}, \tag{6}$$

$$P_{2,2} = \left(\frac{e^R\left(|h_1|^2 P_{1,1} + 1\right)}{|h_2|^2}\right)^{\frac{1}{2}} - \frac{1}{|h_2|^2}.$$

It is straightforward to show that $P_{2,1} = 0$ since

$$P_{2,1} = \left(\frac{e^R e^R}{|h_2|^2}\right)^{\frac{1}{2}} - \frac{e^R}{|h_2|^2} = 0, \tag{7}$$

i.e., U_2 chooses the OMA mode, where the last step follows from the fact that $P_{1,1}$ needs to satisfy the following equality: $\log\left(|h_1|^2 P_{1,1} + 1\right) = R$. As pointed out in [20]–[22], uplink hybrid NOMA can achieve a significant performance gain over OMA, if the durations of different time slots are different. This conclusion does not contradict the one made in this paper, since Corollary 1 is obtained by assuming that the durations of all time slots are the same.

Lemma 3 shows that the solution provided in Lemma 1 is the optimal solution of problem (P2). By using this conclusion and following steps similar to those in the proof of [20, Lemma 6], the following corollary can be obtained.

Corollary 2. *For the considered special case with ordered channel gains, the downlink hybrid NOMA power allocation solution obtained in Lemma 1 is a Pareto optimal solution of the multi-objective optimization problem shown in* (P1).

III. Downlink MISO Hybrid NOMA Transmission

In this section, the downlink SISO hybrid NOMA scheme developed in the previous section is extended to the MISO scenario. Unlike the SISO network, the legacy MISO network is based on both TDMA and SDMA. In particular, it is assumed that there exist two groups of users, denoted by $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, where the users in each group are served simultaneously via SDMA, and TDMA is used to avoid inter-group interference, i.e., the users in $\mathcal{G}_1$ are scheduled to be served earlier than those in $\mathcal{G}_2$. Similar to the SISO case, a user in $\mathcal{G}_2$ can have access to the time slots which belong to $\mathcal{G}_1$. Unlike in the SISO case, the base station also needs to decide whether to design new beamforming vectors for the users in $\mathcal{G}_2$ during the time slots which belong to the users in $\mathcal{G}_1$ [11]. In order to avoid any disruptions to the legacy network, beam sharing is used, i.e., the spatial beams preconfigured for the users in $\mathcal{G}_1$ are used to serve the users in $\mathcal{G}_2$.

The importance of near-field communications in future wireless networks motivates the use of the near-field channel model as an illustrative example. We note that for conventional far-field beamforming, the concept of beam-sharing is straightforward, since many users can share the same beam-steering vector and hence can be served by a single far-field beam [1]. However, the accurate beamfocusing in near-field communications can make beam sharing difficult, which is another motivation for using near-field beamforming for the feasibility study of hybrid NOMA in downlink MISO systems.

A. Near-Field Communication System Model

Consider a near-field MISO downlink network, where a base station is equipped with an N-antenna uniform linear array (ULA). There are M and K single-antenna users in $\mathcal{G}_1$ and and $\mathcal{G}_2$, respectively, which are denoted by $\mathrm{U}_m^{\mathrm{G1}}$ and $\mathrm{U}_k^{\mathrm{G2}}$, respectively. The scenario studied in [29] is a special case with $K = 1$. Similar to [29], the ULA is assumed to be placed

at the center of a 2-dimensional plane. By using Cartesian coordinates, the locations of $\mathrm{U}_m^{\mathrm{G1}}$, $\mathrm{U}_k^{\mathrm{G2}}$, the center of the ULA, and the n-th element of the ULA are denoted by $\boldsymbol{\psi}_m^{\mathrm{G1}}$, $\boldsymbol{\psi}_k^{\mathrm{G2}}$, $\boldsymbol{\psi}_0$, and $\boldsymbol{\psi}_n$, respectively.

It is assumed that $\mathrm{U}_k^{\mathrm{G1}}$'s distance to the base station is much smaller than the Rayleigh distance. Therefore, the resolution of the beamformers of the users in $\mathcal{G}_1$ is almost perfect, i.e., the users' channel vectors are almost orthogonal to each other, which makes the implementation of beamfocusing possible for the users in $\mathcal{G}_1$ [30], [31]. On the other hand, the $\mathrm{U}_k^{\mathrm{G2}}$ could be far-field users, or near-field users whose distances to the base station are larger than those of the users in $\mathcal{G}_1$.

B. An MSIO OMA Benchmark

It is assumed that $\mathrm{U}_k^{\mathrm{G1}}$ and $\mathrm{U}_k^{\mathrm{G2}}$ cannot be simultaneously served by SDMA, which motivates the considered OMA benchmarking scheme based on the combination of SDMA and TDMA[3]. In particular, the OMA transmission consists of two phases. During the first phase which consists of M time slots, the base station serves the M users in $\mathcal{G}_1$ simultaneously via SDMA, and the observation at $\mathrm{U}_m^{\mathrm{G1}}$ is given by

$$y_m^{\mathrm{G1}} = \mathbf{h}_m^H \sum_{m=1}^{M} \sqrt{P^{\mathrm{G1}}} \mathbf{w}_m^{\mathrm{G1}} s_m^{\mathrm{G1}} + n_m^{\mathrm{G1}}, \tag{8}$$

where n_m^{G1} denotes the additive white Gaussian noise with normalized power, s_m^{G1} denotes the symbol intended for $\mathrm{U}_m^{\mathrm{G1}}$, each users in $\mathcal{G}_1$ is assumed to have the same transmit power, denoted by P^{G1}, $\mathbf{w}_m^{\mathrm{G1}}$ denotes the user's beamforming vector, the spherical channel model is adopted, i.e., $\mathbf{h}_m = \sqrt{N}\alpha_m^{\mathrm{G1}}\mathbf{b}\left(\boldsymbol{\psi}_m^{\mathrm{G1}}\right)$, $\mathbf{b}\left(\boldsymbol{\psi}\right) = \frac{1}{\sqrt{N}}\left[e^{-j\frac{2\pi}{\lambda_w}|\boldsymbol{\psi}-\boldsymbol{\psi}_1|} \quad \ldots \quad e^{-j\frac{2\pi}{\lambda_w}|\boldsymbol{\psi}-\boldsymbol{\psi}_N|}\right]^T$, λ_w denotes the wavelength, and $\alpha_m^{\mathrm{G1}} = \frac{\lambda_w}{4\pi|\boldsymbol{\psi}_m^{\mathrm{G1}}-\boldsymbol{\psi}_0|}$ [32]–[35]. Similar to the previous section, the base station is assumed to have access to the users' CSI. Therefore, $\mathrm{U}_m^{\mathrm{G1}}$'s data rate in OMA can be expressed as:

$$\bar{R}_m^{\mathrm{G1}} = \log\left(1 + \frac{P^{\mathrm{G1}}|\mathbf{h}_m^H\mathbf{w}_m^{\mathrm{G1}}|^2}{P^{\mathrm{G1}}\sum_{i=1,i\neq m}^{M}|\mathbf{h}_m^H\mathbf{w}_i^{\mathrm{G1}}|^2 + 1}\right). \tag{9}$$

The second phase of OMA transmission consists of K time slots, where the users in $\mathcal{G}_2$, $\mathrm{U}_k^{\mathrm{G2}}$, are served simultaneously also via SDMA, and their data rates, denoted by $\bar{R}_k^{\mathrm{G2}}$, are similar to those of the users in $\mathcal{G}_1$, i.e.,

$$\bar{R}_k^{\mathrm{G2}} = \log\left(1 + \frac{P_{k,2}^{\mathrm{G2}}|\mathbf{g}_k^H\mathbf{w}_k^{\mathrm{G2}}|^2}{\sum_{i=1,i\neq k}^{K}P_{i,2}^{\mathrm{G2}}|\mathbf{g}_k^H\mathbf{w}_i^{\mathrm{G2}}|^2 + 1}\right), \tag{10}$$

where $\mathbf{g}_k$ and $\mathbf{w}_k^{\mathrm{G2}}$ are defined similar to their counterparts for the users in $\mathcal{G}_1$, and $P_{k,2}^{\mathrm{G2}}$ denotes the transmit power of $\mathrm{U}_k^{\mathrm{G2}}$ during the second phase. Because the use of downlink hybrid NOMA does not affect the users in $\mathcal{G}_1$, we will focus on optimizing the parameters for the users in $\mathcal{G}_2$, i.e., P^{G1} is fixed and $P_{k,2}^{\mathrm{G2}}$ is to be optimized.

[3]This assumption can be justified if $\mathrm{U}_m^{\mathrm{G1}}$ and $\mathrm{U}_k^{\mathrm{G2}}$ are near-field and far-field users, respectively, where there will be strong co-channel interference if the two types of users are served simultaneously. For the case that all users are near-field users, the assumption can be still justified, if the resolution between the channels of $\mathrm{U}_m^{\mathrm{G1}}$ and $\mathrm{U}_k^{\mathrm{G2}}$ is not perfect, i.e., their channels are not perfectly orthogonal to each other [30]–[32].

The assumption that the users in $\mathcal{G}_1$ are very close to the base station makes beamfocusing possible to these users. In particular, the direction vector of a user's spherical channel vector can be used as the user's beamforming vector, i.e., $\mathbf{w}_m^{\mathrm{G1}} = \mathbf{b}\left(\boldsymbol{\psi}_m^{\mathrm{G1}}\right)$, and $|\mathbf{h}_i^H\mathbf{w}_m^{\mathrm{G1}}|^2 \approx 0$, for $i \neq m$, since the resolution for the near-field users very close to the base station is almost perfect, i.e., $|\mathbf{b}\left(\boldsymbol{\psi}_i^{\mathrm{G1}}\right)^H\mathbf{b}\left(\boldsymbol{\psi}_m^{\mathrm{G1}}\right)|^2 \approx 0$ [31]. Alternatively, zero-forcing beamforming vector can also be used by the users in $\mathcal{G}_1$, which can ensure that $|\mathbf{h}_i^H\mathbf{w}_m^{\mathrm{G1}}|^2$, $i \neq m$, is strictly zero, i.e., there is no inter-beam interference. For the users in $\mathcal{G}_2$, conventional zero-forcing beamforming is used, since beamfocusing cannot be applied due to their large distances to the base station.

C. Downlink MISO Hybrid NOMA Transmission

By using downlink hybrid NOMA, the $\mathrm{U}_k^{\mathrm{G2}}$ can have access to the time slots in the first phase, i.e., the base station broadcasts the following signals during the first M time slots:

$$\mathbf{x}^{\mathrm{NOMA}} = \sum_{m=1}^{M}\mathbf{w}_m^{\mathrm{G1}}\left(\sqrt{P^{\mathrm{G1}}}s_m^{\mathrm{G1}} + \sum_{k=1}^{K}s_{m,k}\sqrt{\tilde{P}_{m,k}^{\mathrm{G2}}}s_{m,k}^{\mathrm{G2}}\right), \tag{11}$$

where $s_{m,k}^{\mathrm{G2}}$ denotes the signal sent to $\mathrm{U}_k^{\mathrm{G2}}$ on $\mathbf{w}_m^{\mathrm{G1}}$, $s_{m,k}$ is an indicator, i.e., $s_{m,k} = 1$ if $\mathrm{U}_k^{\mathrm{G2}}$ uses $\mathbf{w}_m^{\mathrm{G1}}$, otherwise $s_{m,k} = 0$, and $\tilde{P}_{m,k}^{\mathrm{G2}}$ is a power allocation coefficient to be optimized.

For illustration purposes, it is assumed that $M > K$, and each user in $\mathcal{G}_2$ selects a single beam. Denote the index of the beam used by $\mathrm{U}_k^{\mathrm{G2}}$ by i^k, e.g., $i^k = 2$ means that on beam $\mathbf{w}_2^{\mathrm{G1}}$, $\mathrm{U}_k^{\mathrm{G2}}$ is active. Recall that during the first phase, $\mathrm{U}_{i^k}^{\mathrm{G1}}$ is scheduled to be served exclusively on $\mathbf{w}_{i^k}^{\mathrm{G1}}$ in TDMA. Therefore, similar to the previous section, $\mathrm{U}_{i^k}^{\mathrm{G1}}$ carries out SIC by first decoding the signal for $\mathrm{U}_k^{\mathrm{G2}}$ with the following data rate:

$$R_{i^k\to k}^{\mathrm{G1}} = \log\left(1 + \frac{|\mathbf{h}_{i^k}^H\mathbf{w}_{i^k}^{\mathrm{G1}}|^2P_{k,1}^{\mathrm{G2}}}{\mathrm{I}_{i^k}^{\mathrm{G1}} + 1}\right), \tag{12}$$

where $P_{k,1}^{\mathrm{G2}} = \tilde{P}_{i^k,k}^{\mathrm{G2}}$, and

$$\mathrm{I}_{i^k}^{\mathrm{G1}} = P^{\mathrm{G1}}\sum_{m=1}^{M}|\mathbf{h}_{i^k}^H\mathbf{w}_m^{\mathrm{G1}}|^2 + \sum_{j\neq k}|\mathbf{h}_{i^k}^H\mathbf{w}_{i^j}^{\mathrm{G1}}|^2P_{j,1}^{\mathrm{G2}}. \tag{13}$$

We note that in the considered MISO context, this SIC decoding order can also be justified by the fact that the effective channel gain of $\mathrm{U}_{i^k}^{\mathrm{G1}}$, $|\mathbf{h}_{i^k}^H\mathbf{w}_{i^k}^{\mathrm{G1}}|^2$, is strong since $\mathbf{w}_{i^k}^{\mathrm{G1}}$ is tailored to the channel vector of $\mathrm{U}_{i^k}^{\mathrm{G1}}$.

Assume that all the users in $\mathcal{G}_2$ have the same target data rate, denoted by R. We note that if $R_{i^k\to k}^{\mathrm{G1}} \geq R$, $\mathrm{U}_{i^k}^{\mathrm{G1}}$ can successfully remove its partner's signal and decode its own signal in the same manner as in OMA, i.e., the data rates for the users in $\mathcal{G}_1$ in NOMA and OMA are the same.

On the other hand, during the first phase, $\mathrm{U}_k^{\mathrm{G2}}$ decodes its signal directly with the following data rate:

$$R_k^{\mathrm{G2}} = \log\left(1 + \frac{|\mathbf{g}_k^H\mathbf{w}_{i^k}^{\mathrm{G1}}|^2P_{k,1}^{\mathrm{G2}}}{\mathrm{I}_k^{\mathrm{G2}} + 1}\right), \tag{14}$$

where

$$\mathrm{I}_k^{\mathrm{G2}} = P^{\mathrm{G1}} \sum_{m=1}^{M} \left|\mathbf{g}_k^H \mathbf{w}_m^{\mathrm{G1}}\right|^2 + \sum_{j\neq k} |\mathbf{g}_k^H \mathbf{w}_{i^j}^{\mathrm{G1}}|^2 P_{j,1}^{\mathrm{G2}}. \quad (15)$$

Therefore, in downlink MISO hybrid NOMA, the achievable data rate for $\mathrm{U}_k^{\mathrm{G2}}$ is given by

$$R_{k,1}^{\mathrm{G2}} = \min\left\{R_k^{\mathrm{G2}}, R_{i^k\to k}^{\mathrm{G1}}\right\}. \quad (16)$$

During the second phase which consists of K time slots, SDMA can be employed again to support the K users in $\mathcal{G}_2$, which means that, during the second phase, the data rate of $\mathrm{U}_k^{\mathrm{G2}}$ in downlink hybrid NOMA, denoted by $R_{k,2}^{\mathrm{G2}}$, is the same as that for OMA, i.e., $R_{k,2}^{\mathrm{G2}} = \bar{R}_k^{G2}$ shown in (10).

Recall that during the first phase, the users in $\mathcal{G}_2$ can use the beams preconfigured to the users in $\mathcal{G}_1$. Hence, beam selection, i.e., how to choose beam $\mathbf{w}_{i^k}^{\mathrm{G1}}$ for $\mathrm{U}_k^{\mathrm{G2}}$, is crucial for the performance of hybrid NOMA. If beamfocusing is used for beamforming, a user's beamforming vector is aligned to its channel vector. According to [30], [31], the resolution of near-field beamforming can be almost perfect in the angle domain, but not necessarily in the distance domain. Given the fact that $\mathrm{U}_m^{\mathrm{G1}}$ is close to the base station and $\mathrm{U}_k^{\mathrm{G2}}$ is far away, $\mathrm{U}_k^{\mathrm{G2}}$ can select the legacy beam whose angle of departure is closest to its own one, i.e., $i^k = \arg\min_m |\theta_k^{\mathrm{G2}} - \theta_m^{\mathrm{G1}}|$, where $\left(\theta_k^{\mathrm{G2}}, r_k^{\mathrm{G2}}\right)$ denotes the polar coordinates of $\mathrm{U}_k^{\mathrm{G2}}$, and $\left(\theta_m^{\mathrm{G1}}, r_m^{\mathrm{G1}}\right)$ denotes the polar coordinates of $\mathrm{U}_m^{\mathrm{G1}}$. If zero-forcing based beamforming is used, $\mathrm{U}_k^{\mathrm{G2}}$ can select the beam on which the user's effective channel gain is the largest, i.e., $i^k = \arg\max_m |\mathbf{g}_k^H \mathbf{w}_m^{\mathrm{G1}}|^2$.

D. Downlink MISO Hybrid NOMA Power Allocation

Similar to the previous section, an energy minimization problem can be formulated as follows:

$$\min_{P_{k,i}^{\mathrm{G2}}\geq 0} \quad \sum_{k=1}^{K} \left(MTP_{k,1}^{\mathrm{G2}} + KTP_{k,2}^{\mathrm{G2}}\right) \quad (\mathrm{P4a})$$

$$s.t. \quad MTR_{k,1}^{\mathrm{G2}} + KTR_{k,2}^{\mathrm{G2}} \geq TKR, 1\leq k\leq K, \quad (\mathrm{P4b})$$

where K is used on the right-hand side of (P4b) to highlight that K time slots are used during the second phase. Similarly to the SISO scenario, the perfect knowledge of the users' CSI is assumed to be available at the base station. In order to gain a better understanding to the properties of problem (P4), the users' data rate expressions need to be simplified as follows.

First, R_k^{G2} can be written as the following explicit function of the power allocation coefficients:

$$R_k^{\mathrm{G2}} = \log\left(1 + \frac{g_{k,k}P_{k,1}^{\mathrm{G2}}}{\sum_{j\neq k} g_{k,j}P_{j,1}^{\mathrm{G2}} + b_k}\right), \quad (17)$$

where $g_{k,j} = \left|\mathbf{g}_k^H \mathbf{w}_{i^j}^{\mathrm{G1}}\right|^2$, and $b_k = P^{\mathrm{G1}} \sum_{m=1}^{M} \left|\mathbf{g}_k^H \mathbf{w}_m^{\mathrm{G1}}\right|^2 + 1$. Similarly, both $R_{i^k\to k}^{\mathrm{G1}}$ and $R_{k,2}^{\mathrm{G2}}$ can be expressed as the following explicit functions of $P_{k,1}^{\mathrm{G2}}$ and $P_{k,2}^{\mathrm{G2}}$:

$$R_{i^k\to k}^{\mathrm{G1}} = \log\left(1 + \frac{h_{k,k}P_{k,1}^{\mathrm{G2}}}{\sum_{j\neq k} h_{k,j}P_{j,1}^{\mathrm{G2}} + d_k}\right), \quad (18)$$

$$R_{k,2}^{\mathrm{G2}} = \log\left(1 + \frac{c_{k,k}P_{k,2}^{\mathrm{G2}}}{\sum_{i=1,i\neq k}^{K} c_{k,i}P_{i,2}^{\mathrm{G2}} + 1}\right),$$

where $h_{k,j} = |\mathbf{h}_{i^k}^H \mathbf{w}_{i^j}^{\mathrm{G1}}|^2$, $d_k = P^{\mathrm{G1}} \sum_{m=1}^{M} |\mathbf{h}_{i^k}^H \mathbf{w}_m^{\mathrm{G1}}|^2 + 1$, and $c_{k,i} = \left|\mathbf{g}_k^H \mathbf{w}_i^{\mathrm{G2}}\right|^2$.

1) OMA Power Allocation: By assuming that $\mathrm{U}_k^{\mathrm{G2}}$ does not use the time slots in the first phase, i.e., $P_{k,1}^{\mathrm{G2}} = 0$, the considered hybrid NOMA optimization problem is degraded to a simple OMA case, as follows:

$$\min_{P_{k,2}^{\mathrm{G2}}\geq 0} \quad \sum_{k=1}^{K} P_{k,2}^{\mathrm{G2}} \quad (\mathrm{P5a})$$

$$s.t. \quad KR_{k,2}^{\mathrm{G2}} \geq KR, 1\leq k\leq K, \quad (\mathrm{P5b})$$

which can be recast by using the simplified expressions of $R_{k,2}^{\mathrm{G2}}$ as follows:

$$\min_{P_{k,2}^{\mathrm{G2}}\geq 0} \quad \sum_{k=1}^{K} P_{k,2}^{\mathrm{G2}} \quad (\mathrm{P6a})$$

$$s.t. \quad \frac{c_{k,k}P_{k,2}^{\mathrm{G2}}}{\sum_{i=1,i\neq k}^{K} c_{k,i}P_{i,2}^{\mathrm{G2}} + 1} \geq e^R - 1, 1\leq k\leq K. \quad (\mathrm{P6b})$$

It is straightforward to show that problem (P6) is a convex optimization problem, and hence can be solved efficiently by applying off-shelf optimization solvers.

2) Downlink Hybrid NOMA Power Allocation: While the power allocation problem can be easily solved for the OMA case, it is more challenging to solve the general downlink hybrid NOMA problem which can be rewritten as follows:

$$\min_{P_{k,i}^{\mathrm{G2}}\geq 0} \quad \sum_{k=1}^{K} \left(MP_{k,1}^{\mathrm{G2}} + KP_{k,2}^{\mathrm{G2}}\right) \quad (\mathrm{P7a})$$

$$s.t. \quad \frac{M}{K}\log\left(1 + \frac{h_{k,k}P_{k,1}^{\mathrm{G2}}}{\sum_{j\neq k} h_{k,j}P_{j,1}^{\mathrm{G2}} + d_k}\right) + \quad (\mathrm{P7b})$$

$$\log\left(1 + \frac{c_{k,k}P_{k,2}^{\mathrm{G2}}}{\sum_{i=1,i\neq k}^{K} c_{k,i}P_{i,2}^{\mathrm{G2}} + 1}\right) \geq R, 1\leq k\leq K$$

$$\frac{M}{K}\log\left(1 + \frac{g_{k,k}P_{k,1}^{\mathrm{G2}}}{\sum_{j\neq k} g_{k,j}P_{j,1}^{\mathrm{G2}} + b_k}\right) + \quad (\mathrm{P7c})$$

$$\log\left(1 + \frac{c_{k,k}P_{k,2}^{\mathrm{G2}}}{\sum_{i=1,i\neq k}^{K} c_{k,i}P_{i,2}^{\mathrm{G2}} + 1}\right) \geq R, 1\leq k\leq K.$$

Unlike the OMA problem in (P6), problem (P7) is non-convex, mainly due to the fact that the optimization variables appear in both the numerators and the denominators of the fractions in (P7b) and (P7c).

In the literature, SCA has been shown to be effective for tackling the non-convex constraints shown in (P7b) and (P7c) [36]. To facilitate the implementation of SCA, the following vectors are defined: $\mathbf{p} = \left[P_{1,1}^{\mathrm{G2}} \quad \cdots \quad P_{K,1}^{\mathrm{G2}}\right]^T$ and $\mathbf{e} = \left[P_{1,2}^{\mathrm{G2}} \quad \cdots \quad P_{K,2}^{\mathrm{G2}}\right]^T$, which means that constraint (P7c) can be rewritten as follows:

$$M\log\left(\bar{\mathbf{g}}_k^T\mathbf{p} + b_k\right) - M\log\left(\tilde{\mathbf{g}}_k^T\mathbf{p} + b_k\right) + K\log\left(\mathbf{c}_k^T\mathbf{e} + 1\right) - K\log\left(\tilde{\mathbf{c}}_k^T\mathbf{e} + 1\right) \geq KR, \quad (19)$$

where $\bar{\mathbf{g}}_k = \left[g_{k,1} \quad \cdots \quad g_{k,K}\right]^T$, the k-th element of $\tilde{\mathbf{g}}_k$ is zero, the other elements of $\tilde{\mathbf{g}}_k$ are the same as those of $\bar{\mathbf{g}}_k$,

$\mathbf{c}_k = \begin{bmatrix} c_{k,1} & \cdots & c_{k,K} \end{bmatrix}^T$, and $\tilde{\mathbf{c}}_k$ is constructed in a similar manner as $\tilde{\mathbf{g}}_k$ by using the elements of $\mathbf{c}_k$.

Therefore, at the i-th iteration of SCA, constraint (P7c) can be approximated as follows:

$$\frac{M}{K}\log\left(\bar{\mathbf{g}}_k^T\mathbf{p}+b_k\right)-\frac{M}{K}\log\left(\tilde{\mathbf{g}}_k^T\mathbf{p}_{i-1}+b_k\right)+\frac{M}{K}\frac{\tilde{\mathbf{g}}_k^T(\mathbf{p}-\mathbf{p}_{i-1})}{\tilde{\mathbf{g}}_k^T\mathbf{p}+b_k}$$
$$+\log\left(\mathbf{c}_k^T\mathbf{e}+1\right)-\log\left(\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1\right)+\frac{\tilde{\mathbf{c}}_k^T(\mathbf{e}-\mathbf{e}_{i-1})}{\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1}\geq R, \tag{20}$$

where $\mathbf{p}_{i-1}$ and $\mathbf{e}_{i-1}$ are obtained from the $(i-1)$-th iteration. Similarly, constraint (P7b) can be first recast as follows:

$$M\log\left(\bar{\mathbf{h}}_k^T\mathbf{p}+d_k\right)-M\log\left(\tilde{\mathbf{h}}_k^T\mathbf{p}+d_k\right)$$
$$+K\log\left(\mathbf{c}_k^T\mathbf{e}+1\right)-K\log\left(\tilde{\mathbf{c}}_k^T\mathbf{e}+1\right)\geq KR, \tag{21}$$

where $\bar{\mathbf{h}}_k = \begin{bmatrix} h_{k,1} & \cdots & h_{k,K} \end{bmatrix}^T$, and $\tilde{\mathbf{h}}_k$ is the same as $\bar{\mathbf{h}}_k$ except its k-th element being zero. Again, at the i-th iteration of SCA, constraint (P7b) can be approximated as follows:

$$\frac{M}{K}\log\left(\bar{\mathbf{h}}_k^T\mathbf{p}+d_k\right)-\frac{M}{K}\log\left(\tilde{\mathbf{h}}_k^T\mathbf{p}_{i-1}+d_k\right)+\frac{M}{K}\frac{\tilde{\mathbf{h}}_k^T(\mathbf{p}-\mathbf{p}_{i-1})}{\tilde{\mathbf{h}}_k^T\mathbf{p}+d_k}$$
$$+\log\left(\mathbf{c}_k^T\mathbf{e}+1\right)-\log\left(\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1\right)+\frac{\tilde{\mathbf{c}}_k^T(\mathbf{e}-\mathbf{e}_{i-1})}{\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1}\geq R. \tag{22}$$

SCA can be carried out iteratively to obtain a suboptimal solution of problem (P7), where the i-th stage of the SCA algorithm requires solving the following convex optimization problem:

$$\min_{\mathbf{p},\mathbf{e}\geq 0} \quad M\mathbf{1}_K^T\mathbf{p}+K\mathbf{1}_K^T\mathbf{e} \tag{P8a}$$

$$s.t. \quad \frac{M}{K}\log\left(\bar{\mathbf{h}}_k^T\mathbf{p}+d_k\right)-\frac{M}{K}\log\left(\tilde{\mathbf{h}}_k^T\mathbf{p}_{i-1}+d_k\right) \tag{P8b}$$
$$+\frac{M}{K}\frac{\tilde{\mathbf{h}}_k^T(\mathbf{p}-\mathbf{p}_{i-1})}{\tilde{\mathbf{h}}_k^T\mathbf{p}+d_k}+\log\left(\mathbf{c}_k^T\mathbf{e}+1\right)$$
$$-\log\left(\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1\right)+\frac{\tilde{\mathbf{c}}_k^T(\mathbf{e}-\mathbf{e}_{i-1})}{\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1}\geq R, 1\leq k\leq K,$$
$$\frac{M}{K}\log\left(\bar{\mathbf{g}}_k^T\mathbf{p}+b_k\right)-\frac{M}{K}\log\left(\tilde{\mathbf{g}}_k^T\mathbf{p}_{i-1}+b_k\right) \tag{P8c}$$
$$+\frac{M}{K}\frac{\tilde{\mathbf{g}}_k^T(\mathbf{p}-\mathbf{p}_{i-1})}{\tilde{\mathbf{g}}_k^T\mathbf{p}+b_k}+\log\left(\mathbf{c}_k^T\mathbf{e}+1\right)$$
$$-\log\left(\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1\right)+\frac{\tilde{\mathbf{c}}_k^T(\mathbf{e}-\mathbf{e}_{i-1})}{\tilde{\mathbf{c}}_k^T\mathbf{e}_{i-1}+1}\geq R, 1\leq k\leq K,$$

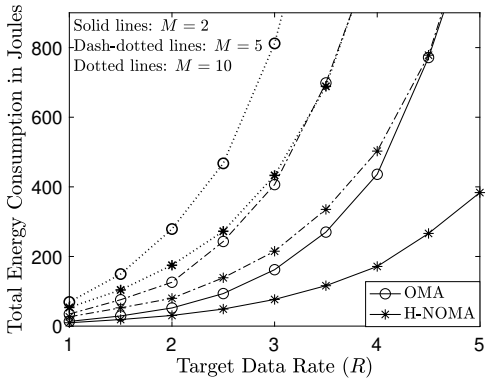
where $\mathbf{1}_K$ is a $K\times 1$ all-one vector.

IV. Simulation Results

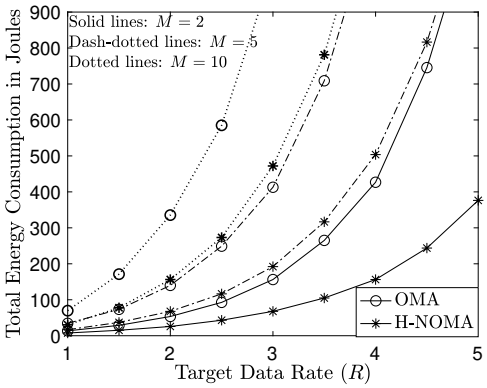
The performance of downlink hybrid NOMA transmission for SISO and MISO systems is evaluated in the following two subsections, respectively.

A. Downlink SISO Hybrid NOMA Transmission

In Fig. 2, the total energy consumption realized by the proposed downlink hybrid NOMA scheme is shown as a function of R, where the performance of OMA is also shown as a benchmark. For Fig. 2, the users' channels are assumed to be independent and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance[4]. Fig. 2(a) focused on the special case that the users' channel gains are ordered, and Fig. 2(b) focuses on the general case that the users' channel gains are not ordered. For the case with the ordered channel gains, the simulation results are obtained by applying optimization solvers to solve problem (P2), and the analytical results are obtained by applying the hybrid NOMA solution provided in Lemma 1. For the general case with unordered channel gains, only simulation results are presented. The two figures in Fig. 2 show that the use of downlink hybrid NOMA can reduce the total energy consumption significantly, compared to OMA, particularly for large R. Fig. 2(a) also shows that the analytical results perfectly match the simulation results, which verifies Corollary 1, i.e., for the special case with ordered channel gains, hybrid NOMA power allocation is the optimal solution of problem (P2) and outperforms OMA.



(a) Special case with ordered channel gains

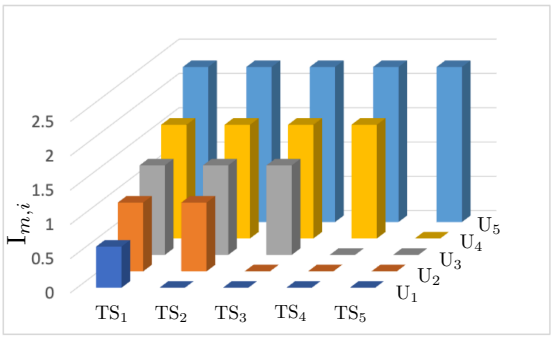


(b) General case with unordered channel gains

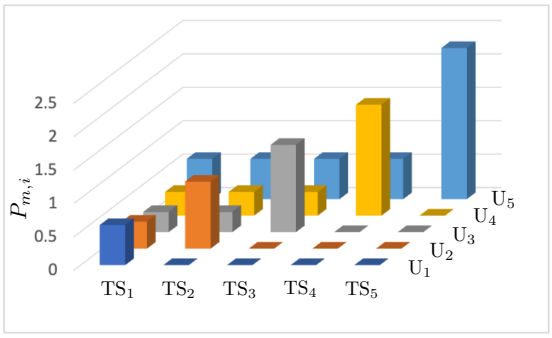
Fig. 2. The total energy consumption realized by the considered transmission schemes in the SISO case.

In Figs. 3 and 4, the properties of hybrid NOMA power allocation are studied by focusing on a deterministic case with $h_m = m$, $1\leq m\leq M$. In particular, Fig. 3(a) shows that the accumulated interference from the first m users at different time slots is identical, i.e., $\sum_{j=1}^{m} P_{j,1} =$

[4]With complex-Gaussian fading, the average energy consumption can be infinite, as explained in the following. Take OMA as an example, where the average energy consumption of $\mathrm{U}_m^{\mathrm{G1}}$ is $\mathcal{E}\left\{\frac{e^R-1}{|h_m|^2}\right\}$, which is infinite since $\mathcal{E}\left\{\frac{1}{x}\right\}\to\infty$ for exponentially distributed x, where $\mathcal{E}\{\cdot\}$ denotes the expectation. Therefore, in the simulation, a constraint of $|h_m|^2\geq 0.01$ is imposed to avoid this singularity issue.



(a) Accumulated interference from the first m users in T_i - $\mathrm{I}_{m,i} = \sum_{j=i}^{m} P_{j,i}$



(b) U_m's power allocation in T_i - $P_{m,i}$

Fig. 3. An illustration of the properties of hybrid NOMA power allocation. $h_m = m$, $1 \leq m \leq M$, and $R = 2$ nats per channel use (NPCU).

$\cdots = \sum_{j=m-1}^{m} P_{j,m-1} = P_{m,m}$, which confirms Lemma 2. Fig. 3(b) shows the users' power allocation coefficients in different time slots, and demonstrates that U_m will use the same transmit power during the first $(m-1)$ time slots, i.e, the NOMA time slots. As discussed in the proof of Lemma 3, the observation from Fig. 3(a) is the reason for the observation from Fig. 3(b). Take U_5 as an example. Fig. 3(a) shows that the interferences experienced by U_5 in the first 4 time slots, i.e., the yellow bars, are same. Furthermore, we note that U_5's channel gains in these time slots are assumed to be the same, which means that U_5 naturally uses the same transmit powers during the first 4 time slots. In Fig. 4, the hybrid NOMA power allocation solution in Lemma 1 is shown to perfectly match the solution obtained by an exhaustive search, which verifies the closed-form expression of the optimal power allocation solution shown in Lemma 1.

B. Downlink MISO Hybrid NOMA Transmission

In this subsection, the performance of downlink MISO hybrid NOMA transmission is focused on, where the carrier frequency is set to 28 GHz, and the antenna spacing is half of the wavelength. In Fig. 5, the total energy consumption realized by the considered transmission schemes is illustrated, by assuming that the users in $\mathcal{G}_1$ are randomly located in the half-ring with radius 10 m and 50 m. The users in $\mathcal{G}_2$, $\mathrm{U}_k^{\mathrm{G2}}$, are at fixed locations 200 m away from the base station, and their angles are equally spaced between $-\frac{\pi}{3}$ and $\frac{\pi}{3}$. The two sub-figures of Fig. 5 demonstrate that the use of downlink hybrid NOMA can realize a significant performance gain over

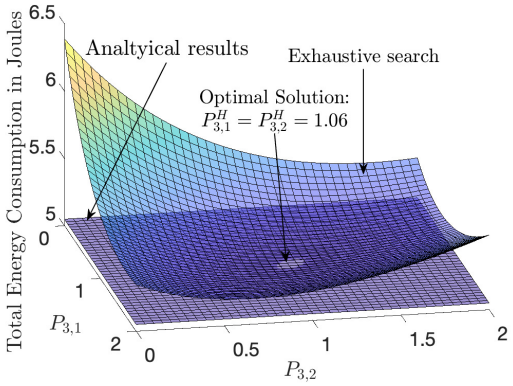


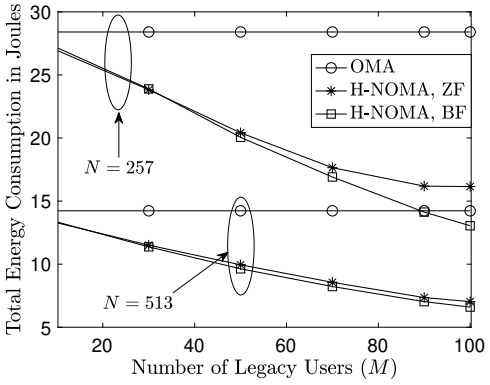
Fig. 4. Verification for the optimality of the hybrid NOMA solution shown in Lemma 1, where $h_m = m$, $1 \leq m \leq M$, and $R = 2$ NPCU.

OMA, particularly if the size of $\mathcal{G}_1$ and the target data rate are large. Fig. 5(a) shows that the use of more antennas can reduce the energy consumption for both OMA and hybrid NOMA, where the performance gap between the two schemes is increased if there are more users in $\mathcal{G}_1$. Fig. 5(b) shows that when the target data rate is small, the performance gain of downlink hybrid NOMA over OMA is small. This is due to the fact that for small target data rates, the time slots available in the second phase are sufficient to serve $\mathrm{U}_k^{\mathrm{G2}}$, and there is no need to employ NOMA and have access to the time slots in the first phase. In addition, Fig. 5 shows that beamfocusing outperforms zero-forcing based beamforming, which is consistent with the conclusion made in [29].

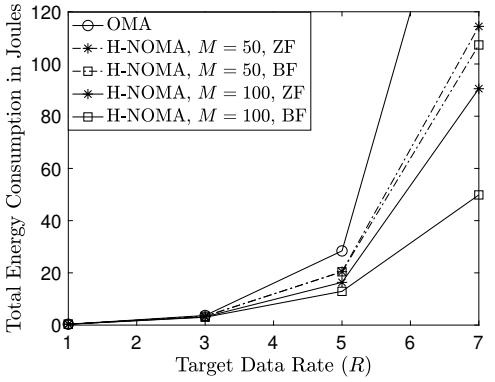
In Fig. 6 and Table I, a deterministic scenario is considered in order to reveal a few interesting properties of downlink hybrid NOMA for near-field communications. In particular, the users in $\mathcal{G}_1$, $\mathrm{U}_m^{\mathrm{G1}}$, are at fixed locations r^{G1} m away from the base station, and their angles are equally spaced by $\frac{\pi}{2M}$. The users in $\mathcal{G}_2$, $\mathrm{U}_k^{\mathrm{G2}}$, are 200 m away from the base station, where $\theta_k^{\mathrm{G1}} = \theta_k^{\mathrm{G2}}$. Fig. 6 is based on the use of beamfocusing, and demonstrates that the use of hybrid NOMA can reduce the energy consumption of OMA significantly, which is consistent to the observations made in Fig. 5. Comparing the two sub-figures of Fig. 6, an interesting observation is that for the case of $r^{\mathrm{G1}} = 50$ m, the use of more antennas at the base station degrades the performance of downlink hybrid NOMA. This is due to the fact that the users in $\mathcal{G}_1$ are very close to the base station, and hence for large numbers of antennas, the resolution of near-field beamforming is almost perfect, i.e., the orthogonality among the users' channel vectors is almost perfect, and one user's beamfocusing based beam covers a small area centered at the user. As a result, it is challenging for a user in $\mathcal{G}_2$ to find a suitable legacy beam, which leads to the performance loss of downlink hybrid NOMA. However, when r^{G1} is increased from 50 m to 100 m, the resolution of near-field beamforming becomes poor, which introduces the opportunity for beam sharing among the users. In Table I, it is shown that for $r^{\mathrm{G1}} = 50$ m, beamfocusing and zero-forcing beamforming achieve practically the same performance, which is again due to the fact that the users in $\mathcal{G}_1$ are very close to the base station, and hence their channel vectors are almost orthogonal. An interesting observation from

TABLE I
TOTAL ENERGY CONSUMPTION FOR THE CONSIDERED DETERMINISTIC SCENARIO OF $r^{\mathrm{G1}} = 50$ M.

R	1	2	3	4	5	6	7
OMA, BF	0.1692	0.6665	2.4081	22.5109	∞	∞	∞
H-NOMA, BF	0.1569	0.3684	0.6103	0.8871	1.2043	1.5678	1.9849
OMA, ZF	0.1678	0.6240	1.864	5.234	14.397	39.3052	107.0106
H-NOMA, ZF	0.1568	0.3675	0.6075	0.8810	1.1926	1.5476	1.9521

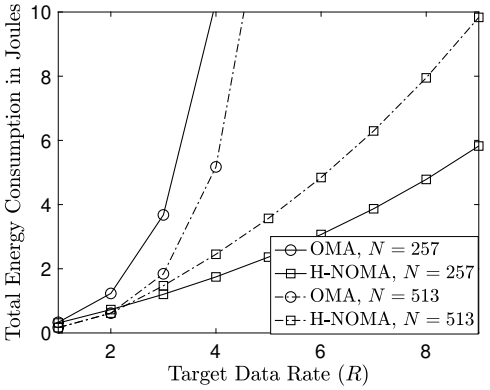


(a) $R = 5$ NPCU

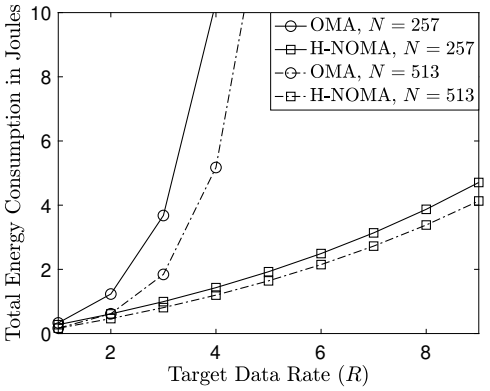


(b) $N = 257$

Fig. 5. The total energy consumption realized by the considered transmission schemes in the MISO case. The users in $\mathcal{G}_1$, $\mathrm{U}_m^{\mathrm{G1}}$, are uniformly located in the half-ring with radius 10 m and 50 m. The users in $\mathcal{G}_2$, $\mathrm{U}_k^{\mathrm{G2}}$, are at fixed locations 200 m away from the base station, and their angles are equally spaced between $-\frac{\pi}{3}$ and $\frac{\pi}{3}$. $K = 3$ and $P^{\mathrm{G1}} = 10$ dBm.



(a) $r^{\mathrm{G1}} = 50$ m



(b) $r^{\mathrm{G1}} = 100$ m

Fig. 6. A deterministic study of the total energy consumption realized by the considered transmission schemes. The users in $\mathcal{G}_1$, $\mathrm{U}_m^{\mathrm{G1}}$, are at fixed locations r^{G1} m away from the base station, and their angles are equally spaced by $\frac{\pi}{2M}$. The users in $\mathcal{G}_2$, $\mathrm{U}_k^{\mathrm{G2}}$, are 200 m away from the base station, where $\theta_k^{\mathrm{G1}} = \theta_k^{\mathrm{G2}}$, $\left(\theta_k^{\mathrm{G2}}, r_k^{\mathrm{G2}}\right)$ denotes the polar coordinates of $\mathrm{U}_m^{\mathrm{G1}}$, and $\left(\theta_m^{\mathrm{G1}}, r_k^{\mathrm{G1}}\right)$ denotes the polar coordinates of $\mathrm{U}_M^{\mathrm{G1}}$. $M = 20$, $K = 3$, and $P^{\mathrm{G1}} = 10$ dBm.

the table is that for OMA, the use of beamfocusing can lead to infinite energy consumption, as explained in the following. Recall that in OMA, $\mathrm{U}_k^{\mathrm{G2}}$ has to rely on the time slots in the second phase, and its achievable data rate is given by $R_{k,2}^{\mathrm{G2}} = \log\left(1 + \frac{c_{k,k}P_{k,2}^{\mathrm{G2}}}{\sum_{i=1,i\neq k}^{K} c_{k,i}P_{i,2}^{\mathrm{G2}}+1}\right)$. With beamfocusing, the inter-beam terms, $c_{k,i}$, are not zero, which means that there exists an upper bound on $R_{k,2}^{\mathrm{G2}}$. When the target data rate is larger than this upper bound, there is no feasible solution for $P_{k,2}^{\mathrm{G2}}$ to realize the large target dada rate. However, by using hybrid NOMA, the user can have access to those time slots in the first phase as well, which avoids the singularity suffered by OMA and reveals an important advantage of downlink hybrid NOMA.

V. CONCLUSIONS

In this paper, hybrid NOMA assisted downlink transmission schemes have been developed for SISO and MISO systems, respectively. For downlink SISO systems, analytical results were derived to reveal several important properties of hybrid NOMA power allocation. For example, in the case that users' channel gains are ordered and the durations of all time slots are the same, downlink hybrid NOMA was shown to always outperform OMA, which is different from the conclusions obtained for uplink hybrid NOMA transmission. For downlink MISO systems, near-field communication was considered to illustrate how NOMA can be used as an add-on in legacy networks based on SDMA and TDMA. Simulation results were presented to verify the developed analytical results and demonstrate the superior performance of downlink hybrid NOMA over conventional OMA. In this paper, MISO hybrid NOMA was implemented between two groups of users, where an important direction for future research is to study how MISO hybrid NOMA can be extended to the case with more than two groups of users. In addition, MISO hybrid NOMA was considered for an ideal near-field system. Thus, an important direction for future research is the consideration of the impact of practical issues, such as non-line-of-sight paths and hybrid beamforming, on the design of downlink hybrid NOMA.

APPENDIX A
PROOF FOR LEMMA 1

It is straightforward to show that U_1's power allocation is the same as that for OMA, i.e., $P_{1,1} = P_1^{\mathrm{OMA}}$. Therefore, the case of $m > 1$ is focused on in this proof.

Recall that $R_{m,i} = \min\left\{R_{m,i}^i, \cdots, R_{m,i}^m\right\}$. We note that $f(x) \triangleq \frac{ax}{bx+1}$ is a monotonically increasing function of x, for positive a, b and x. By using this observation and the fact that $|\bar{h}_{m,i}|^2 = \min\left\{|h_i|^2, \cdots, |h_m|^2\right\}$, the expression for $R_{m,i}$ can be simplified, and problem (P2) can be recast as follows:

$$\min_{P_{m,i}} \quad \sum_{i=1}^{m} P_{m,i} \tag{P9a}$$

$$s.t. \quad \sum_{i=1}^{m} \log\left(1 + b_{m,i}|\bar{h}_{m,i}|^2 P_{m,i}\right) \geq R, \tag{P9b}$$

where the constraints $P_{m,i} \geq 0$ are omitted, and $b_{m,i} = \frac{1}{|\bar{h}_{m,i}|^2 \sum_{j=i}^{m-1} P_{j,i}+1}$.

It is straightforward to show that problem (P9) is a convex optimization problem, and its optimal solution can be found by applying the KKT conditions. In particular, the Lagrangian of problem (P9) is given by

$$L = \sum_{i=1}^{m} P_{m,i} + \lambda\left(R - \sum_{i=1}^{m} \log\left(1 + b_{m,i}|\bar{h}_{m,i}|^2 P_{m,i}\right)\right), \tag{23}$$

where λ is the Lagrangian multiplier. The derivative of the Lagrange is given by

$$\frac{\partial L}{\partial P_{m,i}} = 1 - \lambda \frac{b_{m,i}|\bar{h}_{m,i}|^2}{1 + b_{m,i}|\bar{h}_{m,i}|^2 P_{m,i}} = 0, \tag{24}$$

which means that $P_{i,m}$ can be expressed as follows:

$$P_{m,i} = \lambda - \frac{|\bar{h}_{m,i}|^2 \sum_{j=i}^{m-1} P_{j,i} + 1}{|\bar{h}_{m,i}|^2}. \tag{25}$$

The expression for λ can be found by using the following equality: $\sum_{i=1}^{m} \log\left(1 + b_{m,i}|\bar{h}_{m,i}|^2 P_{m,i}\right) = R$, which yields

$$\lambda = \left(\frac{e^R}{\prod_{i=1}^{m} b_{m,i}|\bar{h}_{m,i}|^2}\right)^{\frac{1}{m}}. \tag{26}$$

By substituting (26) into (25), the optimal power allocation solution is given by

$$P_{m,i} = \left(\frac{e^R}{\prod_{p=1}^{m} b_{m,p}|\bar{h}_{m,i}|^2}\right)^{\frac{1}{m}} - \frac{|\bar{h}_{m,i}|^2 \sum_{j=i}^{m-1} P_{j,i} + 1}{|\bar{h}_{m,i}|^2}. \tag{27}$$

The proof of the lemma is complete.

APPENDIX B
PROOF FOR LEMMA 2

By using the assumption that the users are ordered according to their channel gains, i.e., $|h_m|^2 > |h_{m+1}|^2$, $|\bar{h}_{m,i}|^2$ can be simplified as follows:

$$|\bar{h}_{m,i}|^2 \triangleq \min\left\{|h_i|^2, \cdots, |h_m|^2\right\} = |h_m|^2 \triangleq \gamma_m, \tag{28}$$

which means that the expression of $P_{m,i}^{\mathrm{H}}$ can be simplified as follows:

$$P_{m,i}^{\mathrm{H}} = \left(\frac{e^R}{\prod_{p=1}^{m} \frac{\gamma_m}{\gamma_m \sum_{j=p}^{m-1} P_{j,p}^{\mathrm{H}}+1}}\right)^{\frac{1}{m}} - \frac{\gamma_m \sum_{j=i}^{m-1} P_{j,i}^{\mathrm{H}} + 1}{\gamma_m}. \tag{29}$$

The lemma can be proved by mathematical induction.

A. The Base Case $m = 2$

For the special case of $m = 2$, by using (29), U_2's transmit power during the first two time slots are given by

$$P_{2,1}^{\mathrm{H}} = \left(\frac{e^R}{\frac{\gamma_2^2}{\gamma_2 P_{1,1}^{\mathrm{H}}+1}}\right)^{\frac{1}{2}} - \frac{\gamma_2 P_{1,1}^{\mathrm{H}} + 1}{\gamma_2}, \tag{30}$$

$$P_{2,2}^{\mathrm{H}} = \left(\frac{e^R}{\frac{\gamma_2^2}{\gamma_2 P_{1,1}^{\mathrm{H}}+1}}\right)^{\frac{1}{2}} - \frac{1}{\gamma_2}.$$

In addition, $P_{1,1}^{\mathrm{H}} = \frac{e^R-1}{\gamma_1}$ in order to ensure $\log\left(1 + \gamma_1 P_{1,1}^{\mathrm{H}}\right) = R$.

Therefore, $P_{2,1}^{\mathrm{H}} + P_{1,1}^{\mathrm{H}}$ is given by

$$P_{2,1}^{\mathrm{H}} + P_{1,1}^{\mathrm{H}} = \left(\frac{e^R}{\frac{\gamma_2^2}{\gamma_2 P_{1,1}^{\mathrm{H}}+1}}\right)^{\frac{1}{2}} - \frac{\gamma_2 P_{1,1}^{\mathrm{H}} + 1}{\gamma_2} + \frac{e^R - 1}{\gamma_1} \tag{31}$$

$$= \left(\frac{e^R}{\frac{\gamma_2^2}{\gamma_2 P_{1,1}^{\mathrm{H}}+1}}\right)^{\frac{1}{2}} - \frac{1}{\gamma_2} = P_{2,2}^{\mathrm{H}},$$

which means that the lemma holds for the base case.

B. Inductive Step

Assume that the lemma holds for the case of m, which means that

$$\sum_{i=1}^{m} P_{i,1}^{\mathrm{H}} = \cdots = \sum_{i=m-1}^{m} P_{i,m-1}^{\mathrm{H}} = P_{m,m}^{\mathrm{H}}. \tag{32}$$

The aim of this section is to show that the lemma also holds for the case of $m + 1$, i.e.,

$$\sum_{i=1}^{m+1} P_{i,1}^{\mathrm{H}} = \cdots = \sum_{i=m}^{m+1} P_{i,m}^{\mathrm{H}} = P_{m+1,m+1}^{\mathrm{H}}. \tag{33}$$

Recall that U_{m+1}'s transmit power in the first $(m+1)$ time slots can be written as follows:

$$P_{m+1,i}^{\mathrm{H}} = \left(\frac{e^R}{\prod_{p=1}^{m+1} \frac{\gamma_{m+1}}{\gamma_{m+1} \sum_{j=p}^{m} P_{j,p}^{\mathrm{H}}+1}}\right)^{\frac{1}{m+1}} - \frac{\gamma_{m+1} \sum_{j=i}^{m} P_{j,i}^{\mathrm{H}} + 1}{\gamma_{m+1}}, \tag{34}$$

for $1 \leq i \leq m+1$.

We note that $P_{m+1,i}^{\mathrm{H}} = P_{m+1,p}^{\mathrm{H}}$, for $1 \leq i, p \leq m$, by using the assumption made in (32), as explained in the following.

The key observation is that the first term on the right-hand side of (34) is the same for all $P^{\mathrm{H}}_{m+1,i}$. Therefore, the conclusion that $P^{\mathrm{H}}_{m+1,i} = P^{\mathrm{H}}_{m+1,p}$, for $1 \le i, p \le m$, can be established if the following equality holds

$$\frac{\gamma_{m+1}\sum_{j=i}^{m} P^{\mathrm{H}}_{j,i} + 1}{\gamma_{m+1}} = \frac{\gamma_{m+1}\sum_{j=p}^{m} P^{\mathrm{H}}_{j,p} + 1}{\gamma_{m+1}}, \tag{35}$$

which is true given the assumption made in (32). Because $P^{\mathrm{H}}_{m+1,i} = P^{\mathrm{H}}_{m+1,p}$, for $1 \le i, p \le m$, the use of (32) leads to the following conclusion:

$$\sum_{i=1}^{m+1} P^{\mathrm{H}}_{i,1} = \cdots = \sum_{i=m}^{m+1} P^{\mathrm{H}}_{i,m}, \tag{36}$$

which proves a part of (33). Therefore, the proof of the lemma can be completed by showing that $\sum_{i=1}^{m+1} P^{\mathrm{H}}_{i,1} = P^{\mathrm{H}}_{m+1,m+1}$, which is challenging to prove directly. We note that U_{m+1}'s achievable data rates in the first m time slots are identical, i.e.,

$$\log\left(1 + \frac{\gamma_{m+1}P^{\mathrm{H}}_{m+1,i}}{\gamma_{m+1}\sum_{j=i}^{m} P^{\mathrm{H}}_{j,i}}\right) = \log\left(1 + \frac{\gamma_{m+1}P^{\mathrm{H}}_{m+1,p}}{\gamma_{m+1}\sum_{j=p}^{m} P^{\mathrm{H}}_{j,p}}\right), \tag{37}$$

since $\sum_{j=i}^{m} P^{\mathrm{H}}_{j,i} = \sum_{j=p}^{m} P^{\mathrm{H}}_{j,p}$ as shown in (32) and $P^{\mathrm{H}}_{m+1,i} = P^{\mathrm{H}}_{m+1,p}$, for $1 \le i, p \le m$. Therefore, the $P^{\mathrm{H}}_{m+1,i}$, $1 \le i \le m+1$, are also the optimal solution of the following optimization problem:

$$\min_{P_{m+1,1}, P_{m+1,m+1}} \quad mP_{m+1,1} + P_{m+1,m+1} \tag{P10a}$$

$$s.t. \quad m\log\left(1 + \frac{\gamma_{m+1}P_{m+1,1}}{\gamma_{m+1}\sum_{j=1}^{m} P_{j,1} + 1}\right) + \log\left(1 + \gamma_{m+1}P_{m+1,m+1}\right) \ge R. \tag{P10b}$$

The difference between problems (P10) and (P9) is that in problems (P10), the first m time slots are merged together to become a single time slot with duration mT. In other words, problem (P10) can be viewed as a simple two-user case. By following steps similar to those for solving problem (P9), the alternative expressions for $P^{\mathrm{H}}_{m+1,1}$ and $P^{\mathrm{H}}_{m+1,m+1}$ can be obtained as follows:

$$P^{\mathrm{H}}_{m+1,1} = \left(\frac{e^R}{\psi^m\gamma_{m+1}}\right)^{\frac{1}{m+1}} - \frac{I_m\gamma_{m+1} + 1}{\gamma_{m+1}}, \tag{38}$$

$$P^{\mathrm{H}}_{m+1,m+1} = \left(\frac{e^R}{\psi^m\gamma_{m+1}}\right)^{\frac{1}{m+1}} - \frac{1}{\gamma_{m+1}},$$

where $I_m = \sum_{j=1}^{m} P^{\mathrm{H}}_{j,1}$ and $\psi = \frac{\gamma_{m+1}}{I_m\gamma_{m+1}+1}$.

Proving $\sum_{i=1}^{m+1} P^{\mathrm{H}}_{i,1} = P^{\mathrm{H}}_{m+1,m+1}$ is equivalent to showing $P^{\mathrm{H}}_{m+1,1} + I_m = P^{\mathrm{H}}_{m+1,m+1}$, which holds since

$$P^{\mathrm{H}}_{m+1,1} + I_m = \left(\frac{e^R}{\psi^m\gamma_{m+1}}\right)^{\frac{1}{m+1}} - \frac{I_m\gamma_{m+1} + 1}{\gamma_{m+1}} + I_m \tag{39}$$

$$= \left(\frac{e^R}{\psi^m\gamma_{m+1}}\right)^{\frac{1}{m+1}} - \frac{1}{\gamma_{m+1}} = P^{\mathrm{H}}_{m+1,m+1}.$$

Therefore, the proof of the lemma is complete.

Appendix C
Proof for Lemma 3

Recall that the following observations:

1) The power allocation solution shown in Lemma 1 is an optimal solution of problem (P9);
2) The feasible set of problem (P9) is larger than that of problem (P2), which means that the optimal value of problem (P9) is no larger than that of problem (P2);

By using these facts, the lemma that the power allocation solution shown in Lemma 1 is the optimal solution of problem (P2) can be proved, by showing that this solution is feasible to problem (P2), i.e., $P^{\mathrm{H}}_{m,i} > 0$, $1 \le i \le m$.

We first note that U_m always uses the m-th time slot (the OMA time slot), i.e., $P^{\mathrm{H}}_{m,m} > 0$, which can be straightforwardly established, as shown in the following. Recall from (29) that $P^{\mathrm{H}}_{m,m}$ can be expressed as follows:

$$P^{\mathrm{H}}_{m,m} = \left(\frac{e^R}{\prod_{p=1}^{m}\frac{\gamma_m}{\gamma_m\sum_{j=p}^{m-1}P^{\mathrm{H}}_{j,p}+1}}\right)^{\frac{1}{m}} - \frac{1}{\gamma_m}. \tag{40}$$

If $P^{\mathrm{H}}_{m,m} \le 0$, the following inequality needs to hold

$$\left(\frac{e^R}{\prod_{p=1}^{m}\frac{\gamma_m}{\gamma_m\sum_{j=p}^{m-1}P^{\mathrm{H}}_{j,p}+1}}\right)^{\frac{1}{m}} \le \frac{1}{\gamma_m}. \tag{41}$$

The above inequality is equivalent to the following:

$$e^R \le \prod_{p=1}^{m}\frac{1}{\gamma_m\sum_{j=p}^{m-1}P^{\mathrm{H}}_{j,p}+1}. \tag{42}$$

We note that $e^R > 1$, since $R > 0$. However, $\prod_{p=1}^{m}\frac{1}{\gamma_m\sum_{j=p}^{m-1}P^{\mathrm{H}}_{j,p}+1} \le 1$, which means the inequality in (42) cannot hold, and hence $P^{\mathrm{H}}_{m,m}$ is strictly positive.

Therefore, the lemma can be proved showing that U_m's hybrid NOMA power allocations for the first $m-1$ time slots need to be also strictly positive, i.e., $P^{\mathrm{H}}_{m,i} > 0$, for $1 \le i \le m-1$, which can be proved by mathematical induction.

A. The Base Case $m = 2$

For the special case of $m = 2$, by using (29), U_2's transmit power during the first time slot is given by

$$P^{\mathrm{H}}_{2,1} = \left(\frac{e^R}{\frac{\gamma_2^2}{\gamma_2P^{\mathrm{H}}_{1,1}+1}}\right)^{\frac{1}{2}} - \frac{\gamma_2P^{\mathrm{H}}_{1,1}+1}{\gamma_2}. \tag{43}$$

Therefore, the OMA mode is used if $P^{\mathrm{H}}_{2,1} \le 0$, i.e.,

$$\left(\frac{e^R}{\frac{\gamma_2^2}{\gamma_2P^{\mathrm{H}}_{1,1}+1}}\right)^{\frac{1}{2}} \le \frac{\gamma_2P^{\mathrm{H}}_{1,1}+1}{\gamma_2}, \tag{44}$$

which can be simplified as follows:

$$R \le \log\left(1 + \gamma_2P^{\mathrm{H}}_{1,1}\right) < \log\left(1 + \gamma_1P^{\mathrm{H}}_{1,1}\right), \tag{45}$$

where the last inequality follows by the fact that $\gamma_1 > \gamma_2$. By using the fact that $P_{1,1}^{\mathrm{H}}$ is chosen to ensure $\log\left(1+\gamma_1 P_{1,1}^{\mathrm{H}}\right) = R$, the following contradiction can be established:

$$R < \log\left(1+\gamma_1 P_{1,1}^{\mathrm{H}}\right) = R, \tag{46}$$

which means that $P_{2,1} > 0$, and hence for the special case of $m = 2$, the lemma holds.

B. Inductive Step

Assume that the lemma holds for the case of m, i.e., $P_{j,i}^{\mathrm{H}} > 0$, $i \leq m-1$ and $j \leq m$, which makes Lemma 2 applicable and leads to the following equality:

$$\sum_{j=1}^{m} P_{j,1}^{\mathrm{H}} = \cdots = \sum_{j=m-1}^{m} P_{j,m-1}^{\mathrm{H}} = P_{m,m}^{\mathrm{H}} \triangleq I_m. \tag{47}$$

The aim of this section is to prove that the lemma also holds for the case of $m+1$, i.e., $P_{m+1,i}^{\mathrm{H}} > 0$, $i \leq m$, which is also challenging to prove directly. Recall that U_{m+1}'s achievable data rates during the first m time slots are given by

$$R_{m+1,i} = \log\left(1 + \frac{\gamma_{m+1} P_{m+1,i}^{\mathrm{H}}}{\gamma_{m+1}\sum_{j=i}^{m} P_{j,i}^{\mathrm{H}} + 1}\right), \tag{48}$$

for $1 \leq i \leq m$. The use of (47) indicates that U_{m+1} suffers the same amount of interference (I_m) during each of the first m time slots. In addition, the user's channel gains during the first m time slots are also the same. Therefore, U_{m+1}'s transmit powers during the first m time slots, i.e., $P_{m+1,i}^{\mathrm{H}}$, $1 \leq i \leq m$, must be the same, i.e., $P_{m+1,1}^{\mathrm{H}} = \cdots = P_{m+1,m}^{\mathrm{H}}$. In this case, U_{m+1}'s transmit powers can be alternatively obtained from the following optimization problem:

$$\min_{P_{m+1,1}, P_{m+1,m+1}} \quad m P_{m+1,1} + P_{m+1,m+1} \tag{P11a}$$

$$s.t. \quad m\log\left(1 + \frac{\gamma_{m+1} P_{m+1,1}}{\gamma_{m+1} I_m + 1}\right) + \log\left(1 + \gamma_{m+1} P_{m+1,m+1}\right) \geq R, \tag{P11b}$$

which is identical to problem (P10). Therefore, $P_{m+1,1}^{\mathrm{H}}$ can be expressed as follows:

$$P_{m+1,1}^{\mathrm{H}} = \left(\frac{e^R}{\psi^m \gamma_{m+1}}\right)^{\frac{1}{m+1}} - \frac{I_m \gamma_{m+1} + 1}{\gamma_{m+1}}. \tag{49}$$

To ensure $P_{m+1,1}^{\mathrm{H}} > 0$, the following inequality needs to hold:

$$\left(\frac{e^R}{\psi^m \gamma_{m+1}}\right)^{\frac{1}{m+1}} > \frac{I_m \gamma_{m+1} + 1}{\gamma_{m+1}}, \tag{50}$$

which can be rewritten as follows

$$R > \log\left(\frac{\psi^m \left(I_m \gamma_{m+1} + 1\right)^{m+1}}{h_{m+1}^m}\right). \tag{51}$$

Recall that $\psi = \frac{\gamma_{m+1}}{I_m \gamma_{m+1}+1}$, and hence the inequality in (51) can be expressed as follows:

$$\begin{aligned} R &> \log\left(\frac{\frac{h_{m+1}^m}{\left(I_m\gamma_{m+1}+1\right)^m}\left(I_m\gamma_{m+1}+1\right)^{m+1}}{h_{m+1}^m}\right) \\ &= \log\left(I_m\gamma_{m+1}+1\right). \end{aligned} \tag{52}$$

In order to show that $R > \log\left(I_m\gamma_{m+1}+1\right)$, the fact that the users' channel gains are ordered can be used to show the following inequality:

$$\log\left(I_m\gamma_{m+1}+1\right) < \log\left(I_m\gamma_m+1\right) = \log\left(P_{m,m}^{\mathrm{H}}\gamma_m+1\right), \tag{53}$$

where the last step follows from the equality in (47).

The stationarity of the KKT conditions for problem (P9) leads to the following conclusion:

$$\begin{aligned} &\sum_{i=1}^{m-1}\log\left(1+\frac{\gamma_m P_{m,i}^{\mathrm{H}}}{\gamma_m\sum_{j=i}^{m-1}P_{j,i}^{\mathrm{H}}+1}\right) \\ &+\log\left(P_{m,m}^{\mathrm{H}}\gamma_m+1\right) = R. \end{aligned} \tag{54}$$

By using (54) and also the assumption that $P_{m,i}^{\mathrm{H}} > 0$ for $1 \leq i \leq m-1$, the following inequality can be established:

$$\log\left(P_{m,m}^{\mathrm{H}}\gamma_m+1\right) < R. \tag{55}$$

By combining (53) with (55), $R > \log\left(I_m\gamma_{m+1}+1\right)$ is proved, which means that $P_{m+1,i}^{\mathrm{H}} > 0$, $1 \leq i \leq m$. Therefore, the proof of the lemma is complete.

Appendix D
Proof of Lemma 4

The lemma can be again proved by mathematical induction. For the base case $m = 1$, it is straightforward to show that there is a single optimal solution for problem (P2).

For the inductive step, assume that for the case $m-1$, the lemma holds, i.e., U_i, $1 \leq i \leq m-1$, chooses hybrid NOMA, which makes Lemma 2 applicable. The aim of the proof is to show that the lemma holds for the case of m.

Since problem (P2) is convex, its optimal solution needs to satisfy the KKT conditions. By analyzing the KKT conditions, it is straightforward to show that if there exists another optimal solution for problem (P2), one or multiple $P_{m,i}$ need to be zero. Without loss of generality, denote $\mathcal{S}$ as the subset that collects the indices of $P_{m,i}$, which are zero, i.e., $P_{m,i} = 0$, for $i \in \mathcal{S}$. Define $\mathcal{S}^c$ as the complementary set of $\mathcal{S}$. We note that m must be included in $\mathcal{S}^c$ since $P_{m,m}$ cannot be zero, as discussed in the proof of Lemma 3. Therefore, the use of the KKT condition shown in (24) leads to the following conclusion:

$$1 - \lambda b_{m,i}\gamma_m = 0, i \in \mathcal{S}, \tag{56}$$

which can be rewritten as follows:

$$0 = 1 - \frac{\lambda\gamma_m}{\gamma_m\sum_{j=i}^{m-1}P_{j,i}+1} = 1 - \frac{\lambda\gamma_m}{\gamma_m P_{m-1,m-1}+1}, \tag{57}$$

where the first step follows by $b_{m,i} = \frac{1}{\gamma_m\sum_{j=i}^{m-1}P_{j,i}+1}$, and the last step follows by Lemma 2. Therefore, the fact that $P_{m,i} = 0$, $i \in \mathcal{S}$, leads to the following expression of λ:

$$\lambda = P_{m-1,m-1} + \frac{1}{\gamma_m}. \tag{58}$$

On the other hand, (24) can be used to obtain the following expression for $P_{m,j}$, $j \in \mathcal{S}^c$:

$$P_{m,j} = \lambda - \frac{1}{b_{m,i}\gamma_m}, j \in \mathcal{S}^c. \tag{59}$$

Because of the complementary slackness condition, the $P_{m,j}$, $j \in \mathcal{S}^c$, need to satisfy the following condition:

$$\sum_{j\in\mathcal{S}^c} \log\left(1 + b_{m,j}\gamma_m P_{m,j}\right) = R, \tag{60}$$

which can be rewritten as follows:

$$\prod_{j\in\mathcal{S}^c} \lambda b_{m,j}\gamma_m = e^R. \tag{61}$$

By using the expressions of $b_{m,j}$ and λ in (58), the above equality can be expressed as follows:

$$\left(\gamma_m P_{m-1,m-1} + 1\right)^{|\mathcal{S}^c|} \prod_{j\in\mathcal{S}^c} \frac{1}{\gamma_m \sum_{j=i}^{m-1} P_{j,i} + 1} = e^R. \tag{62}$$

By applying Lemma 2, (62) can be simplified as follows:

$$\gamma_m P_{m-1,m-1} + 1 = e^R. \tag{63}$$

Recall that $P_{m-1,m-1}$ is a function of γ_{m-1} and not related to γ_m. Therefore, the probability for the equality in (63) to hold is zero since the users' channel gains are independent fading. Therefore, the solution shown in Lemma 1 is the only optimal solution of problem (P2), which completes the proof of Lemma 4.

REFERENCES

[1] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple Access Techniques for 5G Wireless Networks and Beyond*. Springer International Publishing, 2019.

[2] X. You, C. Wang, J. Huang *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 110301, pp. 1–74, Feb. 2021.

[3] D. Bepari, S. Mondal, A. Chandra, R. Shukla, Y. Liu, M. Guizani, and A. Nallanathan, "A survey on applications of cache-aided NOMA," *IEEE Commun. Surveys & Tutorials*, vol. 25, no. 3, pp. 1571–1603, 2023.

[4] S. Pakravan, J.-Y. Chouinard, X. Li, M. Zeng, W. Hao, Q.-V. Pham, and O. A. Dobre, "Physical layer security for NOMA systems: Requirements, issues, and recommendations," *IEEE Internet of Things J.*, vol. 10, no. 24, pp. 21 721–21 737, 2023.

[5] X. Pei, Y. Chen, M. Wen, H. Yu, E. Panayirci, and H. V. Poor, "Next-generation multiple access based on NOMA with power level modulation," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1072–1083, 2022.

[6] S. McWade, A. Farhang, and M. F. Flanagan, "Low-complexity reliability-based equalization and detection for OTFS-NOMA," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6779–6792, 2023.

[7] W. Feng, J. Tang, Q. Wu, Y. Fu, X. Zhang, D. K. C. So, and K.-K. Wong, "Resource allocation for power minimization in RIS-assisted multi-UAV networks with NOMA," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6662–6676, 2023.

[8] L. Yuan, Q. Du, N. Yang, F. Fang, and N. Yang, "Performance analysis of IRS-aided short-packet NOMA systems over Nakagami-m fading channels," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 8228–8233, 2023.

[9] X. Mu and Y. Liu, "Exploiting semantic communication for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, 2023.

[10] P. Swami, V. Bhatia, S. Vuppala, and T. Ratnarajah, "User fairness in NOMA-hetnet using optimized power allocation and time slotting," *IEEE Systems Journal*, vol. 15, no. 1, pp. 1005–1014, 2021.

[11] Z. Ding, "NOMA beamforming in SDMA networks: Riding on existing beams or forming new ones?" *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 868–871, Apr. 2022.

[12] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access," *IEEE Trans. Veh. Tech.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[13] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of NOMA toward next generation multiple access (NGMA) for 6G," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1037–1071, Jan. 2022.

[14] J. Choi and J.-B. Seo, "Evolutionary game for hybrid uplink NOMA with truncated channel inversion power control," *IEEE Transa. Commun.*, vol. 67, no. 12, pp. 8655–8665, 2019.

[15] J. Zheng, X. Tang, X. Wei, H. Shen, and L. Zhao, "Channel assignment for hybrid NOMA systems with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1370–1374, 2021.

[16] X. Wen, H. Zhang, H. Zhang, and F. Fang, "Interference pricing resource allocation and user-subchannel matching for NOMA hierarchy fog networks," *IEEE J. Sel. Topics Signal Process*, vol. 13, no. 3, pp. 467–479, 2019.

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

[18] N. Nomikos, T. Charalambous, D. Vouyioukas, G. K. Karagiannidis, and R. Wichman, "Hybrid NOMA/OMA with buffer-aided relay selection in cooperative networks," *IEEE J. Sel. Topics Signal Process*, vol. 13, no. 3, pp. 524–537, 2019.

[19] Q. Wang, H. Chen, C. Zhao, Y. Li, P. Popovski, and B. Vucetic, "Optimizing information freshness via multiuser scheduling with adaptive NOMA/OMA," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1766–1778, 2022.

[20] Z. Ding, D. Xu, R. Schober, and H. V. Poor, "Hybrid NOMA offloading in multi-user MEC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5377–5391, 2022.

[21] L. Liu, B. Sun, Y. Wu, and D. H. K. Tsang, "Latency optimization for computation offloading with hybrid NOMA–OMA transmission," *IEEE Internet of Things J.*, vol. 8, no. 8, pp. 6677–6691, 2021.

[22] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA-MEC offloading," *IEEE Wireless Commun. Lett.*, vol. 68, no. 6, pp. 6207–6211, Jun. 2019.

[23] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient MEC in NOMA-enabled massive IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1015–1027, Apr. 2021.

[24] 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, Sept. 2016.

[25] C. Chaieb, F. Abdelkefi, and W. Ajib, "Deep reinforcement learning for resource allocation in multi-band and hybrid OMA-NOMA wireless networks," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 187–198, 2023.

[26] J. Yu, Y. Li, X. Liu, B. Sun, Y. Wu, and D. Hin-Kwok Tsang, "IRS assisted NOMA aided mobile edge computing with queue stability: Heterogeneous multi-agent reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 7, pp. 4296–4312, 2023.

[27] B. Li, W. Wu, W. Zhao, and H. Zhang, "Security enhancement with a hybrid cooperative NOMA scheme for MEC system," *IEEE Trans. Veh. Tech.*, vol. 70, no. 3, pp. 2635–2648, Mar. 2021.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2003.

[29] Z. Ding and H. V. Poor, "Utilizing imperfect resolution of near-field beamforming: A Hybrid-NOMA perspective," *IEEE Commun. Lett.*, (submitted) Available on-line at arXiv:2311.02451.

[30] Z. Wu and L. Dai, "Multiple access for near-field communications: SDMA or LDMA?" *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1918–1935, 2023.

[31] Z. Ding, "Resolution of near-field beamforming and its impact on NOMA," *IEEE Wireless Commun. Lett.*, Available on-line at arXiv:2308.08159, to appear in 2023.

[32] Y. Liu, Z. Wang, J. Xu, C. Ouyang, X. Mu, and R. Schober, "Near-field communications: A tutorial review," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 1999–2049, 2023.

[33] J. Zhu, Z. Wan, L. Dai, M. Debbah, and H. V. Poor, "Electromagnetic information theory: Fundamentals, modeling, applications, and open problems," Available on-line at arXiv:2209.09562, 2022.

[34] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, M. F. Imani, and Y. C. Eldar, "Beam focusing for near-field multiuser MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7476–7490, Sept. 2022.

[35] X. Zhang, H. Zhang, and Y. C. Eldar, "Near-field sparse channel representation and estimation in 6G wireless communications," Available on-line at arXiv:2212.13527, 2022.

[36] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, 2014.