Load Coupling and Energy Optimization in Multi-Cell and Multi-Carrier NOMA Networks

Lei Lei, *Member, IEEE*, Lei You, *Student Member, IEEE*, Yang Yang, *Member, IEEE*, Di Yuan, *Senior Member, IEEE*, Symeon Chatzinotas, *Senior Member, IEEE*, and Björn Ottersten, *Fellow, IEEE*

***Abstract*—In this paper, we investigate energy optimization in multi-cell and multi-carrier non-orthogonal multiple access (NOMA) networks. We apply a load-coupling model for NOMA networks to capture the coupling relation of mutual interference among cells. With this analytical tool, we formulate an energy minimization problem in a NOMA-based load-coupled system, where optimizing load-rate-power allocation, and determining decoding order and user grouping are the key aspects. Theoretically, we prove that the minimum consumed energy can be achieved by using all the time-frequency resources in each cell to deliver users' demand, and allowing all the users to share resource units. From a practical perspective, we consider three types of NOMA grouping schemes, i.e., all-user grouping, partitioned and non-partitioned grouping. We develop tailored solutions for each grouping scheme to enable efficient load-rate-power optimization. These three algorithmic components are embedded into a power-adjustment framework to provide energy-efficient solutions for NOMA networks. Numerical results demonstrate promising energy-saving gains of NOMA over orthogonal multiple access in large-scale cellular networks, in particular for high-demand and resource-limited scenarios. The results also show fast convergence of the proposed algorithms and demonstrate the effectiveness of the solutions.**

***Index Terms*—Non-orthogonal multiple access (NOMA), load coupling, resource allocation, energy minimization.**

I. Introduction

NON-ORTHOGONAL multiple access (NOMA) for the upcoming fifth generation (5G) cellular systems has attracted considerable research attention from both industry and academia over the past few years. In various applications, NOMA has demonstrated significant performance improvement over orthogonal multiple access (OMA) [1] since it is able to sustain aggressive spectrum reuse and alleviate co-channel interference. A majority of previous NOMA works focus on single-cell optimization where the inter-cell interference (ICI) caused by neighboring cells is not present [2]–[6]. For multi-cell NOMA, resource optimization and performance analysis are challenging [7], [8]. The difficulty arises not only from the presence of ICI but also due to the interplay between ICI and the successive interference cancellation (SIC) process. In SIC, the received ICI is typically treated as a factor in determining the decoding order [1], [7], [8]. Any change of the radiated interference in a cell may influence the SIC process and the resource optimization in all the other cells. This coupling effect also imposes obstacles in problem's decomposition, which leads to a challenging optimization task for jointly determining the decoding order, transmit power, and channel resource allocation.

In this regard, some multi-cell NOMA works have to consider simplified scenarios in order to reduce the complexity in performance analysis, e.g., two-cell single-carrier networks [8], two-user single-carrier networks [8], [9], and multi-cell single-carrier networks [10]–[12]. By considering one channel per cell, the authors in [10] investigate sum-rate maximization in multi-cell NOMA for visible light communications. With the same single-carrier assumption, the authors in [11] use Poisson cluster processes to study the performance of multi-cell NOMA in uplink. For studying single-carrier NOMA in heterogeneous networks, the authors in [12] analyze the performance of coverage probability and spectrum efficiency by modeling the positions of macro base stations (BSs) and small BSs as Poisson point processes. In some works, e.g., [11], [12], stochastic geometry has been considered to model ICI in NOMA. On the one hand, stochastic geometry is capable of capturing the topological randomness of NOMA networks [1], and evaluating the average network performance. On the other hand, the derivation of closed-form expressions for large-scale network analysis remains challenging [1], [12]. In addition, difficulties remain in studying specific network topologies in realistic deployment.

In the context of multi-cell and multi-carrier NOMA, the authors in [13] study a power-minimization and a sum-rate maximization problem for this type of networks. By their

Manuscript received July 13, 2019; revised September 6, 2019; accepted September 9, 2019. Date of publication September 25, 2019; date of current version November 12, 2019. The article was supported by the European Research Council (ERC) under Project AGNOSTIC (742648) and FNR CORE under Project ROSETTA (11632107). The works of D. Yuan and L. You were supported by the Swedish Research Council. This article was presented in part at IEEE GLOBECOM, Abu Dhabi, UAE, Dec. 2018 [20]. The review of this article was coordinated by Dr. Z. Ding. *(Corresponding author: Lei Lei.)*

L. Lei, S. Chatzinotas, and B. Ottersten are with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City 1855, Luxembourg (e-mail: lei.lei@uni.lu; symeon.chatzinotas@uni.lu; bjorn.ottersten@uni.lu).

L. You and D. Yuan are with the Department of Information Technology, Uppsala University, Uppsala 75237, Sweden (e-mail: lei.you@it.uu.se; di.yuan@it.uu.se).

Y. Yang is with Competence Center for High Performance Computing, Fraunhofer ITWM, Kaiserslautern 69663, Germany (e-mail: yang.yang@itwm.fraunhofer.de).

Digital Object Identifier 10.1109/TVT.2019.2943701

assumptions, determining the SIC decoding order is simplified to be independent with ICI, which leads to a suboptimal decoding order and a tractable problem (a linear programming problem for power minimization). In general, there is a lack of analytical tools to facilitate performance analysis and exploit the deep insights of jointly optimizing decoding order, power, and channel allocation. Recently, the authors in [19], [20] extended a load-coupling model from multi-cell and multi-carrier OMA to NOMA to study the load-balance performance. The load-coupling model has been widely used for ICI modeling in OMA networks [14]–[18]. The concept of *load* is the fractional portion of the used resource units (RUs) in a cell, ranging from zero to one [14]–[18]. The received ICI in a cell is depending on and proportional to the other cells' load, thus leading to a *load-coupled system* or an *interference-coupled system*. The model is shown to be able to provide good approximation for the network-level interference characterization [17], [18].

The performance optimization for multi-cell and multi-carrier NOMA networks, and the energy-saving issues in large-scale NOMA systems are studied to a limited extent in the literature. In this work, by adopting the load-coupling model, we explore the insights for energy minimization problems in multi-carrier and multi-cell NOMA systems. As the incremental contributions compared to the existing works, e.g., [13], [19], we provide novel analytical results to answer several open research questions. To minimize power/energy in multi-cell multi-carrier NOMA systems:

- What is the optimal operating load in each cell for energy savings?
- How to achieve target load via optimizing power among BSs?
- What is the optimal user-grouping scheme in NOMA for energy minimization?
- How to jointly determine the SIC decoding order and power-load-rate optimization under various practical user-grouping schemes?

The main contributions of this paper are summarized as follows: Firstly, we formulate the energy minimization problem in a NOMA-based load-coupling system (NLCS), where optimizing load, rate, power allocation, and determining decoding order and user grouping in NOMA are intertwined. The problem appears non-linear and non-convex. We characterize a complete solution by developing a set of reformulation and approximation approaches.

Secondly, through the derived analytical results, we prove that the minimum energy consumption is achieved by operating every cell at the full load status, i.e., consuming all the RUs in each cell. Given full load or any other load level as a target to achieve, we develop an algorithmic framework with fast convergence to enable joint decoding order determination and energy minimization.

Thirdly, the proposed framework is applied under three typical user-grouping schemes in NOMA, i.e., *all-user* (the group including all the users), *partition* (no user overlapped among groups), and *non-partition* (allowing user overlapped among groups). Specifically in the framework, we derive a closed-form solution for the all-user grouping. Theoretically we prove that this grouping scheme is optimal for energy minimization in NLCS though it is not practical. Towards practical NOMA, for partitioned grouping, the framework adopts a bisection-search based method to obtain the power that results in the targeted load. For non-partitioned grouping, the load-rate-power optimization is more complex. By relying on the property of pseudo-convexity, we characterize the convexity of the problem's feasible region, and propose a suboptimal solution in the framework based on Lagrangian relaxation and difference of convex functions (DC) programming.

Finally, the proposed algorithmic framework along with the above three embedded sub-algorithms, provide energy-saving solutions for large-scale NOMA networks. We use numerical studies to demonstrate that the network energy savings can benefit from operating each cell at a higher load level, instead of lower load levels, and allowing more users to share the same RU. The results also show the competitiveness and effectiveness of the proposed algorithmic solution in NOMA over conventional OMA.

TABLE I
NOTATIONS

N	number of RUs per cell
B	bandwidth per RU
I	number of cells
K	number of users
$\mathcal{K}_i$	set of the associated users in cell i
$\mathcal{S}_i$	set of all the possible clusters for cell i, $\lvert\mathcal{S}_i\rvert = 2^{\lvert\mathcal{K}_i\rvert} - 1$
$\mathcal{S}_i^*$	set of the adopted clusters in cell i
$\mathcal{U}_s^i$	set of the users in cluster $s \in S_i$ in cell i
$\mathcal{I}$	set of all the cells $\mathcal{I} = \{1, \ldots, I\}$
p^i	transmit power per RU in cell $i \in \mathcal{I}$
p_{ks}^i	transmit power for user k in cell i's cluster s
g_k^i	channel gain between BS i and user k
r_{ks}^i	data rate of user k in cluster s in cell i
R_k	user k's rate demand
$\boldsymbol{r}_s^i$	vector of collecting all the user rate r_{ks}^i for cell i and cluster s
$\boldsymbol{l}_i$	load vector $[l^1, \ldots, l^{i-1}, l^{i+1}, \ldots, l^I]$
$\boldsymbol{p}$	power vector $[p^1, \ldots, p^i, \ldots, p^I]$
$\boldsymbol{p}_i$	power vector $[p^1, \ldots, p^{i-1}, p^{i+1}, \ldots, p^I]$
l^i	load of cell i, $0 \le l^i \le 1$
l_s^i	load of cluster s in cell i, where $0 \le l_s^i \le 1$, $\sum_{s \in S_i} l_s^i = l_i$
η	noise power
C_k	received ICI plus noise power for user k

II. System Model

A. Multi-Cell NOMA Networks

We consider a downlink NOMA-based network with multiple BSs. The overall bandwidth in a cell is divided into multiple subcarriers or RUs. We denote an RU as the minimum time-frequency resource, e.g., a resource block in long term evolution (LTE) systems. The key notations are summarized in Table I. Note that unlike OMA, each RU in NOMA can be simultaneously accessed by multiple users. We then use a term "cluster" to represent a set of users, e.g., a two-user cluster $\{k, k'\}$, $k, k' \in \mathcal{K}$. A cluster can also be referred to as a user group/set presented in other works. For cell i with $|\mathcal{K}_i|$ users, there are $2^{|\mathcal{K}_i|} - 1$ possible clusters in total, including the clusters with only one user (referred to as the clusters for

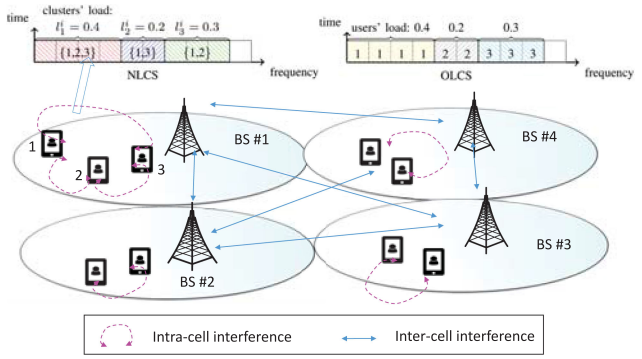


Fig. 1. An illustrative example, 9 out of 10 RUs are used in BS 1, thus the cell's load is $l^1 = 0.9$. In NLCS, cell's load is the sum of clusters' load, e.g., the consumed load for clusters $\{1,2,3\}$, $\{1,3\}$, $\{1,2\}$ ($s = 1,2,3$) is 0.4, 0.2, and 0.3, respectively. NOMA is applied within each cluster. In comparison, cell's load in OLCS is the sum of users' load, e.g., the consumed load for user 1, 2, and 3 is 0.4, 0.2, and 0.3, respectively.

OMA). Each RU can be allocated by up to one cluster to deliver the intra-cluster users' data demand. In this work we consider uniform power p^i over all the RUs in a cell, which is a common assumption adopted in load-coupling systems for network-level performance analysis [14]–[18]. In NLCS, when any cluster is allocated a RU, the total power for this cluster is p^i. With this power budget, we then optimize the power allocation among the users within the cluster, i.e., variable p_{ks}^i. We assume perfect SIC in NOMA, and consider full knowledge of channel state information is available.

B. Interference Modeling in Multi-Cell NOMA

For ICI modeling in multi-cell NOMA, we extend the load-coupling model [15]–[18] from OMA to NOMA. Due to the exclusive user-RU allocation in OMA, i.e., at most one user can access a subcarrier/RU at a time, thus in an OMA based load-coupling system (OLCS), cell's load is the summation of its associated users' load, i.e., the fractional portion of the used RUs for serving a specific user [14]–[16]. Unlike OMA, NOMA has removed this exclusivity. As a result, the load expression for OMA is incorrect for NLCS. The previously derived conclusions in OLCS may not be applicable to NLCS. In NLCS, since one RU/subcarrier can accommodate at most one cluster, we define cell's load by the summation of the clusters' load, i.e., $\sum_{s\in\mathcal{S}_i} l_s^i = l^i$. We use Fig. 1 to show an example of load calculation in OLCS and NLCS.

In multi-cell NOMA, the interference consists of two parts, i.e., the interference among the intra-cluster users, and the ICI among cells. The ICI is treated as noise, and part of the intra-cluster users' interference can be eliminated by applying SIC [1]. The users' decoding order in cell i's cluster s is determined by the descending order of $\frac{g_k^i}{C_k}$ [1], [3], [7], [19], where C_k is the received ICI plus noise power for user k. We define $C_k = \sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_k^j + \eta$ by adopting a widely used ICI modeling approach [14]–[18], where the product $p^j l^j g_k^j$ is introduced to approximate the radiated ICI from cell j to cell i's user k. The rationale of using cell's load to scale the amount of ICI is that when a BS operates at a certain load level over a period of time, e.g., in Fig. 1 BS1 operating at load $= 0.9$ for a few hundred milliseconds (or hundreds/thousands of time slots), the total number of the used RUs for data transmission is fixed, e.g., 9 RUs used in BS1. However, it could be highly random and dynamic from time slot to time slot to decide which 9 RUs are used. In some time slots cell i and cell j may use completely different RUs thus no interference, whereas in some slots the same RU may be used in both cell i and cell j thus introducing interference. Then a follow-up question is how to model/calculate the ICI for the time scale of interest. In general, two approaches can be considered, exact or approximated ICI modeling. The former has to examine every RU's usage in every cell over every time slot since in the same time slot, to exactly know the received ICI of a RU in a cell, the usage of the same RU (with the same frequency band) in all the other cells should be known. However, this exact approach will result in a large amount of signaling overhead and complexity. In practice, it may not be a suitable approach for evaluating average network-level performance. In contrast, the approximated ICI modeling approach has been widely adopted in many papers [14]–[20]. During a period of operation at a certain load level, a RU in cell i may or may not receive interference from cell j over time slot to time slot. Statistically, the probability of a RU in cell i receiving the ICI from cell j is equivalent to the load value l^j (between 0 and 1) [15], [17], [18]. The approach has been proved in [17], [18] that it is able to provide a good approximation for ICI and for the average network performance.

We would like to remark that in this work we focus on the average network-level performance evaluation via introducing load coupling. The time scale of interest in the load-coupling model is typically from a few hundred milliseconds to a few seconds when cell's load is modeled as the fractional usage of cell's resources [21]. For modeling the long-term average cell load, i.e., over hours, other formations are more suited, e.g., by queuing-theoretic formulations to model a cell's load as the ratio of traffic intensity and cell capacity [21].

With the known ICI for cluster s, we use $b_s(k)$ to represent the position of user $k \in \mathcal{U}_s^i$ in the sorted decoding sequence in cluster s. For example, in cell i, if a cluster s consists of three users $\mathcal{U}_s^i = \{1,2,3\}$, and with ratios $\frac{g_3^i}{C_3} > \frac{g_1^i}{C_1} > \frac{g_2^i}{C_2}$, correspondingly the users' positions in this descending order are $b_s(1) = 2$, $b_s(2) = 3$, and $b_s(3) = 1$. The signal-to-interference-and-noise ratio (SINR) of user k in cluster $s \in \mathcal{S}_i$ is presented below.

$$\mathrm{SINR}_{ks}^i = \frac{p_{ks}^i g_k^i}{\sum\limits_{\substack{h\in\mathcal{U}_s^i\setminus\{k\}:\\ b_s(h)<b_s(k)}} p_{hs}^i g_k^i + \sum\limits_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_k^j + \eta} \tag{1}$$

The entity $\sum_{h\in\mathcal{U}_s^i\setminus\{k\}:b_s(h)<b_s(k)} p_{hs}^i g_k^i$ is the intra-cluster interference after performing a successful SIC at user k in cluster s. By applying SIC for the users within cluster s, the interference from the users $h \in \mathcal{U}_s^i\setminus\{k\} : b_s(h) > b_s(k)$ can be decoded and removed [1], [8]. Since all the users in cluster s share a common load l_s^i, thus any user $k \in \mathcal{U}_s^i$ is subject to an equation system (2), where Nl_s^i is the number of used RUs for serving cluster s,

and users' power allocation satisfies $\sum_{k\in\mathcal{U}_s^i} p_{ks}^i = p^i$.

$$BNl_s^i \log\left(1+\frac{p_{ks}^i g_k^i}{\sum\limits_{\substack{h\in\mathcal{U}_s^i\setminus\{k\}:\\ b_s(h)<b_s(k)}} p_{hs}^i g_k^i + \sum\limits_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_k^j + \eta}\right) = r_{ks}^i,\ \forall k\in\mathcal{U}_s^i \tag{2}$$

For illustration, we use the two-user cluster $\{1,2\}$ with cluster load 0.3 in Fig. 1 to explain (2). Suppose $\frac{g_1^i}{C_1}\geq\frac{g_2^i}{C_2}$, where $C_1=\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_1^j+\eta$ and $C_2=\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_2^j+\eta$. According to the descending order of $\frac{g_k^i}{C_k}$, user 2 is assumed to be always able to decode its desired signal x_2. User 1 at its receiver can decode this signal x_2 only if it has higher SINR of signal x_2 at user 1's receiver than at user 2's receiver [7], i.e., $\frac{p_{2\,s}^i g_1^i}{p_{1\,s}^i g_1^i+\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_1^j+\eta}\geq\frac{p_{2\,s}^i g_2^i}{p_{1\,s}^i g_2^i+\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_2^j+\eta}$. As a result, user 2 does not perform SIC, whereas user 1 can decode and remove the intra-cluster interference by applying SIC. Deriving from (2), the rates for user 1 and 2 in cluster $\{1,2\}$ are,

$$B\times 10\times 0.3\log\left(1+\frac{p_{1s}^i g_1^i}{\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_1^j+\eta}\right)=r_{1s}^i$$

$$B\times 10\times 0.3\log\left(1+\frac{p_{2s}^i g_2^i}{p_{1s}^i g_2^i+\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_2^j+\eta}\right)=r_{2s}^i.$$

Then the load equation system for cluster $\{1,2\}$ in Fig. 1 can be expressed as,

$$0.3=\frac{r_{1s}^i}{10B\log\left(1+\frac{p_{1s}^i g_1^i}{\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_1^j+\eta}\right)}=\frac{r_{2s}^i}{10B\log\left(1+\frac{p_{2s}^i g_2^i}{p_{1s}^i g_2^i+\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_2^j+\eta}\right)}, \tag{3}$$

where $p_{1s}^i+p_{2s}^i=p^i$. With the presence of (3), optimizing power p_{1s}^i and p_{2s}^i (or rates r_{1s}^i and r_{2s}^i) should lead to the same resulted load 0.3 for user 1 and 2. In the remaining analysis of the paper, to keep conciseness, we normalize $BN=1$ without loss of any generality.

III. Optimization Problem

By applying the established interference-coupling model, we aim at investigating the optimal energy-saving strategy to satisfy all the users' rate demand in large-scale NOMA networks. Before formulating the optimization problem, we first characterize how the cell's power $p^1,\ldots,p^i,\ldots,p^I$ related to each other in the load equation (2). The characterization will be used to facilitate the problem formulation and the proofs in later sections. We introduce a power vector $\bar{\boldsymbol{p}}_i=[p^1,\ldots,p^{i-1},p^{i+1},\ldots,p^I]$ collecting power $p^1,\ldots,p^I$ except the i-th element p^i. Analogously, we define a load vector by $\bar{\boldsymbol{l}}_i=[l^1,\ldots,l^{i-1},l^{i+1},\ldots,l^I]$.

From the load equations in (2), we can derive p_{ks}^i for each k in cluster s. Without loss of generality, suppose an arbitrary cluster $\mathcal{U}_s^i=\{1,\ldots,K\}$ and the decoding order is consistent with the user indexes $1,\ldots,K$ in cluster s. Then $p_{1s}^i,\ldots,p_{Ks}^i$ read as,

$$\begin{aligned}
p_{1s}^i &= \left(e^{\frac{r_{1s}^i}{l_s^i}}-1\right)\left(\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_1^j+\eta\right)\Big/ g_1^i\\
&\vdots\\
p_{Ks}^i &= \left(e^{\frac{r_{Ks}^i}{l_s^i}}-1\right)\left(\sum_{k'=1}^{K-1} p_{k's}^i g_K^i+\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_K^j+\eta\right)\Big/ g_K^i
\end{aligned} \tag{4}$$

From the above, the expression p_{ks}^i $(k\geq 2)$ contains $p_{1s}^i,\ldots,p_{k-1,s}^i$. Thus in $p_{2s}^i,\ldots,p_{Ks}^i$, we can sequentially substitute each p_{ks}^i with the expressions of $p_{1s}^i,\ldots,p_{k-1,s}^i$. By completing the whole substitution process, we can explicitly express $p^i=p_{1s}^i+,\ldots,+p_{Ks}^i$ by all the other cells' power in (5).

$$p^i=\sum_{k=1}^{K}\left(\frac{C_k}{g_k^i}-\frac{C_{k-1}}{g_{k-1}^i}\right)e^{\frac{\sum_{h=k}^{K} r_{hs}^i}{l_s^i}}-\frac{C_K}{g_K^i} \tag{5}$$

From (2) to (5), the substitution approach has been widely adopted in the literature, e.g., [4], [13], [22], we thus omit the detailed steps to to avoid redundancy but use a two-user case, $\mathcal{U}_s^i=\{1,2\}$, to illustrate. Power $p^i=p_{1s}^i+p_{2s}^i$, where $p_{1s}^i=(e^{\frac{r_{1s}^i}{l_s^i}}-1)\frac{C_1}{g_1^i}$, and by replacing p_{1s}^i in p_{2s}^i, $p_{2s}^i=(e^{\frac{r_{2s}^i}{l_s^i}}-1)((e^{\frac{r_{1s}^i}{l_s^i}}-1)\frac{C_1}{g_1^i}+\frac{C_2}{g_2^i})$. Then $p^i=p_{1s}^i+p_{2s}^i$ in (5) reads $(e^{\frac{r_{1s}^i}{l_s^i}}-1)\frac{C_1}{g_1^i}+(e^{\frac{r_{2s}^i}{l_s^i}}-1)((e^{\frac{r_{1s}^i}{l_s^i}}-1)\frac{C_1}{g_1^i}+\frac{C_2}{g_2^i})=\frac{C_1}{g_1^i}e^{\frac{r_{1s}^i+r_{2s}^i}{l_s^i}}+(\frac{C_2}{g_2^i}-\frac{C_1}{g_1^i})e^{\frac{r_{2s}^i}{l_s^i}}-\frac{C_2}{g_2^i}$.

According to (5), p^i depends on the load allocation among clusters, rate allocation among the users within a cluster, and the received ICI from all the other cells, i.e., power and load vectors $\bar{\boldsymbol{p}}_i=[p^1,\ldots,p^{i-1},p^{i+1},\ldots,p^I]$ and $\bar{\boldsymbol{l}}_i=[l^1,\ldots,l^{i-1},l^{i+1},\ldots,l^I]$. We then define a function f to express p^i in (6). Vector $\boldsymbol{r}_s^i$ is the collection of all rate elements r_{ks}^i for cluster s in cell i, $\forall k\in\mathcal{U}_s^i$.

$$p^i=f(l_s^i,\boldsymbol{r}_s^i,\bar{\boldsymbol{l}}_i,\bar{\boldsymbol{p}}_i),\quad \forall i\in\mathcal{I},\ \forall s\in\mathcal{S}_i \tag{6}$$

In the following, we formulate an energy optimization problem. The variables keep consistence with (6). The optimization tasks are to determine power per RU in each cell, i.e., $p^1,\ldots,p^i,\ldots,p^I$, load level in each cell, i.e., $l^1,\ldots,l^i,\ldots,l^I$, the load allocation among clusters, i.e., l_s^i, and the rate allocation among the users within a cluster, i.e., r_{ks}^i.

$$\text{P0:}\quad \min_{p^i,l_s^i,r_{ks}^i}\ \sum_{i\in\mathcal{I}} p^i\sum_{s\in\mathcal{S}_i} l_s^i \tag{7a}$$

$$\text{s.t.}\quad \sum_{s\in\mathcal{S}_i} r_{ks}^i\geq R_k,\ \forall i\in\mathcal{I},\forall k\in\mathcal{K}_i \tag{7b}$$

$$p^i = f(l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i, \bar{\boldsymbol{p}}_i),\ \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}_i \quad (7c)$$

$$\sum_{s\in\mathcal{S}_i} l_s^i \le 1,\ \forall i \in \mathcal{I} \quad (7d)$$

The objective is to minimize the network energy consumption. The term energy is defined by the product $p^i l^i$ (or $p^i \sum_{s\in\mathcal{S}_i} l_s^i$) which measures the energy consumption at BS i [16]. This is because l^i reflects the normalized amount of RUs (due to $BN = 1$) used in time or frequency domain [14]–[18]. For example, load $l^i = 0.5$ could mean that in cell i half of the frequency resources (e.g., frequency bands) or time resources (e.g., time slots) are consumed. For the former, it is measured in Watt, and for the latter, the metric is energy in Joule. For generality, throughout the paper, we refer to this product as an energy metric. Note that in (7a), originally, the objective is $\sum_{i\in\mathcal{I}} p^i N \sum_{s\in\mathcal{S}_i} l_s^i$. However, we omit N in objectives without loss of optimality since we consider uniform N for every cell, and normalize the term $NB = 1$ in (7a) and (7c). In (7d), the load level of each cell should be no more than one. Constraints (7b) ensure that all the users' rate demands are satisfied. Constraints (7c) characterize the load equation system and confine the feasible region for the load, rate, and power variables.

It is worth noted that the optimization tasks in P0 are not only to optimize the power, rate, load variables, but also jointly determine the SIC decoding order in NOMA. In this paper, the SIC decoding order in NOMA is not predefined but is an outcome of the optimization in P0. The difficulty is that these optimization decisions are intertwined and dependent with each other. To determine one term, the others must be (temporarily) fixed. That is, in order to optimize power, load, and rate, one has to know the decoding order in each cell since the closed-form expression of p^i in (7c) depends on the decoding order in NOMA. However, when power and load, i.e., ICI, are undetermined, the decoding order cannot be derived. This introduces difficulties in jointly determining optimal decoding order and power in each cell.

Remark: The formulated optimization problem and the load-coupling model are proposed to evaluate the average network-level performance without worrying about the slot-by-slot variations in carrier allocation. Unlike the exact ICI modeling, the received ICI in a time slot for a RU in a cell does not need to know whether the the RU with the same frequency band is used in all the other cells. Instead, the received ICI in cell i, $\sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_k^j$, is proportional to the other cells' load and power. This approach simplifies the carrier allocation (the allocation between RUs and clusters) and reduces the signaling overhead. One only needs to determine *how many* RUs (or how much load) to be allocated to a cluster, instead of knowing exactly *which* RUs to be assigned to a cluster in every cell and every time slot. □

The optimal solution for solving P0 is not immediately clear, due to the non-linearity and non-convexity in (7a) and (7c). The product of load and power results in a non-linear objective function. The variables of power, load, and rate in (7c) are intertwined in a non-linear equation system in each cluster. In (7c), one can observe that power p^i is exponential with r_{ks}^i and l_s^i, see (5). Hence, any inappropriate allocation in load and rate, e.g., allocating low load value l_s^i to cluster s but intending to achieve high rate r_{ks}^i, can possibly result in sharp increase in power/energy consumption.

IV. Optimality Characterizations

To optimally solve P0, the following questions need to be addressed. Firstly, what is the optimal load, i.e., l^i and l_s^i, for energy minimization in NLCS? It is not clear whether the energy-saving performance can benefit from using fewer RUs or more RUs to serve users' demands. Secondly, how to optimize each cell's transmit power p^i to achieve the targeted load? Thirdly, what are the optimal clustering scheme and the optimal rates in each cluster? In this section, we provide analyses and solutions for the above questions.

A. Optimal Operating Load in NLCS

We start from dealing with the optimal operating load for energy minimization in NLCS. The result is formalized in Theorem 1.

Theorem 1: At the optimum of P0, $l^i = 1, \forall i \in \mathcal{I}$.

Proof: We prove the conclusion by constructing a contradiction. Suppose at the optimum, there exists at least one cell i with $0 < l^i < 1$. With all the other cells' power p^j and load l^j fixed, $j \neq i$, we increase cell i's load by an arbitrary small value $\beta > 0$, i.e., $l^i + \beta$, to serve the same user demand, or equivalent to adding β to any used cluster's load l_s^i, i.e., $l_s^i + \beta$. According to (5), the resulting power per RU strictly decreases, denoted as $p^i - \beta'$, when load increases. With the proof given below, we claim that the new load $l^i + \beta$ and power $p^i - \beta'$ will result in less energy consumption in cell i and less interference to the other cells.

Based on (5), the product of $p^i l_s^i$ can be seen as a function of l_s^i,

$$f(l_s^i) = l_s^i \left[\sum_{k=1}^{K} \left(\frac{C_k}{g_k^i} - \frac{C_{k-1}}{g_{k-1}^i} \right) e^{\frac{\sum_{h=k}^{K} r_{hs}^i}{l_s^i}} - \frac{C_K}{g_K^i} \right] \quad (8)$$

Note that for presentation convenience, we use the same notation in (5), i.e., $\mathcal{U}_s^i = \{1, \ldots, K\}$ and $\frac{g_1^i}{C_1} \ge, \ldots, \ge \frac{g_K^i}{C_K}$. Then the first-order derivative of $f(l_s^i)$ in l_s^i is,

$$f'(l_s^i) = \sum_{k=1}^{K} \left[\left(\frac{C_k}{g_k^i} - \frac{C_{k-1}}{g_{k-1}^i} \right) e^{\frac{\sum_{h=k}^{K} r_{hs}^i}{l_s^i}} \left(1 - \frac{\sum_{h=k}^{K} r_{hs}^i}{l_s^i} \right) \right] - \frac{C_K}{g_K^i} \quad (9)$$

To see the negativity/positivity of $f'(l_s^i)$, we derive the second-order derivative $f''(l_s^i)$ which is shown to be non-negative.

$$f''(l_s^i) = \frac{\sum_{k=1}^{K} \left[\left(\frac{C_k}{g_k^i} - \frac{C_{k-1}}{g_{k-1}^i} \right) e^{\frac{\sum_{h=k}^{K} r_{hs}^i}{l_s^i}} \left(\sum_{h=k}^{K} r_{hs}^i \right)^2 \right]}{(l_s^i)^3} \ge 0, \quad (10)$$

Thus $f'(l_s^i)$ monotonically increases when l_s^i increases. One can observe that $\lim_{l_s^i \to \infty} f'(l_s^i) = 0$. If l_s^i approaches ∞, $f'(l_s^i)$ in (9) becomes $\frac{C_1}{g_1^i} + (\frac{C_2}{g_2^i} - \frac{C_1}{g_1^i}) +, \ldots, +(\frac{C_K}{g_K^i} - \frac{C_{K-1}}{g_{K-1}^i}) - \frac{C_K}{g_K^i} = 0$. Therefore we can conclude $f'(l_s^i) \leq 0$ for $l_s^i \in (0, 1]$. This implies that when any cluster's load l_s^i in cell i has been increased to $l_s^i + \beta$, the power p^i decreases to $p^i - \beta'$. Then the product $(p^i - \beta')(l_s^i + \beta)$ strictly decreases than $p^i l_s^i$. As a result, less interference is generated from cell i, then the users' resulting rate in all the other cells will increase since higher SINR achieves. Therefore the constrains (7b) can be satisfied. Thus the new pair $(p^i - \beta,' l_s^i + \beta)$ reduces energy without violating any of constraints in P0, which implies that the assumption $l^i < 1$ is not optimal. By contradiction, the minimum energy is obtained until cell's load achieves full load, i.e., $l^i = 1$. Hence the conclusion. ■

In the considered NLCS, load and power are variables and dependent with each other The values of load and power are subject to the load equation in constraints (7c). From the equation, when a cell's load increases, the resulting power monotonically decreases to maintain the operating load level. In order to satisfy users' demand, there are two strategies for power-load allocation. One is to use fewer RUs (lower load) in data transmission but consume more power on each RU, another is to use more RUs (higher load) but less power on each RU. Theorem 1 is used to reveal the fact that the first strategy results in strong ICI but the second strategy, using higher load with less power per RU, will lead to less ICI to other cells.

Regarding the practical meaning of Theorem 1, although theoretically we conclude the optimality of full-load for NLCS, it does not simply suggest the full-load operation for BSs in practice since this will introduce other issues. Instead, a more practical meaning from Theorem 1 is that, if the BSs have choice to operate at either higher or lower load level, then for energy-saving purpose, higher load level (may not necessarily be full load) with lower p^i should be adopted instead of using lower load with higher p^i.

Remark: In Theorem 1, we treat load as a continuous variable to facilitate the analysis. In practice, the load value is discretized by step $1/N$, where N is the total number of RUs per cell, since there are always finite number of RUs available in a cell. Theoretically, by increasing the granularity of the load value, e.g., considering infinite RUs per cell, the performance can ultimately approach the case of continuous load. It is also noted that Theorem 1 in fact reveals the monotonicity that the product of load l^i and power per RU p^i decreases when load l^i increases, no matter how large or small of this increase is. This monotonicity is not changed if one considers, for example, increasing β by 0.1 (10 RUs per cell in the case) or by 0.001 (1000 RUs per cell). Thus the conclusion is applicable to the practical cases with finite RUs per cell. □

B. Optimize Power to Achieve the Target Load

In Theorem 1, we derive the optimal load for P1, and conclude higher load will lead to lower energy consumption. From now, we consider a general l^i, that is, l^i no longer needs to be a variable, and is treated as any feasible load to be achieved. Then a follow-up question is how to find the corresponding power p^i to achieve the full load or any other operating load in all the cells. This is not a trivial task since any change of p^i not only affects the resulting load in cell i but also influences the load in all the other cells due to the presence of ICI. Next, in order to develop a solution for computing optimal p^i, we first characterize the property of power by introducing the concept of standard interference function (SIF), with fixed user rate and cluster load for the moment. If a function $f : \mathbb{R}_+^n \to \mathbb{R}_+^n$ satisfies the following three properties for all input, $\boldsymbol{x} \geq \mathbf{0}$, f is SIF [23].

- Positivity: $f(\boldsymbol{x}) > 0$;
- Monotonicity: If $\boldsymbol{x} \geq \boldsymbol{x}'$, then $f(\boldsymbol{x}) \geq f(\boldsymbol{x}')$.
- Scalability: $\alpha f(\boldsymbol{x}) > f(\alpha \boldsymbol{x})$, for all $\alpha > 1$,.

If $f(\boldsymbol{x})$ is SIF, starting from any initial point and performing fixed-point iteration based algorithm, i.e., the iterative algorithm for power (IAP) proposed in [23], the convergence of the algorithm to the fixed point is guaranteed as long as this point exists. Moreover, this fixed point if exists then it is unique and optimal. We prove that the function $f(\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i)$ is an SIF in $\bar{\boldsymbol{p}}_i$ in Corollary 2 which directly extends the SIF proof in [23]. Note that in (5) or (6), the four factors $l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i$, and $\bar{\boldsymbol{p}}_i$ in $p^i = f(l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i, \bar{\boldsymbol{p}}_i)$ are treated as variables to be determined, whereas in $f(\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i)$, only $\bar{\boldsymbol{p}}_i$ is variable, and $l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i$ are temporarily treated as the fixed terms.

Corollary 2: For any l_s^i, $\boldsymbol{r}_s^i$, $\bar{\boldsymbol{l}}_i$ in cell i, $f(\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i)$ is a standard interference function in $\bar{\boldsymbol{p}}_i$.

Proof: Suppose an arbitrary cluster s, say $\mathcal{U}_s^i = \{1, \ldots, K\}$, is adopted by cell i and $\frac{g_1^i}{C_1} \geq, \ldots, \geq \frac{g_K^i}{C_K}$. From (4), one can observe the positivity. For monotonicity, if we increase any element in vector $\bar{\boldsymbol{p}}_i$ by a positive value $\beta > 0$, e.g., $p^j + \beta$, $\forall j \in \mathcal{I} \backslash \{i\}$, equations in (4) become (11).

$$
\begin{aligned}
p_{1s}^i &= \left(e^{\frac{r_{1s}^i}{l_s^i}} - 1\right)\left(\sum_{j \in \mathcal{I}\backslash\{i\}} (p^j + \beta)\, l^j g_1^j + \eta\right)/g_1^i \\
&\vdots \\
p_{Ks}^i &= \left(e^{\frac{r_{Ks}^i}{l_s^i}} - 1\right)\left(\sum_{k'=1}^{K-1} p_{k's}^i g_K^i + \sum_{j \in \mathcal{I}\backslash\{i\}} (p^j + \beta) l^j g_K^j + \eta\right)/g_K^i
\end{aligned}
\tag{11}
$$

Since all the elements $p_{1s}^i, \ldots, p_{Ks}^i$ are strictly increased, $p^i = p_{1s}^i +, \ldots, + p_{Ks}^i$ increases. Thus $f(\bar{\boldsymbol{p}}_i + \beta; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i) > f(\bar{\boldsymbol{p}}_i; l_i^s, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i)$. In terms of scalability, let

$$
\begin{aligned}
\alpha f(\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i) &= \alpha p^i = \alpha p_{1s}^i +, \ldots, + \alpha p_{Ks}^i \\
&= (e^{\frac{r_{1s}^i}{l_s^i}} - 1)(\alpha \sum_{j \in \mathcal{I}\backslash\{i\}} p^j l^j g_1^j + \alpha\eta)/g_1^i +, \ldots, \\
&+ \left(e^{\frac{r_{Ks}^i}{l_s^i}} - 1\right)\left(\alpha \sum_{k'=1}^{K-1} p_{k's}^i g_K^i + \alpha \sum_{j \in \mathcal{I}\backslash\{i\}} p^j l^j g_K^j + \alpha\eta\right)/g_K^i
\end{aligned}
\tag{12}
$$

Algorithm 1: Alternating Power Adjustment to Achieve the Target Load.

Input: target load $l^1, \ldots, l^i, \ldots, l^I$
Output: power $\boldsymbol{p}^*$
1: Initialize vectors $\boldsymbol{p}'$ and $\boldsymbol{p}^*$
2: **while** $||\boldsymbol{p}^* - \boldsymbol{p}'||_2 > \epsilon$ **do**
3: $\boldsymbol{p}' \leftarrow \boldsymbol{p}^*$
4: **for** $i = 1 : I$ **do**
5: Determine the decoding order in cell i
6: Obtain power p^i that results in the target load l^i
7: $\boldsymbol{p}^* = [p^1, p^2, \ldots, p^I]$

and

$$f(\alpha\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i) = \left(e^{\frac{r_{1s}^i}{l_s^i}} - 1\right)\left(\alpha \sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_1^j + \eta\right)/g_1^i + \ldots,$$
$$+ \left(e^{\frac{r_{Ks}^i}{l_s^i}} - 1\right)\left(\sum_{k'=1}^{K-1} p_{k's}^i g_K^i + \alpha \sum_{j\in\mathcal{I}\setminus\{i\}} p^j l^j g_K^j + \eta\right)/g_K^i \tag{13}$$

We can observe $\alpha f(\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i) > f(\alpha\bar{\boldsymbol{p}}_i; l_s^i, \boldsymbol{r}_s^i, \bar{\boldsymbol{l}}_i)$, hence the conclusion. ■

Motivated by Corollary 2, the corresponding power solution can be obtained by means of an alternating power updating approach (or so called fixed-point iteration approach [16], [23]). That is, we optimize each cell's power p^i one by one for cell $i = 1, \ldots, I$ to achieve the target load l^i in every cell. When p^i is updating, the power variables in the other cells remain unchanged. Observing from (2), cell i's resulting load may fail to achieve the target load when the power has been updated in any other cells. However, having proven the property of SIF, this power adjustment method will converge to a power point that leads to the target load in all cells [23]. We outline a framework in Algorithm 1 for power updating. In Algorithm 1, we declare convergence when the power variation between two successive iterations is no more than a tolerance ϵ. At the convergence, the corresponding power solution is organized in vector $\boldsymbol{p}^*$. The convergence rate of fixed-point iterations is linear [24].

C. Optimal User-Grouping in NLCS

Algorithm 1 provides a general framework for updating p^i from $i = 1$ to I to achieve the target load with convergence. It is worth noting that the key step in Algorithm 1 is Line 6. Thus far, even with known ICI, it is still not clear how to obtain the corresponding p^i in each single cell. This is because different from OLCS, in NLCS one has to determine which clusters to be used in cell i along with the load allocation among the used clusters, i.e., l_s^i, and the users' rate in each cluster, i.e., r_{ks}^i. To obtain optimal p^i as well as l_s^i and r_{ks}^i in each single cell, the optimization task in Line 6 amounts to solving the following problem P1. The optimization variables of P1 are power p^i, clusters' load l_s^i, and rate r_{ks}^i.

$$\text{P1:} \quad \min_{p^i, l_s^i, r_{ks}^i} \; p^i l^i \tag{14a}$$
$$\text{s.t.} \quad \sum_{s\in\mathcal{S}_i} r_{ks}^i \geq R_k, \; \forall k \in \mathcal{K}_i \tag{14b}$$
$$p^i = f\left(\boldsymbol{r}_s^i, l_s^i; \bar{\boldsymbol{l}}_i, \bar{\boldsymbol{p}}_i\right), \; \forall s \in \mathcal{S}_i \tag{14c}$$
$$\sum_{s\in\mathcal{S}_i} l_s^i = l^i \tag{14d}$$

Deriving an optimal solution in P1 is not straightforward as the non-linearity and non-convexity remain. From (14c) (also from (5)), the non-convexity comes from the existence of the fractional formation of variables l_s^i and r_{ks}^i in the exponential operator, i.e., see the term $e^{\sum_{h=k}^{K} r_{hs}^i / l_s^i}$ in (5). However, we observe that this non-convexity can be resolved once the load variables are fixed in the denominator. If there are N RUs in a cell, the minimum load which can be allocated to clusters is $l = \frac{1}{N}$, e.g., $l = 0.1$ in Fig. 1. With the known l^i, the number of used RUs, denoted as $N' = Nl^i$, is fixed. Then the function of p_i in (14c) can be expressed for each used RU below, instead of for each cluster,

$$\sum_{k=1}^{|\mathcal{K}_i|}\left(\frac{C_k}{g_k^i} - \frac{C_{k-1}}{g_{k-1}^i}\right) e^{\frac{\sum_{h=k}^{|\mathcal{K}_i|} r_{hn}^i}{l}} - \frac{C_{|\mathcal{K}_i|}}{g_{|\mathcal{K}_i|}^i} = p^i, \; \forall n \in \mathcal{N}', \tag{15}$$

where $C_0 = 0$, $\frac{g_1^i}{C_1} \geq, \ldots, \geq \frac{g_{|\mathcal{K}_i|}^i}{C_{|\mathcal{K}_i|}}$, and $\mathcal{N}'$ is the set of containing all the used RUs. The notation r_{ks}^i is replaced by r_{kn}^i which represents the rate of user k on RU n in cell i. Note that here we only fix the number of used RUs in cell i, but no need to specify which RUs.

$$\text{P2:} \quad \min_{p^i, r_{kn}^i} \; p^i l^i \tag{16a}$$
$$\text{s.t.} \sum_{n\in\mathcal{N}'} r_{kn}^i \geq R_k, \; \forall k \in \mathcal{K}_i \tag{16b}$$
$$\sum_{k=1}^{|\mathcal{K}_i|}\left(\frac{C_k}{g_k^i} - \frac{C_{k-1}}{g_{k-1}^i}\right) e^{\frac{\sum_{h=k}^{|\mathcal{K}_i|} r_{hn}^i}{l}} - \frac{C_{|\mathcal{K}_i|}}{g_{|\mathcal{K}_i|}^i} = p^i, \; \forall n \in \mathcal{N}' \tag{16c}$$

We then reformulate P1 as a RU-user-rate allocation problem in P2, where the task is to determine rate variables $r_{kn}^i \geq 0$ and the resulting power $p^i \geq 0$. The optimization task of assigning which clusters to RUs in P1 is transformed to allocating which users to RUs in P2. Note that in P2, $|\mathcal{K}_i|$ candidate users are to be allocated on each used RU. This does not mean that we predefine set $\mathcal{K}_i$ as the cluster in each cell, because the optimization procedure is free to determine optimal rates either zero or positive, as well as the formation of clusters on each RU. After the optimization, based on the information of those positive r_{kn}^i, the used clusters on RUs can be derived. It is noticed that the equality constraint functions in (16c) are not affine [25]. However, we show that the optimum of P2, in fact, can be achieved by solving a convex problem. If we relax constraints (16c) in P2 by replacing the equality to inequality (see (17c)),

the relaxed problem P2′ is a convex problem since the objective and constraints (17b) are linear, and the $sum\text{-}exp$ function in (17c) is convex [25].

$$\text{P2}': \min_{p^i, r^i_{kn}} \ p^i l^i \tag{17a}$$

$$\text{s.t.} \sum_{n\in\mathcal{N}'} r^i_{kn} \geq R_k, \quad \forall k \in \mathcal{K}_i \tag{17b}$$

$$\sum_{k=1}^{|\mathcal{K}_i|} \left(\frac{C_k}{g^i_k} - \frac{C_{k-1}}{g^i_{k-1}}\right) e^{\frac{\sum_{h=k}^{|\mathcal{K}_i|} r^i_{hn}}{l}} - \frac{C_{|\mathcal{K}_i|}}{g^i_{|\mathcal{K}_i|}} \leq p^i, \forall n \in \mathcal{N}' \tag{17c}$$

By deriving the Karush-Kuhn-Tucker (KKT) conditions for the optimum of P2′ [25], we show that the inequality constraints (17c) are in fact active at the optimum. Next, we derive the optimal solutions of P2′ in Theorem 3, and summarize the equivalence of P2 and P2′ at optimum in Corollary 4. Firstly, in Theorem 3, we prove that multiplexing all the users to each used RU will lead to the minimum power p^i in P2′. Namely, the optimal cluster in P1 is the cluster consisting of all the users in $\mathcal{K}_i$. We refer to this cluster as the $all\text{-}user$ cluster.

Theorem 3: Multiplexing $|\mathcal{K}_i|$ users to each used RU is optimal for P2′.

Proof: The Lagrangian function of P2′ is written as below.

$$L(p^i, r^i_{kn}; \mu_k, \lambda_n) = l^i p^i + \sum_{k\in\mathcal{K}_i} \mu_k \left(R_k - \sum_{n\in\mathcal{N}'} r^i_{kn}\right) + \sum_{n\in\mathcal{N}'} \lambda_n \left(\sum_{k=1}^{|\mathcal{K}_i|} \left(\frac{C_k}{g^i_k} - \frac{C_{k-1}}{g^i_{k-1}}\right) e^{\frac{\sum_{h=k}^{|\mathcal{K}_i|} r^i_{hn}}{l}} - \frac{C_{|\mathcal{K}_i|}}{g^i_{|\mathcal{K}_i|}} - p^i\right) \tag{18}$$

where $\lambda_n \geq 0, \ \forall n \in \mathcal{N}'$ and $\mu_k \geq 0, \ \forall k \in \mathcal{K}_i$ are the Lagrangian multipliers. Observing the convexity of P2′, its optimal solution can be characterized by the KKT conditions as follows.

$$\frac{\partial L}{\partial p^i} = l^i - \sum_{n\in\mathcal{N}'} \lambda_n = 0, \tag{19a}$$

$$\frac{\partial L}{\partial r^i_{kn}} = \frac{\lambda_n}{l}\left(\sum_{\bar{k}=1}^{k} \left(\frac{C_{\bar{k}}}{g^i_{\bar{k}}} - \frac{C_{\bar{k}-1}}{g^i_{\bar{k}-1}}\right) e^{\frac{\sum_{h=\bar{k}}^{|\mathcal{K}_i|} r^i_{hn}}{l}}\right) - \mu_k = 0, \quad \forall k, n \tag{19b}$$

$$\mu_k \left(R_k - \sum_{n\in\mathcal{N}'} r^i_{kn}\right) = 0, \quad \forall k \in \mathcal{K}_i \tag{19c}$$

$$\lambda_n \left(\sum_{k=1}^{|\mathcal{K}_i|} \left(\frac{C_k}{g^i_k} - \frac{C_{k-1}}{g^i_{k-1}}\right) e^{\frac{\sum_{h=k}^{|\mathcal{K}_i|} r^i_{hn}}{l}} - \frac{C_{|\mathcal{K}_i|}}{g^i_{|\mathcal{K}_i|}} - p^i\right) = 0, \ \forall n \in \mathcal{N}' \tag{19d}$$

Based on (19b), we can derive,

$$r^i_{1n} = l \log \frac{\mu_1 (C_2/g^i_2 - C_1/g^i_1)}{C_1/g^i_1 (\mu_2 - \mu_1)}, \forall n \in \mathcal{N}' \tag{20a}$$

$$\cdots\cdots$$

$$r^i_{|\mathcal{K}_i|n} = l \log \frac{l(\mu_{|\mathcal{K}_i|} - \mu_{|\mathcal{K}_i|-1})}{\lambda_n (C_{|\mathcal{K}_i|}/g^i_{|\mathcal{K}_i|} - C_{|\mathcal{K}_i|-1}/g^i_{|\mathcal{K}_i|-1})}, \forall n \in \mathcal{N}'. \tag{20b}$$

The above equations give the optimal rate solutions as $r^i_{11} = \cdots = r^i_{1\,N'}, r^i_{21} = \cdots = r^i_{2\,N'}, \ldots,$ and $r^i_{|\mathcal{K}_i|-1,1} = \cdots = r^i_{|\mathcal{K}_i|-1,N'}$, which means user's rate demand $R_1, \ldots, R_{|\mathcal{K}_i|-1}$ will be uniformly allocated over all the used RUs. Regarding the last user $k = |\mathcal{K}_i|$, we conclude that at the optimum the equation $r^i_{|\mathcal{K}_i|,1} = \cdots = r^i_{|\mathcal{K}_i|N'}$ also holds. The reason is explained as follows. In (19b) we can observe that the term $\sum_{\bar{k}=1}^{k} (\frac{C_{\bar{k}}}{g^i_{\bar{k}}} - \frac{C_{\bar{k}-1}}{g^i_{\bar{k}-1}}) e^{\frac{\sum_{h=\bar{k}}^{|\mathcal{K}_i|} r^i_{hn}}{l}}$ is positive. If any of multipliers μ_k or λ_n becomes zero, all the other multipliers have to be zeros in order to satisfy (19b), but this violates condition (19a). Thus, the multipliers will be positive at the optimum. As a consequence of the strictly positive multipliers, to achieve the equalities in (19c) and (19d), also considering the uniform rate allocation for users $k = 1, \ldots, |\mathcal{K}_i| - 1$, the optimal rate solution for user $k = |\mathcal{K}_i|$ has to be $r^i_{|\mathcal{K}_i|,1} = \cdots = r^i_{|\mathcal{K}_i|N'}$. Thus the conclusion. ■

Corollary 4: At the optimum, P2′ is equivalent to P2.

From the proof of Theorem 3, all the multipliers λ_n are strictly positive at the optimum. By the complementary slackness condition, the equality of (19d) must hold. As a result, the entity $(\sum_{k=1}^{|\mathcal{K}_i|} (\frac{C_k}{g^i_k} - \frac{C_{k-1}}{g^i_{k-1}}) e^{\frac{\sum_{h=k}^{|\mathcal{K}_i|} r^i_{hn}}{l}} - \frac{C_{|\mathcal{K}_i|}}{g^i_{|\mathcal{K}_i|}} - p^i)$ in (19d) becomes 0, which is equivalent to constraints (16c) in P2. Hence the conclusion.

Corollary 5: At the optimum of P1, $|\mathcal{K}_i|$ users are allocated to each used RU.

The optimal allocation in P1 is consistent with Theorem 3 for P2 and P2′. From P1 to P2, we discretize the continuous load value l^i_s by load step $l = 1/N$. By increasing N, better granularity of l achieves. The two problems are equivalent when N becomes infinite. From the proof of Theorem 3, the derived KKT conditions are independent of N and l. Hence, the same conclusion holds for P1.

The derived analysis in this section now can enable us to outline a complete solution for optimally solving P0, that is, setting full load as the target load and adopting the all-user cluster in each cell, then iteratively updating each cell's power $p^1, \ldots, p^i, \ldots, p^I$ by applying the result of Theorem 3 to achieve the full load. When a cell is processed, the other cells' power stays unchanged. The iterations eventually converge to a power vector which leads to minimum network energy consumption and full load in all the cells.

From a practical perspective, adopting the all-user cluster for each cell may not always be a realistic choice in NOMA systems since more users participating in SIC can result in longer signal processing delay, higher complexity in receiver design, and higher error probability in decoding [8], [26]. In the literature, various clustering schemes are considered according to application scenarios and system requirements. We classify them into three types: **Type-A**: all-user, **Type-B**: partition, and **Type-C**: non-partition. Firstly, the Type-A cluster is used in the scenarios with few users, e.g., two-user NOMA systems [8]. However, if

many users are associated in a cell, this scheme may not be a practical solution [8]. The second type, partition, means that all the scheduled clusters in a cell have no common users, forming a partition of set $\mathcal{K}_i$. For example, if two clusters s and s' are scheduled for cell i, then $\mathcal{U}_s^i \cap \mathcal{U}_{s'}^i = \emptyset$ and $\mathcal{U}_s^i \cup \mathcal{U}_{s'}^i = \mathcal{K}_i$. This type of clustering is widely adopted in multi-antenna NOMA systems. For example in [27]–[29], K users in a cell are partitioned into $\frac{K}{2}$ clusters based on their channel correlation and gain differences to perform beamforming. Unlike the first two types, the Type-C clusters can overlap by some common users, e.g., the clusters $\{1,2,3\}$, $\{1,3\}$, $\{1,2\}$ in Fig. 1. The user-subcarrier allocation (or clusters allocation) in some previous works [2], [14] falls into this domain. The Type-C can be seen as a general and flexible clustering scheme for NOMA.

V. Power Minimization for Three Types of Clustering Schemes

In this section, for the considered three types of clusters, we characterize a set of tailored computation methods or algorithmic solutions for dealing with the corresponding power minimization problems. Note that for Type-B and Type-C, determining the optimal clusters for partition and non-partition schemes is beyond the scope of this paper. We focus on power, rate, and load optimization for the given clusters which can be obtained by the proposed practical and well-known schemes in the literature, e.g., the user-grouping schemes in [5] and [6] to form partition and non-partition clusters, respectively.

A. Power Minimization for All-User Clustering Schemes

In Type-A, only one cluster containing all the users is adopted for cell i. Thus all the associated users' demand R_k will be delivered in this single cluster. Then rate variables are fixed and replaced by parameters R_k. Applying the results from Theorem 3 and Corollary 4, the calculation of p^i is straightforward, given by a closed-form expression:

$$p^i = \sum_{q=1}^{|\mathcal{K}_i|} \left(\frac{C_{(q)}}{g^i_{(q)}} - \frac{C_{(q-1)}}{g^i_{(q-1)}} \right) e^{\frac{\sum_{h=q}^{|\mathcal{K}_i|} R_{(h)}}{l^i}} - \frac{C_{(|\mathcal{K}_i|)}}{g^i_{(|\mathcal{K}_i|)}} \tag{21}$$

where index $1,\ldots,q,\ldots,|\mathcal{K}_i|$ refers to users in descending order of $\frac{g^i_k}{C_k}$, $\forall k \in \mathcal{K}_i$, and (q) is the user in the q-th position in the order.

B. Power-Load Optimization for Partitioned Clustering Schemes

In Type-B, the adopted clusters in cell i form a partition, and power p^i in cell i is subject to the following expression,

$$p^i = \sum_{q=1}^{|\mathcal{U}_s^i|} \left(\frac{C_{(q)}}{g^i_{(q)}} - \frac{C_{(q-1)}}{g^i_{(q-1)}} \right) e^{\frac{\sum_{h=q}^{|\mathcal{U}_s^i|} R_{(h)}}{l^i_s}} - \frac{C_{(|\mathcal{U}_s^i|)}}{g^i_{(|\mathcal{U}_s^i|)}},\ \forall s \in \mathcal{S}_i^*, \tag{22}$$

where $\mathcal{S}_i^*$ is the set containing all the given clusters for cell i. Since $\mathcal{U}_s^i \cap \mathcal{U}_{s'}^i = \emptyset, \forall s, s' \in \mathcal{S}_i^*$ in a partition, one user's demand is delivered by only one cluster. Thus, similar to the all-user cluster, the rates can be fixed as $R_k,\ \forall k \in \mathcal{U}_s^i$. Then the remaining optimization task is to determine the optimal load allocation l_s^i among the clusters of $\mathcal{S}_i^*$. Since p^i is uniform over RUs in a cell, optimizing this single variable can be carried out by a bisection search method. With the power value p^i, the corresponding load l_s^i can be calculated for each cluster by (22). The bisection search for power p^i terminates when the gap between the sum load $\sum_{s\in\mathcal{S}_i^*} l_s^i$ and the target load l^i is less than a predefined tolerance ϵ.

C. Power-Load-Rate Optimization in Non-Partitioned Clustering Schemes

In Type-C, one user's rate demand can be delivered in multiple clusters. Then a follow-up question is how to split a user's rate among the clusters, and how much load or how many RUs should be allocated to each cluster. These aspects make the optimization much more complicated than Type-A and Type-B. Next, with the target load l^i and the given non-partitioned clusters in $\mathcal{S}_i^*$, we formulate the power minimization problem in P3. The optimization tasks are to determine per-RU power p^i, load portion l_s^i for each given cluster, and the rate split r_{ks}^i.

$$\text{P3:} \quad \min_{p^i, l_s^i, r_{ks}^i} \ p^i l^i \tag{23a}$$

$$\text{s.t.} \sum_{s\in\mathcal{S}_i^*} N l_s^i r_{ks}^i \geq R_k,\ \forall k \in \mathcal{K}_i \tag{23b}$$

$$\sum_{q=1}^{|\mathcal{U}_s^i|} \left(\frac{C_{(q)}}{g^i_{(q)}} - \frac{C_{(q-1)}}{g^i_{(q-1)}} \right) e^{\frac{\sum_{h=q}^{|\mathcal{U}_s^i|} r^i_{(h)s}}{l}} - \frac{C_{(|\mathcal{U}_s^i|)}}{g^i_{(|\mathcal{U}_s^i|)}} \leq p^i,\ \forall s \in \mathcal{S}_i^* \tag{23c}$$

$$0 < r_{ks}^i,\ \forall s \in \mathcal{S}_i^*, \forall k \in \mathcal{U}_s^i \tag{23d}$$

$$\frac{1}{N} \leq l_s^i,\ \forall s \in \mathcal{S}_i^* \tag{23e}$$

$$\sum_{s\in\mathcal{S}_i^*} l_s^i = l^i \tag{23f}$$

Analogous to the transformation approach from P1 to P2, the constraints (23b) and (23c) are transformed from their original version (24a) and (24b) by introducing the load granularity $l = 1/N$. As aforementioned, it is difficult to directly address the non-convexity in (24b). Then we transfer this difficulty from (24b) to (23b) in order to solve the problem. One can observe that the non-convex constraints (24b) become convex in (23c), while the linear constraints (24a) become bilinear (23b) which is non-linear but can be addressed by several well-established optimization methods, e.g., McCormick envelope method [30].

$$\sum_{s\in\mathcal{S}_i^*} r_{ks}^i \geq R_k,\ \forall k \in \mathcal{K}_i \tag{24a}$$

$$\sum_{q=1}^{|\mathcal{U}_s^i|} \left(\frac{C_{(q)}}{g^i_{(q)}} - \frac{C_{(q-1)}}{g^i_{(q-1)}} \right) e^{\frac{\sum_{h=q}^{|\mathcal{U}_s^i|} r^i_{(h)s}}{l^i_s}} - \frac{C_{(|\mathcal{U}_s^i|)}}{g^i_{(|\mathcal{U}_s^i|)}} \leq p^i,\ \forall s \in \mathcal{S}_i^* \tag{24b}$$

In (23b), a user's total rate is expressed by $\sum_{s\in\mathcal{S}_i^*} N l_s^i r_{ks}^i$. For example, in Fig. 1, user 1's sum rate over three clusters

$\{1,2,3\}$, $\{1,3\}$, $\{1,2\}$ with respective load 0.4, 0.2, and 0.3, is $10\times 0.4r^i_{11}+10\times 0.2r^i_{12}+10\times 0.3r^i_{13}$. Note that, strictly speaking, the value of l^i_s should be the multiples of $l=1/N$, e.g., 0.1 in Fig. 1. To facilitate the analysis, we consider l^i_s as a continuous variable. This can be approximately achieved when N is sufficient large, e.g., more than thousands RUs per cell, such that the performance loss between continuous and discrete load can be negligible. Due to the uniform p^i in the objective function, analogous to Corollary 4, the inequality constraints (23c) will be active at the optimum. In (23d) and (23e), since the optimization is based on the given clusters then all the included clusters in $\mathcal{S}^*_i$ are required to be used. The load l^i_s for each cluster is at least $1/N$, i.e., allocated by at least one RU, and each rate value should be positive. In constraints (23f), the target load in cell i should be achieved.

The difficulty of solving P3 is the bilinear terms in (23b). In general, a conventional relaxation-and-approximation method can be used to deal with the bilinear term by bounding each variable with lower and upper bounds. The approximation performance will be largely dependent on the tightness of the bounds on the variables l^i_s and r^i_{ks}. Although l^i_s is bounded by a relatively tight interval, i.e., (0,1), the bounds for $r^i_{ks}\in(0,R_k]$ have a much wider rage of variation in practice especially for the scenarios with high data demand. Applying McCormick envelopes for P3 could return a weak lower bound and a poor-quality suboptimal solution (upper bound). It is also possible to express the envelopes as piecewise functions. That is, dividing the domain between the lower and upper bounds into multiple sections, and applying envelopes at each section. The approximation performance can be improved by increasing the number of breakpoints in the envelopes. However, the computational complexity increases exponentially, which imposes obstacles in addressing the large-scale problems in practice. Therefore, for P3 we analyze and exploit the problem's property firstly, and then derive a tailored solution to solve the problem.

$$\text{P3}': \min_{p^i,l^i_s,r^i_{ks},t_{ks}} p^i l^i-\sum_{k\in\mathcal{K}_i}\lambda_k\left(\sum_{s\in\mathcal{S}^*_i}t^2_{ks}-R_k\right) \tag{25a}$$

$$\text{s.t. } t^2_{ks}-Nl^i_s r^i_{ks}\le 0,\ \forall s\in\mathcal{S}^*_i,\forall k\in\mathcal{U}^i_s \tag{25b}$$

$$0<r^i_{ks}\le R_k,\ \forall s\in\mathcal{S}^*_i,\forall k\in\mathcal{U}^i_s \tag{25c}$$

$$0<t_{ks},\ \forall s\in\mathcal{S}^*_i,\forall k\in\mathcal{U}^i_s \tag{25d}$$

$$0<l^i_s,\ \forall s\in\mathcal{S}^*_i \tag{25e}$$

$$\sum_{s\in\mathcal{S}^*_i}l^i_s=l^i \tag{25f}$$

$$\sum_{q=1}^{|\mathcal{U}^i_s|}\left(\frac{C_{(q)}}{g^i_{(q)}}-\frac{C_{(q-1)}}{g^i_{(q-1)}}\right)e^{\frac{\sum_{h=q}^{|\mathcal{U}^i_s|}r^i_{(h)s}}{l}}-\frac{C_{(|\mathcal{U}^i_s|)}}{g^i_{(|\mathcal{U}^i_s|)}}\le p^i,\ \forall s\in\mathcal{S}^*_i \tag{25g}$$

We circumvent the bilinear terms by introducing a set of auxiliary variables $t_{ks},\forall s\in\mathcal{S}^*_i,\forall k\in\mathcal{U}^i_s$, then equivalently expressing the constraints (23b) in the following form,

$$\begin{aligned} &Nl^i_s r^i_{ks}\ge t^2_{ks},\quad \forall s\in\mathcal{S}^*_i,\forall k\in\mathcal{U}^i_s\\ &\sum_{s\in\mathcal{S}^*_i}t^2_{ks}\ge R_k,\quad \forall k\in\mathcal{K}_i \end{aligned} \tag{26}$$

where $\sum_{s\in\mathcal{S}^*_i}t^2_{ks}$ is convex but the constraint $\sum_{s\in\mathcal{S}^*_i}t^2_{ks}\ge R_k$ is not. We develop a solution based on Lagrangian relaxation with DC programming. Firstly, we absorb the non-convex constraints $\sum_{s\in\mathcal{S}^*_i}t^2_{ks}\ge R_k,\forall k\in\mathcal{K}_i$ into the objective function by Lagrangian multipliers $\lambda_k\ge 0,\forall k\in\mathcal{K}_i$. The subproblem of the Lagrangian relaxation is constructed in P3′. The objective function in P3′ can be seen as the difference of two convex functions when the multipliers are given. If the set of feasible region is convex then the problem falls into the domain of DC programming which belongs to a tractable class in non-convex optimization [30].

Next, we characterize the feasible region of P3′. In order to outline the properties of the constraint functions (25b), we first introduce the definition and the second order condition for pseudo-convex functions. The gradient and the Hessian matrix of a function $f(\boldsymbol{x}),\ \boldsymbol{x}\in\mathcal{X}$ evaluated at a point (vector) $\boldsymbol{x}$ are denoted by $\nabla f(\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x})$, respectively, where $\mathcal{X}$ is the convex feasible set.

Definition 1: A function $f(\boldsymbol{x})$ is pseudo-convex if $(\boldsymbol{y}-\boldsymbol{x})^T\nabla f(\boldsymbol{x})\ge 0\Longrightarrow f(\boldsymbol{y})\ge f(\boldsymbol{x}),\ \forall \boldsymbol{x},\boldsymbol{y}\in\mathcal{X}$ [30], [31].

In practice, it may not be easy to identify a pseudo-convex function by its definition, we hence prove the pseudo-convexity of function $f(t_{ks},l^i_s,r^i_{ks})=t^2_{ks}-Nl^i_s r^i_{ks}$ by applying the second order condition, given in Definition 2.

Definition 2: A sufficient condition for a function $f(\boldsymbol{x})$ to be pseudo-convex on $\mathcal{X}$ is that, there exists a real number $0\le\alpha\le\infty$, such that the symmetric matrix $M=\nabla^2 f(\boldsymbol{x})+\alpha\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^T$ is positive (semi-)definite for all $\boldsymbol{x}\in\mathcal{X}$ [32].

Lemma 6: In (25b), the constraint function $f(t_{ks},l^i_s,r^i_{ks})=t^2_{ks}-Nl^i_s r^i_{ks},\quad \forall s\in\mathcal{S}^*_i,\quad \forall k\in\mathcal{U}^i_s$ is pseudo-convex for $Nl^i_s r^i_{ks}-t^2_{ks}>0$.

Proof: For notation-wise simplicity in the proof, we present the function $f(t_{ks},l^i_s,r^i_{ks})$ as $f(t,l,r)$, and assume $N=1$, i.e., $f(t,l,r)=t^2-lr$. By applying the sufficient condition in Definition 2, the symmetric matrix M reads,

$$\begin{aligned} M&=\nabla^2 f(t,l,r)+\alpha\nabla f(t,l,r)\nabla f(t,l,r)^T\\ &=\begin{bmatrix}2&0&0\\0&0&-1\\0&-1&0\end{bmatrix}+\alpha\begin{bmatrix}2t\\-r\\-l\end{bmatrix}[2t,-r,-l]\\ &=\begin{bmatrix}2+4t^2\alpha&-2tr\alpha&-2tl\alpha\\-2tr\alpha&r^2\alpha&rl\alpha-1\\-2tl\alpha&lr\alpha-1&l^2\alpha\end{bmatrix} \end{aligned} \tag{27}$$

The matrix M is positive definite only if all the three leading principal minors are positive, i.e., D_1,D_2,D_3 for order 1, 2, and 3, respectively. We then derive D_1,D_2,D_3 as $D_1=2+4t^2\alpha$, $D_2=(2+4t^2\alpha)r^2\alpha-(2tr\alpha)^2=2r^2\alpha$, and $D_3=2(2\alpha(lr-t^2)-1)$. In D_3, with the strict inequality $lr-t^2>0$, we can

always find a positive and bounded α such that $2\alpha(lr - t^2) - 1 > 0$ for any l, r, and t, e.g., $\alpha > \frac{1}{2(lr-t^2)}$, then D_3 is positive, and D_1, D_2 are therefore positive. Hence the lemma. ■

Motivated by the result of Lemma 6, in order to achieve the pseudo-convexity, constraints (25b) should be further restricted as $t_{ks}^2 - Nl_s^i r_{ks}^i < 0,\ \forall s \in \mathcal{S}_i^*, \forall k \in \mathcal{U}_s^i$. To keep the feasible region as a closed set, we slightly relax the problem P3′ into P3″ by introducing a positive parameter $\epsilon > 0$. In practice, we keep ϵ to be small, thus the optimality gap between P3′ and P3″ can be negligible compared to the total energy consumption.

$$\text{P3}'' :\ \min\ (25a) \tag{28a}$$

$$\text{s.t. } (25f), (25g) \tag{28b}$$

$$t_{ks}^2 - Nl_s^i r_{ks}^i \le -\epsilon,\ \forall s \in \mathcal{S}_i^*,\ \forall k \in \mathcal{U}_s^i \tag{28c}$$

$$\epsilon \le r_{ks}^i \le R_k,\ \forall s \in \mathcal{S}_i^*,\ \forall k \in \mathcal{U}_s^i \tag{28d}$$

$$\epsilon \le t_{ks},\ \forall s \in \mathcal{S}_i^*,\ \forall k \in \mathcal{U}_s^i \tag{28e}$$

$$\epsilon \le l_s^i,\ \forall s \in \mathcal{S}_i^* \tag{28f}$$

We summarize the convexity of the solution set of the reformulation P3″ in Proposition 7.

Proposition 7: Constraints (28b)–(28f) in P3″ form a closed convex set.

Proof: The constraint functions $sum\text{-}exp$ in (25g) are convex [25], and the equality constraint $\sum_{s\in\mathcal{S}_i^*} l_s^i = l^i$ in (25f) is affine. In constraints (28c), as shown in Lemma 6, the function $f(t_{ks}, l_s^i, r_{ks}^i) = t_{ks}^2 - Nl_s^i r_{ks}^i$ is pseudo-convex in the set $\{(t_{ks}, l_s^i, r_{ks}^i) | Nl_s^i r_{ks}^i - t_{ks}^2 \ge \epsilon\}$. The level sets of a pseudo-convex function are convex, and the intersection of convex sets is also convex [30]. Thus, the intersection of convex sets formed by constraints (28b)–(28f) is a closed and convex set. ■

In P3″, once the multipliers λ_k are fixed, the remaining problem in fact minimizes a concave function over a convex set. Although the problem is non-convex, the developed transformation and reformulation approaches make the problem belong to a tractable class in non-convex optimization. Some exact algorithms or efficient suboptimal algorithms can be applied to solve it. The global optimum can be obtained by some exact methods, e.g., by successively enclosing the convex set within a tightening polyhedron [33], but it has exponential computational complexity. To enable an efficient sub-optimal solution, DC programming can be applied. In DC programming, solving a non-convex problem: $\min_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) = f(\boldsymbol{x}) - g(\boldsymbol{x})$ is replaced by successively solving a set of approximated convex problems: $\min_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) = f(\boldsymbol{x}) - [g(\boldsymbol{x}^{(m)}) + \nabla g^T(\boldsymbol{x}^{(m)})(\boldsymbol{x} - \boldsymbol{x}^{(m)})]$ (at the m-th iteration), where $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are convex functions, and $\mathcal{X}$ is a convex set [34]. The procedure of successive convex approximation eventually converges to a stationary point, typically leading to a suboptimal solution in general [3], [34].

To deal with the cases of Type-C clusters, we propose an algorithm based on Lagrangian relaxation with DC programming to provide a feasible and suboptimal solution to P3. The steps are summarized in Algorithm 2. For a set of multipliers, we solve P3″ by DC programming from Lines 5 to 8, where the vector $\boldsymbol{x}$ collects all the variables of P3″, and the convex set $\mathcal{X}$ is formed by constraints (28b)–(28f), and $h(\boldsymbol{x}) = f(\boldsymbol{x}) - g(\boldsymbol{x}) = p^i l^i - \sum_{k\in\mathcal{K}_i} \lambda_k (\sum_{s\in\mathcal{S}_i^*} t_{ks}^2 - R_k)$. In Line 6, the convex approximation problem can be efficiently solved by applying standard convex optimization tools [25]. Once the DC programming converges, we apply subgradient method [30] to update Lagrangian multipliers λ_k in Line 11. When the Lagrangian optimization terminates at Line 13, the optimized load allocation $\hat{l}_s^i$ among the given clusters is a feasible solution, i.e., $\sum_{s\in\mathcal{S}_i^*} \hat{l}_s^i = l^i$, but some users' demand may not be necessarily satisfied due to the application of Lagrangian relaxation. To ensure the solution feasibility, we postprocess the result in Lines 14 to 17. One may notice that once load l_s^i is known, the bilinear term $\sum_{s\in\mathcal{S}_i^*} l_s^i r_{ks}^i$ in (23b) along with its non-convexity are dissolved. The remaining problem of P3: $\min_{p^i, r_{ks}^i} p^i l^i$, s.t. (23b)–(23f) is convex. Then a suboptimal and feasible solution $\boldsymbol{x}^*$ for P3 can be efficiently obtained.

Algorithm 2: Lagrangian Relaxation with DC Programming for Solving P3.

1: Initialize vectors $\boldsymbol{x}^{(0)}$, $|\hat{h} - \bar{h}| > \epsilon$, and set $\hat{m} = 0$.
2: **while** $|\hat{h} - \bar{h}| > \epsilon$ or $\hat{m} \le M_{\max}$ **do**
3: $\bar{h} \leftarrow \hat{h}$
4: $m \leftarrow 0$
5: **while** $|h(\boldsymbol{x}^{(m+1)}) - h(\boldsymbol{x}^{(m)})| > \epsilon$ **do**
6: Solve the convex approximation problem: $\boldsymbol{x}^{(m+1)} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) - g(\boldsymbol{x}^{(m)}) - \nabla g^T(\boldsymbol{x}^{(m)})(\boldsymbol{x} - \boldsymbol{x}^{(m)})$
7: $m \leftarrow m + 1$
8: **end while**
9: $\{\hat{p}^i, \hat{l}_s^i, \hat{t}_{ks}, \hat{r}_{ks}^i,\ \forall s \in \mathcal{S}_i^*, \forall k \in \mathcal{U}_s^i\} \leftarrow \boldsymbol{x}^{(m)}$
10: $\hat{h} \leftarrow h(\boldsymbol{x}^{(m)})$
11: Update multipliers λ_k by the subgradient method
12: $\hat{m} = \hat{m} + 1$
13: **end while**
14: **if** $\sum_{s\in\mathcal{S}_i^*} N\hat{l}_s^i \hat{r}_{ks}^i < R_k,\ \forall k \in \mathcal{K}_i$ **then**
15: Fix the load variables l_s^i to $\hat{l}_s^i,\ s \in \mathcal{S}_i^*$ in P3
16: Solve P3 to obtain a suboptimal and feasible solution $\boldsymbol{x}^*$:
17: $\boldsymbol{x}^* = \arg\min_{p^i, r_{ks}^i} p^i l^i$, s.t. (23b)–(23f)

D. Energy Optimization Framework in NLCS

Based on the framework of Algorithm 1 and the characterizations derived earlier in this section, we develop Algorithm 3 for network energy minimization for the three user-clustering schemes. With a target load $\bar{l}^i$ in each cell, from Line 2 to 18, the power optimization is carried out for cell $i = 1, \ldots, I$ one by one. In each cell i, the SIC decoding order is determined firstly in Line 5, then the proposed three methods are adopted to deal with Type-A, Type-B, and Type-C clustering, respectively:

- If the all-user cluster (Type-A) is used in cell i, based on the result of Theorem 3, optimal p^i can be directly derived by the closed-form expression in Line 7.
- If the clusters used in cell i form a partition (Type-B), the computation for optimal p^i in each iteration is done in Lines 10 to 14.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Cell radius	200 m
Carrier frequency	2 GHz
Bandwidth per cell	9 MHz
Bandwidth per RU	15 KHz
Number of cells	20
Number of users per cell	4, 12
Path loss	COST-231-HATA
Shadowing	Log-normal, 8 dB standard deviation
Fading	Rayleigh flat fading [27]
Noise power spectral density	-173 dBm/Hz
Tolerance ϵ in Algorithm 1–3	10^{-5}
$M_{\max}$ in Algorithm 2	200
Clustering schemes in NOMA	[5], [6]

Algorithm 3: Energy Minimization Framework for NLCS.

Given: target load $\bar{l}^i$, clusters $s \in \mathcal{S}_i^*$ for each cell i
Output: $p^1, \ldots, p^i, \ldots, p^I$
Initialize: load $l^1, \ldots, l^I$, power vectors $\boldsymbol{p}'$ and $\boldsymbol{p}^* = [p^1, \ldots, p^I]$, $(||\boldsymbol{p}^* - \boldsymbol{p}'||_2 > \epsilon)$
repeat
  $\boldsymbol{p}' \leftarrow \boldsymbol{p}^*$
  for $i = 1 : I$ **do**
    Arrange user indexes, such that for indexes $1, \ldots, q, \ldots, |\mathcal{K}_i|$, $\frac{g_k^i}{C_k}$, $\forall k \in \mathcal{K}_i$ is in descending order
    if $\mathcal{S}_i^*$ is **Type-A** clustering **then**
      $p^i = \sum_{q=1}^{|\mathcal{K}_i|} \left(\frac{C_{(q)}}{g_{(q)}^i} - \frac{C_{(q-1)}}{g_{(q-1)}^i} \right) e^{\frac{\sum_{h=q}^{|\mathcal{K}_i|} R^i(h)}{l^i}} - \frac{C_{(|\mathcal{K}_i|)}}{g_{(|\mathcal{K}_i|)}^i}$
    if $\mathcal{S}_i^*$ is **Type-B** clustering **then**
      repeat
        Bisection search for p^i. For a searched p^i **do**
        for each $s \in \mathcal{S}_i^*$ **do**
          Calculate the resulting load l_s^i:
          $p^i = \sum_{q=1}^{|\mathcal{U}_s^i|} \left(\frac{C_{(q)}}{g_{(q)}^i} - \frac{C_{(q-1)}}{g_{(q-1)}^i} \right) e^{\frac{\sum_{h=q}^{|\mathcal{U}_s^i|} R(h)}{l_s^i}} - \frac{C_{(|\mathcal{U}_s^i|)}}{g_{(|\mathcal{U}_s^i|)}^i}$
        until $|\sum_{s \in \mathcal{S}_i^*} l_s^i - \bar{l}^i| \leq \epsilon$
    if $\mathcal{S}_i^*$ is **Type-C** clustering **then**
      Apply Algorithm 2 to obtain p^i
  $\boldsymbol{p}^* = [p^1, \ldots, p^i, \ldots, p^I]$
until $||\boldsymbol{p}^* - \boldsymbol{p}'||_2 \leq \epsilon$

- If the clusters for cell i are in non-partitioned format (Type-C), computation for p^i is done by applying Algorithm 2 which provides a suboptimal and feasible solution.

The algorithm terminates when the distance of the power vector between two successive iteration is less than a tolerance, meanwhile all the cells' load values achieve the target load.

VI. PERFORMANCE EVALUATION

In this section, we present numerical studies to: 1) illustrate the derived theoretical results in previous sections; 2) evaluate the energy-saving gains of network NOMA by the proposed analytical model and algorithm; 3) illustrate the effectiveness of the proposed algorithmic solutions. Table II summarizes the key parameters. In performance evaluation, all the users in each cell are randomly and uniformly distributed. We generate two hundreds instances and consider the average performance.

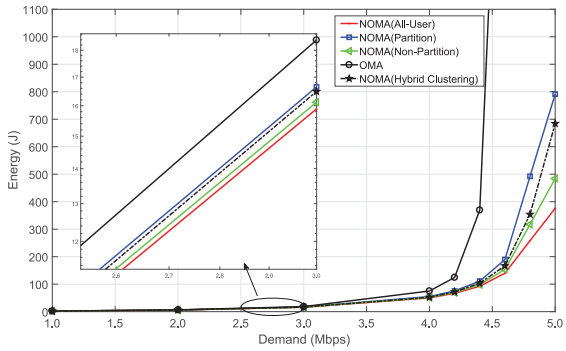


Fig. 2. Energy consumption with respect of demand (four users per cell, load = 1).

TABLE III
PERFORMANCE COMPARISON BETWEEN NOMA AND OMA

Rate Demand (Mbps):	2	3	4	4.2	4.4
$\frac{E_O - E_H}{E_H} \times 100\%$	7.6%	11.5%	44.1%	72.8%	257.3%
$\frac{E_O - E_A}{E_A} \times 100\%$	9.7%	15.7%	50.1%	86.8%	299.3%
$\frac{E_C - E_A}{E_A} \times 100\%$	1.1%	1.4%	3.3%	3.3%	6.3%

For performance comparison, we implement the algorithm proposed in [16] to compute the optimal energy for OMA networks, where the clusters with one user have been used only in OMA, and are excluded from the NOMA schemes. In NOMA, we evaluate the performance from 2-user clustering to 6-user clustering. We adopt the grouping scheme proposed in [5] for Type-B clustering (partition). In the scheme, for example in 2-user clusters, the best-worst user paring/grouping is applied. That is, the ratios $\frac{g_1^i}{C_1}, \ldots, \frac{g_{|\mathcal{K}_i|}^i}{C_{|\mathcal{K}_i|}}$ for all the users in cell i are sorted. The highest-ratio user and the lowest-ratio user are paired into one cluster, while the second highest-ratio user and the second-lowest ratio user are grouped into another cluster, and so on [5]. For Type-C clustering (non-partition), we adopt the proposed "fast optimal user-clustering algorithm" in [6] to generate non-partitioned groups.

In Fig. 2 and Table III, the network energy consumption E_A, E_B, E_C, E_H are obtained for four NOMA clustering schemes, Type-A, Type-B, Type-C, and Hybrid clustering, respectively. The first three NOMA clustering schemes apply the homogeneous type clustering in all the cells, whereas "NOMA (Hybrid Clustering)" is implemented by randomly adopting Type-A, Type-B, Type-C among cells. For Type-B and Type-C in NOMA, 2-user clustering is used in Fig. 2, Table III, and Fig. 3. In addition, E_O stands for the energy consumption in OMA scheme, which is obtained by applying the proposed algorithm in [16]. We remark that the energy values E_A, E_B and E_O are optimal, whereas E_C and E_H are suboptimal in general.

In Fig. 2, we evaluate the energy consumption with respect to user demand. In general, energy consumption in both NOMA and OMA increases exponentially as demand increases. The reason is that due to the demand constraint $\sum_{s \in \mathcal{S}_i} r_{ks}^i \geq R_k$, $\forall i \in$

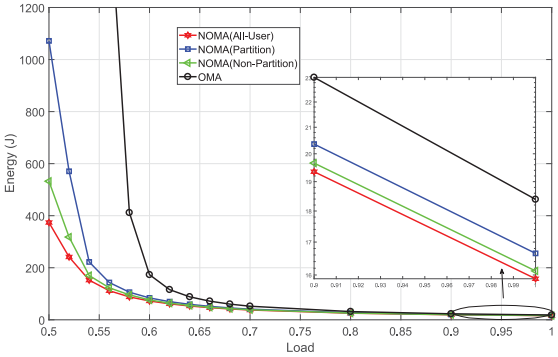


Fig. 3. Energy consumption with respect of load (4 users per cell, demand = 3 Mbps).

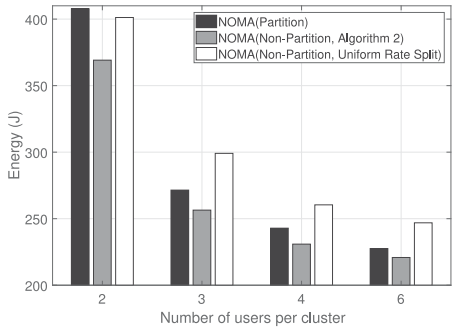


Fig. 4. Energy consumption with respect of the number of users per cluster (12 users in each cell, demand = 4 Mbps, load = 1).

$\mathcal{I}, \forall k \in \mathcal{K}_i$ in optimization, when a user's demand R_k increases, the user's rate split among clusters, i.e., r_{ks}^i, has to increase. From eq. (5), one can observe that when rate r_{ks}^i increases *linearly*, the resulting power p^i increases *exponentially* in NLCS. Analogously, the power-rate function in OLCS also follows the similar exp formation [14]–[18]. Thus, in high-demand cases, e.g., 4–4.5 Mbps, the consumed energy in all the schemes could surge, whereas due to the effect of exp operator, in low-demand cases, e.g., 1–3 Mbps, the energy increases moderately. It is worth noted that power optimization could be infeasible in some cases, that is, even infinite energy is consumed, the target demand and load are not able to deliver/satisfy. For example, if one uses arbitrary small load to serve arbitrary high demand, the optimization problems in NLCS and OLCS become infeasible. Therefore, to enable a feasible or practical power solution in optimization, both load and demand should vary within a certain region.

In addition, we summarize several key information of Fig. 2: Firstly as a result of Theorem 3, Type-A clustering (All-user) indeed yields the minimum energy consumption among the four NOMA schemes. Secondly, applying NOMA for network energy savings is more effective in high-demand scenarios than low-demand cases. Table III further summarizes the performance gaps in Fig. 2. For low-demand cases, the energy-saving gains of NOMA over OMA are marginal, e.g., less than 5% for 1 Mbps, whereas in high-demand instances, all the NOMA schemes demonstrate superior performance, e.g., 2–3 times energy decease in NOMA over OMA at 4.4 Mbps. Thirdly, compared to OMA, NOMA is able to support higher demand with power in its practical range. In Fig. 2, the energy consumption of OMA increases to an unrealistic value ($>10^{10} J$) for 4.4–4.5 Mbps. Although the energy increases dramatically with the demand for all schemes, the rate of increase in all the NOMA schemes is much more moderate than OMA. Fourthly, from the last line of Table III, one can observe that the gaps between E_A and E_C is small. Together with Theorem 3, E_A can be treated as a lower bound and used as a benchmark for gauging the performance of the proposed Algorithm 2 for E_C. The results imply that Algorithm 2 is able to provide a close-to-optimal solution for non-partition clustering.

In Fig. 3, we examine the energy consumption with respect to load. Several observations can be noted. Firstly, the results are in line with Theorem 1. As expected, the minimum energy is achieved with load = 1 in all the NOMA schemes (as well as in OMA). Secondly, NOMA is able to satisfy users' demand by using few bandwidth resources than OMA. For serving the same amount of demand, e.g., 3 Mbps in Fig. 3, all the NOMA schemes consume less than half of RUs (with load between 0.3 and 0.5), thus more RUs can be released for serving the upcoming user demand, whereas the solution in OMA becomes infeasible when the load is less than 0.58. Thirdly, splitting one user's rate demand into multiple clusters, i.e., non-partition clustering, may result in less energy consumption than partition mainly due to the former's diversity and flexibility in cluster format. The performance gaps between the two schemes dramatically increase in scenarios with limited resource (low-load region in Fig. 3) and in high-demand cases, see Fig. 2.

Next in Fig. 4, we illustrate the impact of cluster's size on the energy consumption in NOMA, and show the necessity of performing Algorithm 2 for non-partition clustering. When we introduce more users in each used cluster in Type-B and Type-C, the total energy consistently decreases. This impact is significant from 2-user to 3-user clustering, and becomes marginal for the cases with larger cluster size. In addition, for Type-C (non-partition) clustering, optimizing the rate split for each overlapped user appearing in multiple clusters is of importance. As a simple comparison, when a user's demand is uniformly distributed in multiple clusters, the performance becomes degraded, compared to the result optimized by Algorithm 2.

In Fig. 5, we use NOMA (hybrid-clustering) scheme to illustrate the typical convergence evolution of Algorithm 3. From the result, the required number of iterations to converge largely depends on users' demand, where an iteration is defined by the execution from Lines 4 to 16 in Algorithm 3. For the lower-demand cases, e.g., 1 Mbps and 3 Mbps, the power adjustment procedure can efficiently achieve the target load with 10–20 iterations. For high-demand cases, the convergence could have long-tail effect. It should be remarked that the distance $||\boldsymbol{p}^* - \boldsymbol{p}'||_2$ may not decrease monotonically over iterations, see the small protrusion in blue line. This is because when cells' power is updated, the decoding order could be changed in some cells. Then the expression of p^i will be also adjusted under the new decoding order, possibly resulting in increased distance $||\boldsymbol{p}^* - \boldsymbol{p}'||_2$. On the other hand, with the progress of iterations, the variation

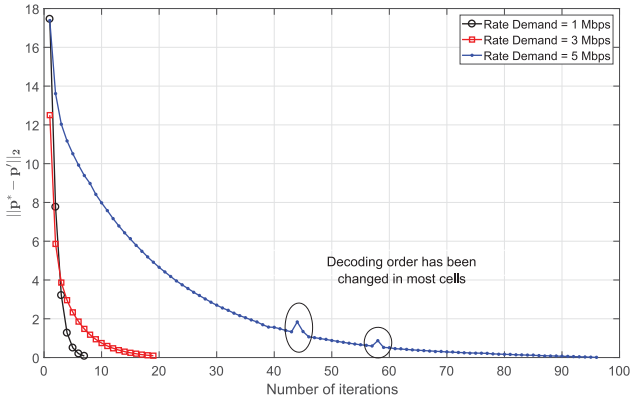


Fig. 5. The convergence evolution of Algorithm 3: the Euclidean distance of cells' power between two successive iterations (12 users per cell, cell's load = 1).

of power between two successive iterations diminishes, and then the decoding order tends to be fixed, and leads to a final convergence.

VII. Conclusion

We have extended an analytical tool, i.e., load-coupling model, from OMA to NOMA for studying the performance of multi-cell and multi-carrier NOMA networks. Towards energy minimization in NOMA networks, we have concluded that operating at full load is optimal for energy savings, and the minimum energy can be achieved by applying the aggressive all-user clustering scheme. We have designed tailored power-adjustment and load-rate optimization algorithms for three types of NOMA clustering schemes. The numerical studies have illustrated the superior performance of NOMA over OMA in network energy savings, particularly in high-demand and low-load instances.

References

[1] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[2] L. Lei, D. Yuan, and P. Värbrand, "On power minimization for non-orthogonal multiple access (NOMA)," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2458–2461, Dec. 2016.

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

[4] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016.

[5] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[6] J. M. Kang and I. M. Kim, "Optimal user grouping for downlink NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 724–727, Oct. 2018.

[7] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun.*, Sep. 2013, pp. 611–615.

[8] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[9] Y. Fu, Y. Chen, and C. W. Sung, "Distributed downlink power control for the non-orthogonal multiple access system with two interfering cells," in *Proc. IEEE Int. Conf. Commun.*, May 2016, pp. 1–6.

[10] X. Zhang, Q. Gao, C. Gong, and Z. Xu, "User grouping and power allocation for NOMA visible light communication multi-cell networks," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 777–780, Apr. 2017.

[11] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using Poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[12] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[13] Z. Yang, C. Pan, W. Xu, Y. Pan, M. Chen, and M. Elkashlan, "Power control for multi-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 927–942, Feb. 2018.

[14] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Optimal cell clustering and activation for energy saving in load-coupled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6150–6163, Nov. 2015.

[15] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.

[16] C. K. Ho, D. Yuan, L. Lei, and S. Sun, "Power and load coupling in cellular networks for energy optimization," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 509–519, Jan. 2015.

[17] A. J. Fehske and G. P. Fettweis, "Aggregation of variables in load models for interference-coupled cellular data networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2012, pp. 5102–5107.

[18] H. Klessig, D. Öhmann, A. J. Fehske, and G. P. Fettweis, "A performance evaluation framework for interference-coupled cellular data networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 938–950, Feb. 2016.

[19] L. You, L. Lei, D. Yuan, S. Sun, S. Chatzinotas, and B. Ottersten, "A framework for optimizing multi-cell NOMA: Delivering demand with less resource," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–7.

[20] L. Lei, L. You, Y. Yang, D. Yuan, S. Chatzinotas, and B. Ottersten, "Power and load optimization in interference-coupled non-orthogonal multiple access networks," in *Proc. IEEE GLOBECOM*, Dec. 2018.

[21] A. J. Fehske, I. Viering, J. Voigt, C. Sartori, S. Redana, and G. P. Fettweis, "Small-cell self-organizing wireless networks," in *Proc. IEEE*, vol. 102, no. 3, pp. 334–350, Mar. 2014.

[22] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. K. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12244–12258, Dec. 2018.

[23] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.

[24] H. R. Feyzmahdavian, M. Johansson, and T. Charalambous, "Contractive interference functions and rates of convergence of distributed power control laws," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4494–4502, Dec. 2012.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

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

[27] Z. Liu, L. Lei, N. Zhang, G. Kang, and S. Chatzinotas, "Joint beamforming and power optimization with iterative user clustering for MISO-NOMA systems," *IEEE Access*, vol. 5, pp. 6872–6884, 2017.

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

[29] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.

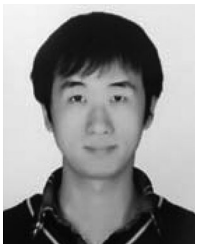
[30] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[31] Y. Yang and M. Pesavento, "A unified successive pseudoconvex approximation framework," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3313–3328, Jul. 2017.

[32] P. Mereau and J.-G. Paquet, "Second order conditions for pseudo-convex functions," *SIAM J. Appl. Math.*, vol. 27, no. 1, pp. 131–137, 1974.

[33] K. L. Hoffman, "A method for globally minimizing concave functions over convex sets," *Math. Program.*, vol. 20, no. 1, pp. 22–32, Dec. 1981.

[34] R. Horst and N. V. Thoai, "DC programming: Overview," *J. Optim. Theory Appl.*, vol. 103, no. 1, pp. 1–43, Oct. 1999.



Lei Lei (S'12–M'17) received the B.Eng. and M.Eng. degrees from Northwestern Polytechnic University, Xi'an, China, in 2008 and 2011, respectively, and the Ph.D. degree from Linköping University, Linköping, Sweden, in 2016. Since November 2016, he has been a Research Associate with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg City, Luxembourg. He was a Research Assistant with the Institute for Infocomm Research, A*STAR, Singapore, from June 2013 to December 2013. His current research interests include resource allocation and optimization in 5G-satellite networks, energy-efficient communications, and deep learning in wireless communications. He received the IEEE Sweden Vehicular Technology-Communications-Information Theory (VT-COM-IT) joint chapter Best Student Journal Paper Award in 2014. He was a co-recipient of the IEEE SigTelCom 2019 Best Paper Award.

Lei You (S'15) received the B.Eng. and M.Eng. degrees from Qingdao University, Qingdao, China, in 2012 and 2014, respectively. He is currently working toward the Ph.D. degree with the Department of Information Technology, Uppsala University, Uppsala, Sweden. He was a Visiting Researcher with Ranplan Wireless Network Design Ltd., U.K., from 2015 to 2016, and Converge ICT, Athens, in 2016. He was a Project Manager of the EU Horizon 2020 Marie-Sklodowska Curie Project DECADE, from 2015 to 2016. His current research interests include mathematical optimization and machine learning in the domain of mobile communications and networking.

Yang Yang (S'09–M'13) received the B.S. degree from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2009, and the Ph.D. degree from The Hong Kong University of Science and Technology. From November 2013 to November 2015, he had been a Research Associate with the Communication Systems Group, Technische Universität Darmstadt, Darmstadt, Germany. From December 2015 to October 2017, he had been a Research Scientist with Intel. From November 2017 to June 2019, he had been a Research Associate with the University of Luxembourg. Since July 2019, he has been a Research Scientist with the Fraunhofer Institute for Applied Mathematics, Fraunhofer, Germany. His research interests are in parallel and distributed solution methods in convex optimization and nonlinear programming with applications in large-scale signal processing.

Di Yuan (M'03–SM'15) received the M.Sc. degree in computer science and engineering, and the Ph.D. degree in optimization from the Linköping Institute of Technology, Linköping, Sweden, in 1996 and 2001, respectively. After the Ph.D. degree, he has been an Associate Professor and then a Full Professor with the Department of Science and Technology, Linköping University, Linköping, Sweden. In 2016, he joined Uppsala University, Sweden, as a Chair Professor. His current research mainly addresses network optimization of 4G and 5G systems, and capacity optimization of wireless networks. He has been a Guest Professor with the Technical University of Milan (Politecnico di Milano), Italy, in 2008, and Senior Visiting Scientist at Ranplan Wireless Network Design Ltd, U.K., in 2009 and 2012. In 2011 and 2013, he was part time with Ericsson Research, Sweden. In 2014 and 2015, he was a Visiting Professor with the University of Maryland, College Park, MD, USA. He is an Area Editor for the *Computer Networks* journal. He has been on the management committee of four European Cooperation in field of Scientific and Technical Research (COST) actions, Invited Lecturer of European Network of Excellence EuroNF, and Principal Investigator of several European FP7 and Horizon 2020 projects. He is a co-recipient of IEEE ICC'12 Best Paper Award, and was the supervisor of the Best Student Journal Paper Award by the IEEE Sweden Joint VT-COM-IT Chapter in 2014.

Symeon Chatzinotas (S'06–M'09–SM'13) received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, Surrey, U.K., in 2006 and 2009, respectively. He is currently the Deputy Head of the SIGCOM Research Group, Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, Luxembourg City, Luxembourg, and a Visiting Professor with the University of Parma, Parma, Italy. He was involved in numerous research and development projects for the Institute of Informatics Telecommunications, National Center for Scientific Research Demokritos, the Institute of Telematics and Informatics, Center of Research and Technology Hellas, and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has authored or coauthored more than 300 publications, 2500 citations, and an H-Index of 27 according to Google Scholar. His research interests include multiuser information theory, co-operative/cognitive communications, and wireless networks optimization. He was a co-recipient of the 2014 Distinguished Contributions to Satellite Communications Award, and the Satellite and Space Communications Technical Committee, the IEEE Communications Society, and the CROWNCOM 2015 Best Paper Award.

Björn Ottersten (S'87–M'89–SM'99–F'04) was born in Stockholm, Sweden, in 1961. He received the M.S. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He has held research positions with the Department of Electrical Engineering, Linköping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, Belgium, and the University of Luxembourg, Luxembourg. From 1996 to 1997, he was the Director of Research with ArrayComm, Inc., a start-up in San Jose, CA, USA, based on his patented technology. In 1991, he was appointed a Professor of Signal Processing with the Royal Institute of Technology (KTH), Stockholm, Sweden. From 1992 to 2004, he was the Head of the Department for Signals, Sensors, and Systems, KTH, and from 2004 to 2008, he was the Dean of the School of Electrical Engineering, KTH. He is currently the Director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City, Luxembourg. As Digital Champion of Luxembourg, he acts as an Adviser to the European Commission. He was a recipient of the IEEE Signal Processing Society Technical Achievement Award in 2011 and the European Research Council advanced research grant twice, in 2009–2013 and in 2017–2022. He has coauthored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, and 2013, and seven other IEEE conference papers best paper awards. He was the Editor-in-Chief of *EURASIP Signal Processing Journal*, Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the Editorial Board of the IEEE SIGNAL PROCESSING MAGAZINE. He is currently a member of the editorial boards of *EURASIP Journal of Advances Signal Processing* and *Foundations and Trends of Signal Processing*. He is a Fellow of EURASIP.