Resource Allocation for Layered Multicast Video Streaming in NOMA Systems

DOI:
10.1109/TVT.2022.3193122

Document Version
Accepted author manuscript

Resource Allocation for Layered Multicast Video Streaming in NOMA Systems

Muhammad Norfauzi Dani, Daniel K. C. So, *Senior Member, IEEE,* Jie Tang, *Senior Member, IEEE,* and Zhiguo Ding, *Fellow, IEEE*

***Abstract*—The application of non-orthogonal multiple access (NOMA) to multi-layer multicast video streaming is envisioned to address the explosively growing demand for capacity which is dominated mostly by multimedia content. In this paper, a joint power allocation and subgrouping scheme is developed to enhance the performance of NOMA-based multi-layer multicast systems. An optimization problem is formulated to maximize the overall sum multicast rate whilst satisfying the maximum transmission power and proportional rate constraints. Due to the complexity of the optimization problem, we first derive two power allocation techniques for the 2-layer case considering arbitrary subgrouping which are based on the iterative implementation and closed-form analysis, respectively. We then generalize the low-complexity closed-form solution for the general multi-layer case. This scheme successively allocates power to each layer stream while assuring that the minimum target rate and proportionality are guaranteed, particularly for the transmission rate of the high-priority base layer stream. A sub-optimal joint power allocation and subgrouping scheme is also designed by incorporating the power allocation scheme into the proposed iterative subgrouping techniques. Simulation results show the effectiveness of the power allocation and subgrouping schemes in enhancing the sum multicast rate performance. In addition, the proposed power allocation scheme ensures substantially higher rates for the base layer stream, which is crucial in robust delivery of standard quality video to all users.**

***Index Terms*—Non-Orthogonal Multiple Access (NOMA), power allocation, subgroup formation, multicast, sum multicast rate maximization**

This paper was presented in parts at IEEE Globecom Workshops 2017 [1]. This work has been supported in part by National Key Research and Development Project under Grant 2019YFB1804100, in part by the National Natural Science Foundation of China under Grant 61971194, 62071364, in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011607, and in part by the Research Fund Program of Guangdong Key Laboratory of Aerospace Communication and Networking Technology under Grant 2018B030322004. (*Corresponding author: Jie Tang.*)

M. N. Dani, D. K. C. So and Z. Ding are with the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, UK. (e-mail: norfauzi.dani@manchester.ac.uk, d.so@manchester.ac.uk, zhiguo.ding@manchester.ac.uk). M. N. Dani is also with the Electrical and Electronic Engineering programme area, Universiti Teknologi Brunei, BE1410, Brunei Darussalam (e-mail: norfauzi.dani@utb.edu.bn).

J. Tang is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710126, China (e-mail: eejtang@scut.edu.cn).

I. Introduction

THE evolved multimedia broadcast multicast service (eMBMS) has been designed for LTE-Advanced to address the capacity requirement due to the increasing mobile data traffic dominated by multimedia services (e.g. video downloading, live streaming, etc) [2]. The multicast features incorporated in eMBMS offer high resource utilization efficiency by allowing the same multimedia content to be delivered simultaneously to multiple users on shared allocated resources. Nevertheless, the demand for high quality video and the emergence of enhanced content application (e.g. Virtual Reality (VR), augmented reality (AR), etc) poses greater challenges for the current eMBMS system to cope with the capacity requirements. The single-rate multicast system used in current eMBMS is associated with low data rate since it is limited by the user with the least channel gain. Furthermore, this technique is unfair towards the users with better channel quality as they potentially can achieve higher data rate with unicast system. This problem is addressed by implementing layered multicast system which allows each user to receive different video quality depending on their channel quality. Moreover, layered multicast system potentially achieves higher spectral efficiency compared to conventional single-rate system [3], [4]. Therefore, the layered multicast video streaming is one of the promising features for next generation mobile networks [3], [5].

Another candidate solution which caters the capacity requirement for future mobile network is non-orthogonal multiple access (NOMA). NOMA offers enhanced spectral efficiency, improved cell-edge transmission rates, low latency and fairness towards all users [6]–[8]. Due to these benefits, several works on the application of NOMA in multicast systems have emerged in literature. In particular, the system design and coverage performance of NOMA-based broadcasting/multicasting in cellular networks was studied in [9]. Moreover, NOMA was applied to multicast system in wireless content caching networks in [10] to reduce the bandwidth usage for both multicasting and content pushing, where resource allocation was performed by jointly optimizing the power allocation and content matching. User assignment and discrete power control were jointly considered in [11] for scalable NOMA multicast system. In [12], a two-stage beamforming method was designed to support multiple multicast groups in which the beamforming vector and power were jointly optimized to achieve minimum transmission power and target rates. Sophisticated beamforming was also applied

in [13] to allow the delivery of multiple multicast streams to different multicast groups. These streams are multiplexed with common broadcast information using NOMA with fixed power allocation. Meanwhile, the authors in [14] considered the application of NOMA in content-centric multicast delivery in cloud radio access networks (C-RANs). In that work, the beamforming vectors and the quantization noise covariance matrix were jointly optimized to maximize the minimum delivery rate of contents considering finite-capacity fronthaul links. By applying fixed power allocation, the works in [15] and [16] investigated the NOMA performance in terms of signal-to-interference and noise ratio (SINR) coverage probability, average number of served users and sum rate for NOMA-based multicast systems in cooperative millimeter-wave and heterogeneous networks, respectively. Power allocation problems for relay assisted NOMA-based multicast system for vehicle-to-everything (V2X) networks were investigated in [17]. On the other hand, subgrouping in multicasting is a promising technique which enhances network performance by exploiting multi-user diversity [3]. User subgrouping for two multicast groups was studied in [18] to maximize the spectral efficiency of NOMA multi-service scheme. Other works considered the application of NOMA in mixed unicast-multicast transmission. For example, in [19], a joint beamforming and power allocation scheme was designed considering spectral efficiency and security aspects. The work in [20] solely focuses on the design of power allocation schemes. On the other hand, energy-efficient resource allocation was investigated in [21] for massive MIMO systems with simultaneous wireless information and power transfer (SWIPT). Meanwhile, [22] and [23] considered cooperative communications for multicast NOMA to enhance users' reliability.

Resource allocation has been recognized as a key technique to maximize a variety of network performance including sum rate maximization. A significant amount of literature exist on resource allocation schemes for multicast systems in orthogonal frequency division multiple access (OFDMA) networks as highlighted in [4]. In particular, the multi-rate technique has gained much interest in literature due to its potential in achieving user rate differentiation and higher spectral efficiency [3], [4]. Meanwhile, the transmission rate of conventional multicast systems is often restricted by the user with worst channel gain in a multicast group to ensure successful detection of content by all users. By exploiting layered coding such as Multiple Description Coding (MDC), Fine Granularity Scalable (FGS) coding and Scalable Video Coding (SVC), multi-rate multicast allows the video content to be encoded into multiple streams of different transmission rates. Cell-edge users only receive the high-priority basic-quality video stream while cell-center users with higher channel gains receive high quality video as more enhancement layer streams (including the high-priority stream) can be successfully decoded. However, additional frequency/time resources are required to deliver the multiple layer streams in an orthogonal multiple access-based network. This issue can be addressed by applying NOMA which allows the utilization of whole bandwidth allocated to a single multicast group by multiplexing all the layer streams in the power domain. Moreover, resource allocation in an OFDMA-based multi-layer system involves the assignment of each subcarrier or resource block (RB) to the layer streams, which is not required in NOMA. Therefore, in a multi-layer multicast network, resource optimization for NOMA is less complex compared to that of OFDMA.

Applying NOMA to video multicasting with layered coding, the weakest users' signal can be the base layer stream (high-priority basic-quality video data) while stronger users are sent the enhancement layer streams (additional data to improve video quality). Instead of discarding the weaker user's signals as in conventional NOMA, the stronger users will utilize them to enhance the overall sum multicast rate. Motivated by these benefits of exploiting the nature of NOMA and layered coding, we investigate joint power allocation and subgrouping for multi-layer multicast streaming in NOMA networks, which is crucial to further enhance the system performance, particularly the overall throughput. The optimization problem is formulated with the aim of maximizing the sum multicast rate performance while satisfying the maximum transmission power and proportional rate constraints. We consider the proportional rate constraint in our work to guarantee substantial rates and fairness for all the layer streams while maximizing the sum multicast rates. In addition, this constraint also ensures that the high-priority base layer stream (i.e., the weakest user's signal) is successfully decoded by all users by allocating sufficient power to this stream [24].

There are not many literature that consider optimizing the subgrouping in NOMA-based layered multicast system, which is different from the pairing/clustering problem in unicast NOMA system. For example, the block-level utility maximization and distortion minimization problems in [25] and [26] respectively are solved without optimizing the subgrouping. To the best of our knowledge, [27] and [28] are the only works which also consider optimizing the subgrouping. However, [28] focused only on a single channel and two-layer case, and therefore it is easier to obtain the closed form solution. On the contrary, our work considers multiple layers and multiple RBs case, which is difficult to solve as more variables are optimized. Similar to our work, [27] also considered both multiple RBs and multiple layers. However, the closed-form power allocation in [27] was derived to exactly meet the target rate of lower layers and allocate the remaining power to the uppermost layer. This strategy does not necessarily maximize the sum multicast rate. On the other hand, our work considers maximizing the sum multicast rate with proportional rate constraint, which is a more complicated optimization problem. Furthermore, the subgrouping technique used in [27] is based on exhaustive search which is associated with high computational complexity. In our work, we have developed low-complexity subgrouping methods which are suitable for practical implementation. Meanwhile, the work in [29] investigated a complex non-convex discrete optimization problem based on quality of experience (QoE) and

statistical channels, but subgrouping was not optimized. Other works such as [30]–[32], considered MISO case and focused on the optimization of the beamforming weights with the aim of minimizing the transmission power. Meanwhile, [33] and [34] incorporate cooperative strategy to allow the weak users to receive the enhancement layer streams, but resource allocation was not investigated. The main contributions of this paper are summarized as follows

- The joint optimization of power allocation and subgrouping is a mixed integer non-linear programming (MINLP) problem which is generally difficult to solve and requires computationally-intensive numerical solution. Therefore, we first consider solving the problem for 2-layer case with arbitrary subgrouping which is a non-linear programming (NLP). Sum multicast rate maximization problem with proportional rate constraint can be solved using Lagrangian dual decomposition (LDD) technique. Based on LDD, we derive two sub-optimal power allocation schemes: an iterative subgradient technique and a closed-form solution. Simulation results show that the subgradient method offers better performance compared to the closed-form solution at the expense of much higher complexity. On the other hand, the closed-form solution is more suitable for practical implementation due to its low complexity. In addition, it is shown that the proposed schemes outperform all existing low-complexity schemes in terms of sum multicast rate performance and fairness towards the users with poor channel gains.
- In order to solve the joint optimization problem at low-complexity, we split the problem into two subproblems. The power allocation subproblem for a general case with multiple layers is solved by modifying the 2-layer based closed-form solution into a low-complexity power allocation scheme which successively allocates power for each layer. Despite the modification of the solution, the proposed scheme maintains its property of achieving the minimum target rate and proportionality, particularly, in determining the transmission rate for the high-priority base layer stream. This scheme offers robust delivery of basic quality video to all users by ensuring a high transmission rate for the base layer stream, which is vital in multicast video streaming. Moreover, this scheme offers superior sum multicast rate performance over [27].
- Finally, we propose three low-complexity iterative subgrouping methods in order to solve the subgrouping subproblem and then incorporate the multi-layer power allocation scheme to develop a technique which jointly optimizes the power allocation and subgroup formation. Simulation results demonstrate that the proposed low-complexity subgrouping methods offer performance comparable to that of exhaustive search.

The remainder of this paper is organized as follows. Section II describes the system model of the NOMA-based multi-layer multicast system and the formulation of the optimization problem. Section III presents the power allocation scheme for the 2-layer case which considers arbitrary subgrouping. Section IV details the joint power allocation and subgroup formation technique for the general multi-layer case. The simulation results are presented in Section V and, finally, this paper is concluded in Section VI.

II. System Model And Problem Formulation

A downlink NOMA-based multicast system, which consists of a single-antenna[1] base station (BS) and K users who request the same video content[2], is considered and is illustrated in Fig. 1. The BS delivers the video data to all users with a maximum transmit power budget of P_T over a total bandwidth of B_T which is divided into N RBs of equal bandwidth B. The bandwidth B is assumed to be smaller than the coherent bandwidth such that each individual RB will experience frequency flat fading. Channel impairments due to path loss, shadowing, and fading are also taken into account and hence, the channel gain[3] for the k^{th} user at the n^{th} RB is expressed as $|h_{k,n}|^2 = \frac{\xi_k |H_{k,n}|^2}{PL_k}$ where ξ_k is the log-normal shadowing factor for user k, $|H_{k,n}|^2$ is the effect of fading, and PL_k is the path loss effect experienced by user k. Let the set of all users in the multicast group at the n^{th} RB be $\mathcal{M}_n = \{1, 2, ..., K\}$ sorted in an ascending order of their channel gains[4]. These users are divided into a number of subgroups that optimizes the sum multicast rate.

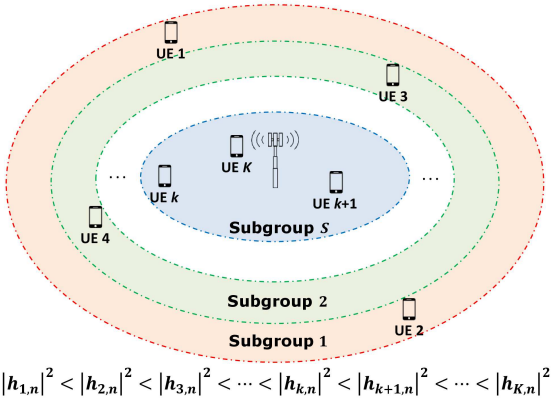


Fig. 1. System model.

[1]This work can be extended to multiple-antenna BS case (i.e. MISO) by applying the beamforming approach similar to [30] and [31].

[2]In order to focus on power allocation and subgrouping problems, we assume that there are no other users requesting for different video contents in this work. In practice, users in the same cell requesting other multimedia contents can be served by allocating different subbands or RBs. The allocation of subbands or RBs to different group of users can be treated as an independent optimization problem which has been dealt with extensively in the literature on [4].

[3]Note that, in this work, we assume that the BS obtains perfect channel state information (CSI) for all users. In practice, channel estimation will be needed and the estimation error will degrade the performance. Channel estimation for multicast system is also challenging due to multiple users in the cell. Nonetheless, the amount of performance degradation will be similar to the existing and proposed schemes considered in this work. The lack of accurate CSI or the use of statistical channels such as in [29] can be considered as a future work.

[4]Note that sorting the users according to channel gains for each RB is a common practice in power allocation for NOMA [24], [27]. This allows the users to be sorted according to the optimal SIC decoding order for each RB.

By applying layered coding, the BS first encodes the video content into S layers of data streams. Layer 1 is the basic quality video layer served to all users. Each subsequent layer is then intended for all users in the corresponding and higher ranked subgroups (i.e., the data in the s^{th} layer is for the users in subgroups s to S). We define the set of users and the total number of users in subgroup s at the n^{th} RB as $\mathcal{G}_n^{(s)}$ and $G_n^{(s)}$ respectively. The first (base) layer data stream $X^{(1)}$ consists of the most important data elements for basic quality video content and therefore must be successfully decoded in order to obtain the desired content. The enhancement layer data streams $X_n^{(s)}$ (for $s \in \{2, ..., S\}$) improve the video quality if successfully decoded by the users. All the layer streams are then multiplexed in the power domain by employing superposition coding (SC) and the received signal at user k at RB n is represented by

$$Y_{k,n} = h_{k,n} \sum_{s=1}^{S} \sqrt{P_n^{(s)}} X_n^{(s)} + W_{k,n} \tag{1}$$

where $P_n^{(s)}$ is the power allocated to layer stream s at RB n and $W_{k,n}$ is the additive white Gaussian noise (AWGN). It is noted from (1) that all the users will receive the superposed signal which consists of the base layer and all the remaining enhancement layer data streams. However, not all the users are guaranteed to successfully detect all the layers due to their channel conditions. Notably, the users in subgroup 1 directly decode the base layer stream by treating the remaining enhancement layers as noise. This is only possible if the power allocated to the base layer stream is sufficiently larger compared to the combined power of the enhancement layer streams. However, these users will not be able to decode the enhancement layer streams through the successive interference cancellation (SIC) technique due to the poor channel qualities and lower power allocated to the enhancement layer streams. On the other hand, the users in other subgroups (for $s \in \{2, ..., S\}$) will be able to extract the intended enhancement layer streams from the received signal through the cancellation of the base layer and lower-level enhancement layers components (for $s \in \{1, ..., s-1\}$). Instead of discarding the weaker users' streams (for $s \in \{1, ..., s-1\}$) as in conventional unicast NOMA, these signal streams are combined together to improve the quality of the video content. Consequently, the sum multicast rate performance is enhanced due to the combined transmission rates of all the layered streams. In conventional layered video streaming, the enhancement layers should be dropped when the base layer is corrupted or lost during transmission which leads to the inefficient utilization of bandwidth and power [35]. Therefore, it is necessary to guarantee the successful detection of base layer first before decoding the enhancement layers. This detection order complies with the conventional SIC decoding scheme in NOMA as the base layer (being the weakest user's signal) should be successfully detected first prior to the detection of enhancement layer through SIC. Fig. 2 shows the SIC technique for the users in each subgroup in a 4-layer NOMA-based multicast system.

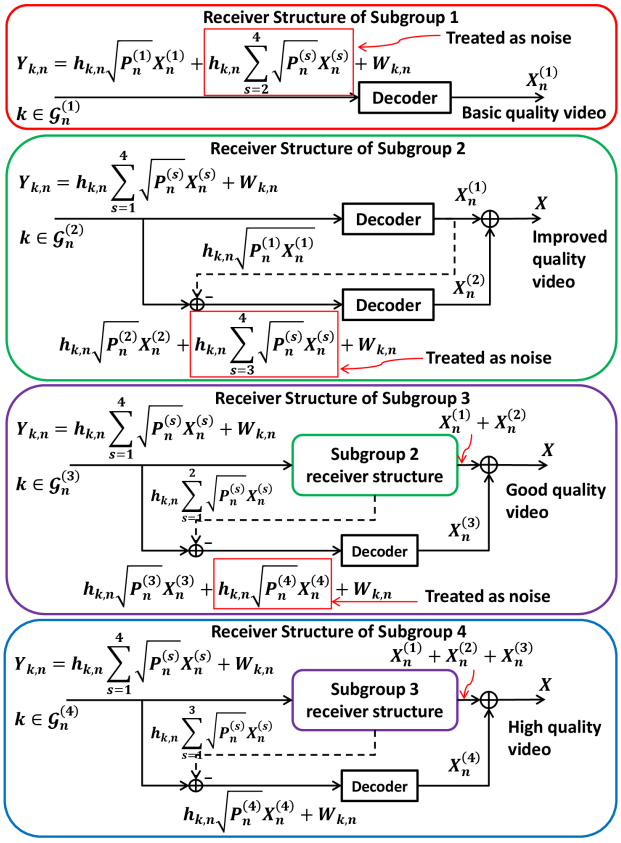


Fig. 2. SIC in multicast users' receivers for 4-layer multicast video streaming.

Considering that the users in the s^{th} subgroup successfully decode the lower layers (for $s \in \{1, ..., s-1\}$), the achievable rates for the s^{th} and S^{th} layers at RB n are given respectively by

$$R_n^{(s)} = B \log_2 \left(1 + \frac{P_n^{(s)} \gamma_n^{(s)}}{\gamma_n^{(s)} \sum_{i=s+1}^{S} P_n^{(i)} + BN_0} \right) \tag{2}$$

$$R_n^{(S)} = B \log_2 \left(1 + \frac{P_n^{(S)} \gamma_n^{(S)}}{BN_0} \right) \tag{3}$$

where N_0 is the noise power spectral density and $\gamma_n^{(s)}$ is the channel gain of the weakest user in subgroup s due to multicasting, which is given by $\gamma_n^{(s)} = \min |h_{k,n}|^2$ in which $k \in \mathcal{G}_n^{(s)}$. First of all, this allows the weaker users to receive better standard quality video content. Secondly, in this multicast system, a higher rate for the base layer will benefit all multicast users. Moreover, if the base layer is not correctly detected, the higher layer data will be useless due to layered coding and made worst by error propagation in SIC. Therefore, unlike conventional NOMA, the power allocation solution here must offer higher power for the base layer (i.e., weak user signal). The commonly used minimum rate constraint will allocate sufficient power for the weak user to achieve the minimum rate, and allocate as much power as possible for the strong user to maximize the sum multicast rate. In this work, the proportional rate constraint is used as it will proportionally allocate power

to the various users after the minimum rate is achieved [24].

Supposing that all multicast users K successfully decode the base layer streams and only the members of subgroup $s \in \{2, ..., S\}$ are able to decode the intended enhancement layer data streams, the sum multicast rate[5] of the NOMA multicast system can be mathematically represented as

$$R_{sum} = \sum_{n=1}^{N} \left[\left(\sum_{s=1}^{S-1} K_n^{(s)} R_n^{(s)} \right) + K_n^{(S)} R_n^{(S)} \right] \tag{4}$$

where $K_n^{(s)}$ is the total number of users who successfully decode layer stream s at RB *n,* which directly relates to the total number of users in subgroup s by the expression $K_n^{(s)} = \sum_{i=s}^{S} G_n^{(i)}$. It is worth mentioning that subgrouping influence the value of $K_n^{(s)}$ which affects the considered channel gain $\gamma_n^{(s)}$ in (2) and (3), and hence the transmission rate $R_n^{(s)}$. Therefore, the sum multicast rate can be optimized by appropriately choosing the value of $K_n^{(s)}$ through subgrouping while allocating power to each layer streams.

From (2), it could be noted that the transmission rate is limited by the weakest channel gain of each respective subgroup and the intereference caused by the upper-level enhancement layers. In particular, the rate of the first (base) layer is restricted by the weakest channel gain among all the users as well as the interference from all the enhancement layers. However, in multi-layered video coding, it is crucial to maintain a considerably high achievable rate for the base layer stream in order to offer robust delivery of the essential elements of the video content for all users. Hence, the optimization problem here is aimed towards maximizing the sum multicast rate performance of the NOMA-based multi-layer multicast system by designing a joint power allocation and subgrouping scheme which can satisfy both total transmission power and minimum rate proportional fairness constraints. The joint power allocation and subgrouping optimization problem is formulated as

$$\underset{P_n^{(s)}, K_n^{(s)}}{\text{maximize}} \quad R_{sum} \tag{5a}$$

$$\text{subject to} \quad \sum_{n=1}^{N} \sum_{s=1}^{S} P_n^{(s)} = P_T \tag{5b}$$

$$P_n^{(s)} \geq 0, \ \forall n, \ \forall s \tag{5c}$$

$$\sum_{n=1}^{N} R_n^{(1)} : \ldots : \sum_{n=1}^{N} R_n^{(S)} = \Phi_{min}^{(1)} : \ldots : \Phi_{min}^{(S)} \tag{5d}$$

$$K_n^{(s)} \in \{1, 2, ..., K\} \tag{5e}$$

$$K_n^{(1)} \geq K_n^{(2)} \geq \cdots \geq K_n^{(S)} \tag{5f}$$

where constraints (5b) and (5c) are the total power constraint and non-negative power constraint respectively, whereas, constraint (5d) is the proportional rate constraint that guarantees each layer stream s achieving the minimum target rates $\Phi_{min}^{(s)}$ whilst maintaining a proportionality among all the layer streams. Note that once the minimum rate requirements are met, the power resources are distributed among the layer streams in a proportional manner and therefore the obtained rates are not necessarily restricted to the target rates ratio [24]. The application of this constraint in NOMA ensures sufficiently higher power is allocated to the weaker users' signals to allow the successful detection of their own signals while treating the stronger users' signals as noise. Moreover, adequate power is also assigned to the stronger users to enhance the achievable rate performance. Constraint (5e) assures that $K_n^{(s)}$ can only take integer values. Finally, constraint (5f) implies that not all users can successfully detect the enhancement layer s (for $s \in \{2, ..., S\}$) and also ensures the expression $K_n^{(s)} = \sum_{i=s}^{S} G_n^{(i)}$ is complied.

III. Power Allocation Schemes for Two-layer with Arbitrary Subgrouping

The optimization problem in (5) is an MINLP problem which is generally difficult to solve and often associated with computationally-intensive numerical solution. Hence, we first simplify the problem in (5) by considering only two layer streams and arbitrarily group the users based on the order of their channel gains[6]. Based on (2), (3) and (4), the sum multicast rate is given by

$$R_{sum} = K^{(L)} \sum_{n=1}^{N} R_n^{(L)} + K^{(H)} \sum_{n=1}^{N} R_n^{(H)} \tag{6}$$

where $K^{(L)}$ and $K^{(H)}$ are the numbers of users who can decode the base layer and enhancement layer respectively. Note that the superscripts (L) and (H) replace the number-based layer/subgroup index to highlight that the base layer (equivalent to layer 1) and enhancement layer (equivalent to layer 2) streams are always being treated as weak and strong users' signals respectively. In addition, the users are arbitrarily grouped in such a manner that the $G^{(L)}$ weakest users belong to subgroup L and the remaining $G^{(H)}$ strong users join subgroup H. It is assumed that all the users will decode the base layer stream and therefore $K^{(L)}$ always equals to the total number of users, that is, $K^{(L)} = G^{(L)} + G^{(H)} = K$. The users in subgroup H can also decode the enhancement layer stream; hence $K^{(H)} = G^{(H)}$. The achievable rates for the base layer and enhancement layer at RB n are represented respectively as

$$R_n^{(L)} = B \log_2 \left(1 + \frac{P_n^{(L)} \gamma_n^{(L)}}{P_n^{(H)} \gamma_n^{(L)} + B N_0} \right) \tag{7}$$

[5]Throughout this paper, the sum multicast rate is defined as the total data rates of all users who succesfully decode the base layer streams, where it is important to note that not all of them are able to decode the enhancement layer streams [15]. In other words, only some users are able to decode base and enhancement layer streams while the remaining users can only decode the base layer to retrieve basic quality video. The sum multicast rate used in this paper should not be confused with the conventional sum rate used in communications, that is defined as $R_{sum} = \sum_{n=1}^{N} \left[\left(\sum_{s=1}^{S-1} R_n^{(s)} \right) + R_n^{(S)} \right]$.

[6]In arbitrary subgrouping, the users are first sorted in descending order of their channel gains. The number of users in each subgroup $G_n^{(s)}$ is arbitrarily specified. For example, in a 2-layer case ($S = 2$) where the number of users for each subgroup is arbitrarily chosen as $G_n^{(1)} = 2$ and $G_n^{(2)} = 3$, the users ranked 4th and 5th are selected for subgroup 1, and those ranked 1st to 3rd are in subgroup 2.

$$R_n^{(H)} = B\log_2\left(1+\frac{P_n^{(H)}\gamma_n^{(H)}}{BN_0}\right) \tag{8}$$

where $\gamma_n^{(L)}$ is the channel gain of the weakest user in the multicast group which is given by $\gamma_n^{(L)} = \min|h_{k,n}|^2$ in which $k \in \mathcal{K}_n$ and $\gamma_n^{(H)}$ is the smallest channel gain in subgroup H by considering $\gamma_n^{(H)} = \min|h_{l,n}|^2$ where $l \in \mathcal{G}_n^{(H)}$. Since the weakest channel gain is considered, the rate in (7) is achievable to all the users. Therefore, SIC can be successfully applied by the strong users in subgroup H to achieve the rate for enhancement layer in (8).

The optimization problem is transformed into a non-linear programming and is formulated as

$$\underset{P_n^{(L)},P_n^{(H)}}{\text{maximize}} \quad R_{sum} \tag{9a}$$

$$\text{subject to} \quad \sum_{n=1}^{N} P_n^{(L)} + \sum_{n=1}^{N} P_n^{(H)} \le P_T \tag{9b}$$

$$P_n^{(L)} \ge 0,\ P_n^{(H)} \ge 0,\ \forall n \tag{9c}$$

$$\sum_{n=1}^{N} R_n^{(L)} : \sum_{n=1}^{N} R_n^{(H)} = \Phi_{min}^{(L)} : \Phi_{min}^{(H)} \tag{9d}$$

Note that the constraints (5b) - (5d) in the original optimization problem are retained as these constraints are directly related to power allocation, whereas, the subgrouping-related constraints (5e) and (5f) are removed as arbitrary subgrouping is considered.

A. Subgradient Method

According to [36], the objective function for weighted sum rate maximization problem in NOMA system, which is similar to that of (6), is concave under some conditions in general multiple users/layers case. In the following, we proof specifically that our objective function is concave for the two-layer case.

Proposition 1: The objective function (6) *for the two-layer case is concave without any conditions.*

Proof: See Appendix A. ■

Although (6) is proven to be concave, a high-complexity numerical tool is required to obtain the optimal power allocation [37], [38]. Therefore, we first propose a near-optimal iterative subgradient method to determine the power allocated to both the base and enhancement layer in all RBs with lower computational complexity. The power allocation can be solved using the LDD technique [38], [39]. By applying the LDD technique, the power allocation problem can be solved using the Karush-Kuhn-Tucker (KKT) conditions as detailed in Appendix B and are expressed in terms of dual Lagrange multipliers μ and τ as follows

$$P_n^{(L)} = \left[\left(K^{(L)}\Phi_{min}^{(L)} - \tau\right)\left(\frac{1}{\mu\Phi_{min}^{(L)}\ln 2} - \frac{BN_0\Phi_{min}^{(H)}\left(\gamma_n^{(H)}-\gamma_n^{(L)}\right)}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(L)}\gamma_n^{(H)}}\right)\right]^+ \tag{10}$$

$$P_n^{(H)} = \left[\frac{BN_0\left(K^{(H)}\Phi_{min}^{(H)}+\tau\right)\Phi_{min}^{(L)}}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(L)}} - \frac{BN_0\left(K^{(L)}\Phi_{min}^{(L)}-\tau\right)\Phi_{min}^{(H)}}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(H)}}\right]^+ \tag{11}$$

where $[x]^+ = \max(x,0)$ which ensures constraint (9c) is satisfied. The dual variables μ and τ are then solved by using the iterative subgradient algorithm as summarized in Algorithm 1. In the first iteration ($t = 0$), the power allocated to both the base and enhancement layers is calculated using specified initial values of the Lagrange multipliers $\mu(0)$ and $\tau(0)$. In each iteration t, the dual variables are updated in steps 4 and 5 respectively by taking into account the dual variables in the previous iteration, the positive step sizes for μ and τ (denoted as α_μ and α_τ respectively), and the valid subgradients which are derived according to [40] as

$$\nabla\mu = \sum_{n=1}^{N} P_n^{(L)} + \sum_{n=1}^{N} P_n^{(H)} - P_T \tag{12}$$

$$\nabla\tau = \frac{\sum_{n=1}^{N} B\log_2\left(1+\frac{P_n^{(L)}\gamma_n^{(L)}}{P_n^{(H)}\gamma_n^{(L)}+BN_0}\right)}{\Phi_{min}^{(L)}} - \frac{\sum_{n=1}^{N} B\log_2\left(1+\frac{P_n^{(H)}\gamma_n^{(H)}}{BN_0}\right)}{\Phi_{min}^{(H)}}. \tag{13}$$

Algorithm 1 Subgradient Algorithm

1: Initialization: set $t = 0$ and ϵ, initialize $\mu(0)$ and $\tau(0)$
2: **while** $|\mu(t-1)-\mu(t)| \ge \epsilon$ **or** $|\tau(t-1)-\tau(t)| \ge \epsilon$ **do**
3: solve $P_n^{(L)}$ and $P_n^{(H)}$ using (10) and (11) respectively
4: update $\mu(t+1) = [\mu(t) + \alpha_\mu\nabla\mu]^+$
5: update $\tau(t+1) = \tau(t) + \alpha_\tau\nabla\tau$
6: $t \leftarrow t+1$
7: **end while**
8: output the optimal solutions $P_n^{(L)*}$ and $P_n^{(H)*}$

Note that the initial values and the step sizes affect the convergence towards optimal solution. In this work, diminishing step size $\alpha = a/\sqrt{t}$ or $\alpha = a/(b+t)$ is applied to guarantee near-optimal solution [40] where a and b are fixed non-negative values. The iteration terminates when the dual variables μ and τ converge, that is, when the change in values between the dual variables in the current iteration and that in previous iteration is considerably small and achieve a tolerance value of ϵ.

It is worth pointing out that the local optimum obtained by subgradient method is also globally optimal if the optimization problem is convex [40]. However, if the problem is not feasible (i.e., not convex), it is not guaranteed that the local optimum determined by subgradient algorithm is the optimal solution. Subgradient method may trapped in a local optimum instead of global solution.

B. Multicast-based Equal RB Power Allocation (M-ERPA)

Despite the simplicity of subgradient method, the convergence towards the optimal solution may be slow and hence may involve a significant amount of iterations. Therefore, in order to further reduce the complexity of power allocation, we

attempt to derive a closed-form solution by assuming that the power is equally allocated to each RB [24] and transforming the KKT condition related to the proportional rate constraint to another form. The total power for each RB is determined by dividing the maximum power budget P_T by the total number of RBs N and hence, is shared between the base layer and enhancement layer streams in each RB as follows

$$P_n^{(L)} + P_n^{(H)} = \frac{P_T}{N} = P_{RB} \tag{14}$$

where P_{RB} is the power allocated to each RB. By applying this assumption in the process of solving the optimization problem (9) as detailed in Appendix C, the closed-form suboptimal power for the base layer and enhancement layers in each RB are respectively derived as (15) and (16) in the following page where the variables ψ_1, ψ_2, ψ_3 and ψ_4 are defined in Appendix C. It is worth mentioning that despite all RBs have the same total power, the power allocation ratio $P_n^{(L)}/P_n^{(H)}$ vary over all RBs depending on the considered channel gains $\gamma_n^{(L)}$ and $\gamma_n^{(H)}$ as observed in (15) and (16).

IV. Power Allocation And Subgroup Formation Schemes for Multi-Layer

The performance of the NOMA-based multicast system can be further enhanced by optimizing the subgrouping and multiplexing multiple (more than two) layer streams in the power domain. Multi-user diversity will be exploited in order to achieve user transmission rate differentiation which consequently enhances the overall throughput. Nevertheless, the optimal joint power allocation and subgroup formation of the multi-layer multicast scheme is often associated with excessive computational complexities. Therefore, we split the joint problem into two subproblems and focus initially on solving the power allocation subproblem for the multi-layer case.

A. Successive Layer-based Power Allocation (SLPA)

In order to solve the power allocation problem in (5) at a low complexity, we modify the 2-layer based sub-optimal closed-form solutions derived in Section III-B. The solutions in (15) and (16) directly allocate the power to layer 1 (base) and layer 2 (enhancement) for the 2-layer case. For the multi-layer case, the power for each layer is successively allocated starting with the lowest layer (layer 1) and hence referred to as the successive layer-based power allocation (SLPA) scheme. The SLPA scheme in a 4-layer multicast system is illustrated in Fig. 3.In the first stage, all the layers except for layer 1 (which is treated as layer L) are grouped together to form a single imaginary upper layer H. Based on the power budget of P_{RB}, the solutions in (15) and (16) are then applied to obtain the power allocated to layer 1 (base layer) $P_n^{(1)}$ and the imaginary upper layer P_n^{GA}. The latter is to be shared among all the upper layers[7]. As such, the channel gains of the base layer and the upper layers should not be used directly as $\gamma_n^{(L)}$ and $\gamma_n^{(H)}$. This is because if the imaginary upper layer has a higher channel gain, it will be allocated with less power, which will be insufficient to be shared among all these layers. In order to compensate for this problem, the considered channel gains for layer L (base layer) and layer H (group of enhancement layers) are respectively defined as

$$\gamma_n^{(L)} = \min_{q \in \Omega_n^{(L)}} \gamma_n^{(q)} \times W^{-1} \tag{17}$$

$$\gamma_n^{(H)} = \min_{r \in \Omega_n^{(H)}} \gamma_n^{(r)} \times W \tag{18}$$

where $\Omega_n^{(L)}$ and $\Omega_n^{(H)}$ are the set of layer(s) which form the lower layer and imaginary upper layer respectively, and W is a weight which is defined as $W = \frac{\sum_{r \in \Omega_n^{(H)}} \gamma_n^{(r)}}{\sum_{s=1}^{S} \gamma_n^{(s)}}$. This weight is a fraction of the combined channel gain in the imaginary upper layer to the total channel gain of all layers. Since W is always less than 1, it effectively lowers the considered channel gain for the imaginary upper layer and increases that for the base layer. This means the base layer will give up some of its power in order to compensate for the power to be shared among the enhancement layers. At each stage, the variables $K^{(L)}$ and $K^{(H)}$ in (15) and (16) are defined as $K^{(L)} = \max_{q \in \Omega_n^{(L)}} K^{(q)}$ and $K^{(H)} = \max_{r \in \Omega_n^{(H)}} K^{(r)}$. By starting with the lowest layer, a high data rate is guaranteed for the high-priority base layer since both allocated power $P_n^{(1)}$ and the interference power P_n^{GA} is already determined in the first stage. However, higher rate is not guaranteed for the upper layers since the remaining power P_n^{GA} is shared. Note that, in layered coding, it is useless to obtain upper layers with higher rates if the base layer is in outage. Therefore, this approach ensures robust delivery of basic quality video data to all users by maintaining high transmission rate for the base layer stream. This is critical in multilayer video coding as a higher rate for the base layer can ensure all users have a good basic video quality, while the enhancement layers improve the quality of service for users with better channel.

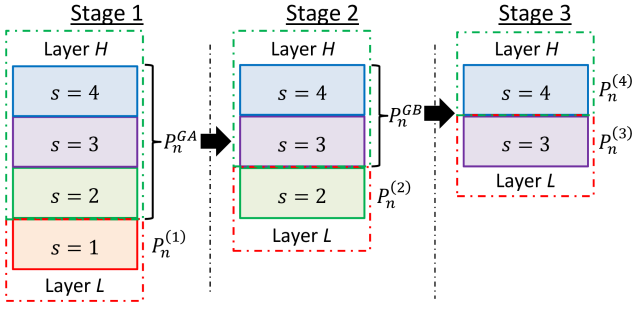


Fig. 3. Illustration of the power allocation stages in SLPA for 4-layer system.

In the second stage, layer 2 is being treated as the lower layer L while the remaining upper enhancement layers are grouped together as the imaginary upper layer H. The power allocated to layer 2 $P_n^{(2)}$ and to the newly-formed imaginary upper layer P_n^{GB} is obtained by applying (15) and (16) respectively using the power budget allocated in the previous stage (i.e.,

[7]Note that, from (2), the interference power is calculated as the sum of power allocated to all the remaining upper layers. Therefore, the interference of the base layer is P_n^{GA}. The rate of the base layer can be guaranteed in the first stage based on allocated power $P_n^{(1)}$ and interference P_n^{GA}.

$$P_n^{(L)} = P_{RB} + \frac{BN_0\Big(\big(\gamma_n^{(L)}\Phi_{min}^{(H)}+\gamma_n^{(H)}\Phi_{min}^{(L)}\big)\big(2\big(K^{(L)}-K^{(H)}\big)\psi_1+\psi_2+BN_0\psi_3\big)+2\big(K^{(H)}\gamma_n^{(H)}-K^{(L)}\gamma_n^{(L)}\big)\psi_4\Big)}{\gamma_n^{(L)}\gamma_n^{(H)}\Big(\big(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\big)\big(2\big(K^{(L)}-K^{(H)}\big)\psi_1+\psi_2+BN_0\psi_3\big)+2\big(K^{(H)}-K^{(L)}\big)\psi_4\Big)} \quad (15)$$

$$P_n^{(H)} = -\frac{BN_0\Big(\big(\gamma_n^{(L)}\Phi_{min}^{(H)}+\gamma_n^{(H)}\Phi_{min}^{(L)}\big)\big(2\big(K^{(L)}-K^{(H)}\big)\psi_1+\psi_2+BN_0\psi_3\big)+2\big(K^{(H)}\gamma_n^{(H)}-K^{(L)}\gamma_n^{(L)}\big)\psi_4\Big)}{\gamma_n^{(L)}\gamma_n^{(H)}\Big(\big(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\big)\big(2\big(K^{(L)}-K^{(H)}\big)\psi_1+\psi_2+BN_0\psi_3\big)+2\big(K^{(H)}-K^{(L)}\big)\psi_4\Big)}. \quad (16)$$

$P_{RB} = P_n^{GA}$). Here, the considered channel gains are also determined by (17) and (18) in order to ensure sufficient power to be shared among the upper enhancement layers. The same procedure is repeated in the subsequent stages until there are only two layers left for consideration in which (15) and (16) are directly applied, taking into account the power budget allocated in the preceding stage.

In the following section, we propose three low complexity iterative subgroup formation schemes and incorporate the proposed power allocation scheme to develop schemes that jointly determine the power allocation and subgrouping solutions.

B. Subgroup Formation

As defined in (4), the number of users receiving layer s at RB n, $K_n^{(s)}$, is an integer in which the possible values can be represented as a set $\mathcal{K}_n^{(s)}$. From (4), it can be observed that the layer transmission rate $R_n^{(s)}$ grows linearly with $K_n^{(s)}$. However, selecting a higher value for $K_n^{(s)}$ does not necessarily optimize the sum multicast rate performance as increasing $K_n^{(s)}$ implies that a user with very weak channel condition will be included into the same subgroup. The total rate for every user in the same subgroup is thus reduced, which could reduce the sum multicast rate. In other words, the optimization of $K_n^{(s)}$ depends on the weakest user in the subgroup which affect the transmission rate of layer stream s and hence the sum multicast rate.

We consider that all users successfully retrieve the video content by receiving the base layer data and therefore, in our work, the number of users receiving layer 1 ($K_n^{(1)}$) always equals to the total number of users (K). Consequently, the number of users receiving the remaining layers s (for $s \in \{2, ..., S\}$) must be obtained to maximize the sum multicast rate. Consider an example of a 3-layer NOMA-based multicast system with 5 users, the number of users receiving layer 1 is 5 and the possible number of users receiving layer 2 is represented by the set $\mathcal{K}_n^{(2)} \in \{1, 2, 3, 4\}$. If for instance user 2 and 3 are in subgroup 2, $\mathcal{K}_n^{(2)}$ will be 4 as user 4 and 5 (who has better channel gains) will also be able to receive this layer. In this case, only user 4 and 5 could be in subgroup 3 and thus the set of possible number of users in layer 3 will be $\mathcal{K}_n^{(3)} \in \{1, 2\}$. From this example, it can be observed that the search space of the set $\mathcal{K}_n^{(s)}$ is finite and hence, it is possible to optimize the subgrouping through an exhaustive search. However, this optimal search algorithm is computationally intensive particularly when the number of users (K), the number of RBs (N) and the number of layer streams (S) are large. Therefore, we propose three sub-optimal subgrouping schemes which offer much lower computational complexity in this section.

Algorithm 2 Subgroup Formation Method 1

1: **Initialize**: the set $\mathcal{K}_n^{(s)}$ and $\tilde{K}_n^{(s)}$ which is the j^{th} element of the set $\mathcal{K}_n^{(s)}$.
2: **for** $s = 2$ **to** S
3: Determine the sum multicast rates, $\mathcal{R}_n = \{R_n(j-1), R_n(j), R_n(j+1)\}$
4: **if** $\arg\max \mathcal{R}_n = j+1$
5: $t = j+1$
6: **while** $R_n(t) > R_n(t-1)$ & $t < K_n^{(s-1)*}$
7: $t = t+1$
8: Determine $R_n(t)$ and update $\mathcal{R}_n$
9: **endwhile**
10: update new $K_n^{(s)*} = \arg\max \mathcal{R}_n$
11: **elseif** $\arg\max \mathcal{R}_n = j-1$
12: $t = j-1$
13: **while** $R_n(t) > R_n(t+1)$ & $t > 1$
14: $t = t-1$
15: Determine $R_n(t)$ and update $\mathcal{R}_n$
16: **endwhile**
17: update new $K_n^{(s)*} = \arg\max \mathcal{R}_n$
18: **elseif** $\arg\max \mathcal{R}_n = j$
19: $K_n^{(s)*} = \tilde{K}_n^{(s)}$
20: **end if**
21: **end for**
22: **output**: the sub-optimal solutions $K_n^{(s)*}$ & $P_n^{(s)*}$

The main idea of subgrouping Method 1 is to examine whether removing or adding users into an arbitrarily-formed subgroup will increase the sum multicast rate. This method is summarized in Algorithm 2, which starts at subgroup 2 as subgroup 1 is the base layer stream for all users. Without loss of generality, we consider the subgrouping for the s^{th} subgroup, i.e., subgroup 2 to $s-1$ has already been assigned. First, the subgroup is arbitrarily formed by setting a channel-to-noise ratio (CNR) limit, Q_s. All users with a CNR between Q_{s-1} and Q_s will be admitted into this s^{th} subgroup, i.e., $\mathcal{G}_n^{(s)} = \big\{k : 1 \le k \le K,\, Q_{s-1} < {|h_{k,n}|^2}/{BN_0} < Q_s\big\}$. Let the number of users receiving layer s in this arbitrary subgroup be $\widetilde{K}_n^{(s)}$, and let it be the j^{th} element of the set $\mathcal{K}_n^{(s)}$. Then, the algorithm calculates the sum multicast rate according to this arbitrarily-formed subgroup, denoted as $R_n(j)$. The sum multicast rates for removing and adding a user in this subgroup (i.e., the sum multicast rates $R_n(j-1)$ and $R_n(j+1)$ according to the $(j-1)^{\text{th}}$ and $(j+1)^{\text{th}}$ elements in $\mathcal{K}_n^{(s)}$ respectively) are also calculated. Based on these calculated sum multicast rates, the trend of the objective function can be estimated in order to decide on the search direction. If the trend is increasing (i.e., adding a user increases the sum multicast rate), the algorithm will keep adding more users to the subgroup until there is no more improvement in sum multicast rate (i.e., a local/global maximum is met) or the maximum value of $K_n^{(s)}$

is reached (i.e., the last element in $\mathcal{K}_n^{(s)}$). On the contrary, the algorithm will keep removing users from the subgroup if a decreasing trend (i.e., eliminating a user increases the sum multicast rate) is observed at the start of the search. If the sum multicast rates $R_n(j-1)$ and $R_n(j+1)$ are smaller than $R_n(j)$ (i.e., a concave trend), the search is immediately stopped and the current arbitrarily-formed subgroup (i.e. j^{th} element) is considered to be the solution. Note that, in this algorithm, the sum multicast rates are calculated by applying the proposed power allocation scheme and thus, the power allocation is sub-optimized in all sum multicast rate evaluations.

Despite the reduced computational complexity, the solution for Method 1 may easily be trapped in a local maximum. Therefore, we improve this method by allowing the algorithm to search for the possibility of another local maximum. Upon reaching the first local maximum (represented as the $\tilde{j}^{\text{th}}$ element of $\mathcal{K}_n^{(s)}$), the sum multicast rate of the $\left(\tilde{j}-2\right)^{\text{th}}$ or $\left(\tilde{j}+2\right)^{\text{th}}$ element is examined and if found to be larger than that of $\left(\tilde{j}-1\right)^{\text{th}}$ or $\left(\tilde{j}+1\right)^{\text{th}}$ one, the algorithm resumes the iteration in search of the second local maximum. The best solution among the two local optima is then opted. This Method 2 outperforms Method 1 as it offers a higher possibility in finding the global optimum, but at the expense of higher complexity.

The final approach of lowering the computational complexity involves the reduction of the search space by eliminating some elements in $\mathcal{K}_n^{(s)}$. In a multicast system, the total sum multicast rate will be reduced when an additional user has a much lower channel gain is added into the subgroup. On the other hand, if the channel gain of the additional user is similar to others, including it in the subgroup will almost certainly increase the sum multicast rate. Therefore, if the channel gain for some users are similar, there is no need to compute the sum multicast rate for each additional user in the grouping assignment. Hence, to reduce the computational complexity, those users with similar channel gains can be considered together (i.e. they will be added or excluded together into a layer). Consider the example of 2-layer NOMA multicast system with 5 users. If the first two strongest users (i.e. user index 1 and 2 in $\mathcal{K}_n^{(2)}$) have similar channel quality, they can be grouped together and thus $K_n^{(2)} = 1$ can be removed from $\mathcal{K}_n^{(2)}$. These two users will either be selected together for a layer or none will be selected at all. Based on the new set $\mathcal{K}_n^{(s)}$, the solution is obtained by using Method 1. Note that this method (Method 3) potentially eliminates local optimum and therefore facilitate the search for the global optimal. However, there is also a risk of losing the global optimal due to the removal of similar elements from $\mathcal{K}_n^{(s)}$.

C. Complexity of Subgroup Formation Algorithm

For exhaustive search, the number of power allocation and sum multicast rate computations required is $N\frac{K!}{(K-(S-1))!}$ [27]. Meanwhile, Method 1 and 3 require $N(S-1)(3+I)$ computations where I is the total number of iterations. The additional three computations are accounted for the evaluation of the trends of the objective function at the start of the algorithm. Method 2 requires two rounds of search in order to improve the possibility of finding global maximum. Hence, Method 2 needs $N(S-1)(3+I_1+1+I_2)$ computations where I_1 and I_2 are the total number of iterations required for the first and second round of search respectively. Here, the first round of search is similar to Method 1 and the second search need only one more computation to examine the objective function's trend. The three proposed subgrouping methods only require small number of iterations I and thus offer much lower complexity as demonstrated in Section V.B.

V. Simulation And Results

We consider a downlink NOMA-based multicast network that consists of K users uniformly distributed within a circular cell radius of 500 m with a BS located at the centre. The effects of path loss, shadowing effect, noise and frequency selective fading are considered throughout all the simulations and the associated parameters are listed in Table I, unless stated otherwise.

TABLE I
Simulation Parameters

Paramaters	Values/Model
No. of RB, N	25
Total Bandwidth, B_T	5 MHz
Cell radius	500 m
Minimum distance from BS	10 m
Carrier frequency, f_c	2 GHz
Path loss, PL	$38.46+10v\log_{10}(\text{d})$
Path loss exponent, v	3
Log normal shadowing	σ=8dB
Noise power spectral density, N_0	-174 dBm/Hz
Frequency Selective Fading	ITU Pedestrian B
$\Phi_{min}^{(L)}$	0.5 Mbps
$\Phi_{min}^{(H)}$	1 Mbps

The performance of the proposed power allocation schemes are compared to other existing low-complexity NOMA power allocation schemes as well as the optimal scheme, which is solved using a numerical nonlinear optimization tool in MATLAB. For Fixed Power Allocation (FPA) scheme, the power allocation ratio of $0.8:0.2$ and $0.8:0.16:0.032:0.008$ are applied for 2-layer and 4-layer cases respectively. It should be noted that for the 4-layer FPA scheme, the sum of the power allocation ratio to the upper three layers is 0.2, which is identical to that of layer 2 in the 2-layer case. In other words, the power allocation ratio between layer 1 and the combined upper three layers is $0.8:0.2$. The power for each layer is successively determined by using this $0.8:0.2$ ratio in a manner similar to that of the proposed SLPA scheme. The main difference is that the power allocation ratio is always fixed regardless of the channel conditions of users. Thus, this FPA scheme potentially offer comparable performance to SLPA.

A. Two-Layer Case with Arbitrary Subgrouping

In this subsection, we consider the simulation of $K = 5$ multicast users who are arbitrarily grouped according to their

channel conditions such that the two weakest users belongs to subgroup 1 ($k \in \mathcal{G}_n^{(1)}$) and the remaining users join subgroup 2 ($k \in \mathcal{G}_n^{(2)}$). This arrangement ensures that more than 50% of the users will attain excellent quality of service. As listed in Table I, the minimum target rate[8] for base and enhancement layer streams are set as $\Phi_{min}^{(L)} = 0.5\,\text{Mbps}$ and $\Phi_{min}^{(H)} = 1.0\,\text{Mbps}$ respectively. This implies that the users in subgroup 2 will achieve a combined target rate of at least $1.5\,\text{Mbps}$. Fig. 4 presents the sum multicast rate performance of the proposed power allocation schemes (subgradient method and M-ERPA) which are compared to optimal NOMA, OFDMA and other NOMA power allocation schemes including FPA, Fractional Transmit Power Allocation (FTPA) [6] and Power Allocation for Intra-Group Scalable Multicast Scheduling (IGSMS-PA) [27]. The iterative subgradient method performs very closely to the optimal solution at the expense of higher complexity compared to other sub-optimal NOMA power allocation schemes. Nonetheless, this subgradient algorithm converges quickly (i.e., requires only around 10 iterations according to our simulations, and is not presented due to page limit) which shows that the complexity of this method is not too high. Meanwhile, the closed-form M-ERPA method performs better than other low-complexity schemes (FPA, FTPA and IGSMS-PA) since M-ERPA is derived based on a sum multicast rate maximization problem. In addition, the performance of NOMA is significantly better compared to optimal OFDMA, even with the application of sub-optimal power allocation schemes.

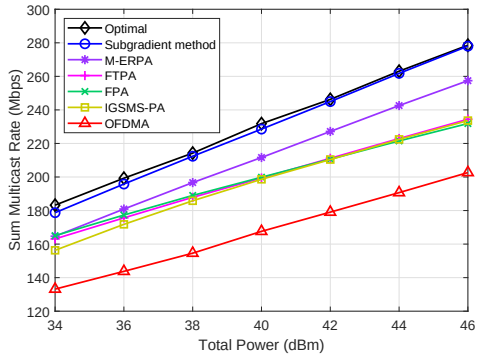


Fig. 4. Performance in terms of sum multicast rate versus total transmission power.

Fig. 5 illustrates the performance of the power allocation schemes in terms of the individual user's rates in each subgroup. Both the subgradient method and M-ERPA offer considerably higher transmission rates for the users in subgroup 1 at the expense of only small degradation to the users in subgroup 2. Consequently, the difference in individual user's rates between subgroup 1 and subgroup 2 is relatively smaller compared to that of other schemes, which is due to the proportional fairness feature in both proposed schemes. These results indicate that the proposed schemes do not only offer better sum multicast rate performance, but also yield a good degree of fairness. More importantly, this shows that the proportional rate constraint can achieve a higher rate for the base layer, which can benefit all users in this multicast multilayered video coding system.

[8]In our work, we arbitrarily chose the minimum target rate that will achieve a decent video quality for the base layer. In practice, the minimum rate requirement will be based on the minimum acceptable video quality for the base layer, which can be quite flexible. If in exceptional condition that the weakest user cannot achieve the minimum rate, that user can be excluded as outage user, and the algorithm will be applied to the remaining users.

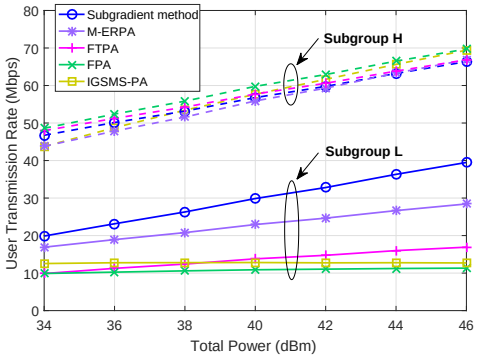


Fig. 5. Transmission rate of individual users in each subgroup versus total transmision power.

B. General Multi-Layer Case

For the general multi-layer case which considers joint optimization of power allocation and subgrouping, a significant performance gain can only be anticipated with larger number of users due to the heterogeneity of users' channel conditions. Therefore, we set $K = 20$ users in this simulation. Using SLPA, we first compare the sum multicast rate performance of the three proposed subgrouping methods with the optimal scheme and fixed arbitrary subgrouping for 2-layer ($S = 2$) and 4-layer ($S = 4$) cases as illustrated in Fig. 6. For the arbitrary subgrouping, the subgrouping configurations of $\{K_{1,n}, K_{2,n}\} = \{20, 10\}, \forall n$ and $\{K_{1,n}, K_{2,n}, K_{3,n}, K_{4,n}\} = \{20, 15, 10, 5\}, \forall n$ are implemented for 2-layer and 4-layer cases respectively. The results for both 2-layer and 4-layer cases show that significant performance gains are achieved through the implementation of the proposed subgrouping techniques over fixed arbitrary subgrouping. In addition, all the proposed subgrouping methods perform very closely to the optimal exhaustive search solution for both 2-layer and 4-layer cases. Method 1 is the worst among all the proposed schemes as it is easily trapped in local optima. Both Method 2 and 3 achieve better performance due to their capabilities in finding the global optimum through the search for second local maximum and the elimination of local optima respectively. Note that, for Method 3, the users are grouped together when the difference in CNR in dB is less than 5%, which is found to be a reasonable threshold for all maximum total power P_T scenarios. Fig. 7 shows the effect of selecting different CNR percentage differences for Method 3. The sum multicast rate performance improves as the percentage increases until around 5-6% as more local optima are possibly eliminated from the search space. Nevertheless, the performance starts to drop after 6% due to the elimination of global optimum and better local optima. More importantly,

the sum multicast rate at 5% difference is very close to the exhaustive search result, which shows the effectiveness of Method 3.

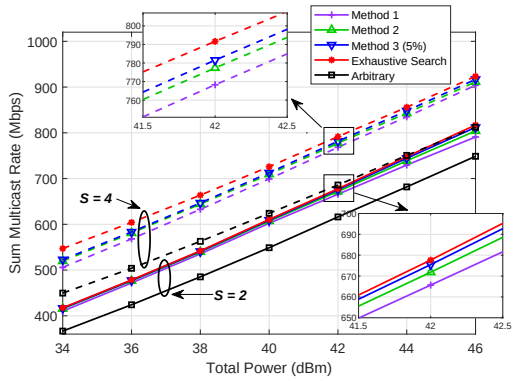


Fig. 6. Sum multicast rate performance of different subgrouping methods utilizing SLPA versus total transmission power for $S = 2$ and $S = 4$ cases.

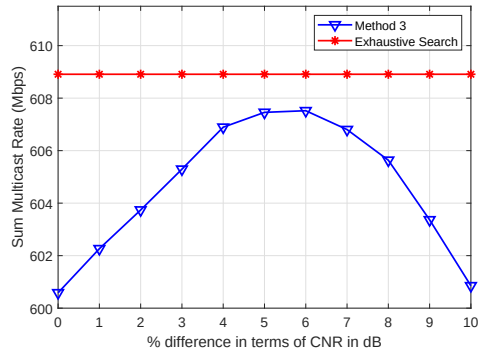


Fig. 7. The effect of varying the percentage difference in CNR (in dB) of Method 3 on sum multicast rate performance when $P_T = 40\,dBm$ for the $S = 2$ case.

Fig. 8 presents the complexity of the various proposed subgrouping schemes for the 2-layer ($S = 2$) and 4-layer ($S = 4$) cases, measured by the average number of power allocation and sum multicast rate computations. Method 2 offers good sum multicast rate performance at the expense of higher complexity due to its attempt in finding the second local maximum. The computational complexity of Method 1 and 3 is much lower as Method 1 is easily trapped in a local maximum and the search space in Method 3 is significantly reduced. It can also be observed that the complexity of the 4-layer case is approximately three times higher compared to the 2-layer case. This is because in the 4 layer case, the algorithm is required to determine the number of users receiving the remaining three layers $K_n^{(s)}$ (for $s \in \{2, 3, 4\}$) while only the number of users receiving the single enhancement layer $K_n^{(2)}$ is optimized in the 2-layer case. Note that the order of complexity for the proposed methods is almost consistent for all total transmission power P_T. The above results show that subgrouping Method 3 offers the best performance when both sum multicast rate and complexity are jointly considered, and therefore is implemented thereafter to evaluate the performance of different power allocation schemes. Moreover, in order to show results for a more general case, only 4-layer case is considered.

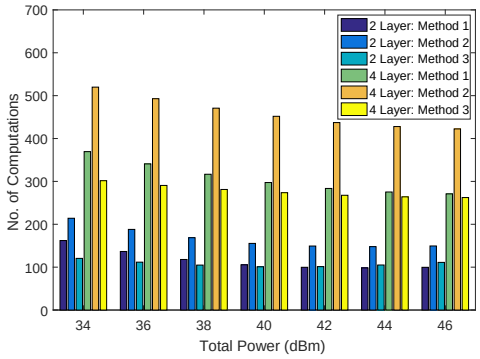


Fig. 8. Computational complexity of different subgrouping methods for $S = 2$ and $S = 4$ cases.

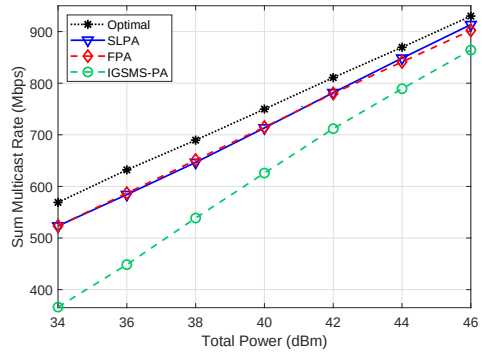
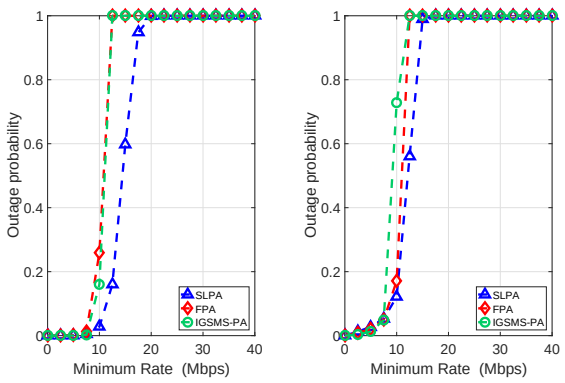


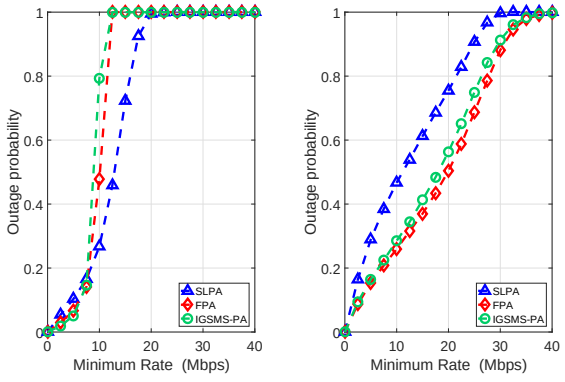
Fig. 9. Sum multicast rate performance of different power allocation schemes utilizing subgrouping Method 3 versus total transmission power for $S = 4$ case.

Using subgrouping Method 3 for the 4-layer case, we compare the sum multicast rate performance of the proposed SLPA scheme with the FPA, IGSMS-PA and the numerically optimized solution in Fig. 9. Both SLPA and FPA outperforms IGSMS-PA. When the transmission power is below $42\,dBm$, the performance of SLPA is comparable to that of FPA. However, at higher transmission power, the performance of SLPA is enhanced and is much closer to the optimal solution. This is because, in the multi-layer case, the allocation of higher power to the uppermost enhancement layer will incur a higher level of interference to the lower layers and therefore does not necessarily increase the overall rates. The proposed SLPA prioritizes the accomplishment of minimum target rates and proportional fairness which guarantees substantial rates for the lower layers, and thus potentially enhances the sum multicast rate performance. On the other hand, FPA scheme does not obligate to any minimum rate requirement and therefore the target rate for some layers may not be achieved, particularly for the high-priority base layer which is essential for the successful detection of video data. Meanwhile, IGSMS-PA allocates the power sufficiently to meet the minimum target rates for the lower layers and offer the remaining power to the uppermost layer. This strategy achieves a much lower sum multicast rate

as the lower layers are affected by higher interference in the uppermost layer.In multicast video streaming, it is vital to guarantee robust delivery of basic quality video to all the users for improved user experience.



(a) Layer 1/Base layer ($s = 1$) (b) Layer 2 ($s = 2$)



(c) Layer 3 ($s = 3$) (d) Layer 4 ($s = 4$)

Fig. 10. Outage probability for each layer stream s in $S = 4$ case against different target rate at $P_T = 46\,dBm$.

The outage probability plots for SLPA, FPA and IGSMS-PA schemes in the 4-layer case are presented in Fig. 10 to examine the robustness of the delivery of each layer stream when the total transmission power is fixed at $P_T = 46\,dBm$. As shown in Fig. 10(a), the outage probability performance for the base layer (layer 1) stream delivery in SLPA is superior compared to that of FPA and IGSMS-PA. This indicates that, in SLPA, the base layer stream can be delivered at higher target rates with lower outage which is essential for enhanced basic quality video. Moreover, the outage performances of SLPA for layer 2 and 3 are better than FPA and IGSMS-PA as indicated in Fig. 10(b) and 10(c). Nevertheless, the performance for the uppermost layer is degraded as shown in Fig. 10(d). However, it must be noted that if the base layer is in outage, receiving high rates for the upper enhancement layers are useless in layered video coding. Therefore, although FPA can have a comparable sum multicast rate performance to that of SLPA, the base video quality is poor and the users are often in outage. On the other hand, SLPA guarantees the robust delivery of lower layers, particularly for the base layer. As for IGSMS-PA, successful delivery of the layer streams are guaranteed by achieving the minimum target rates, albeit at a lower sum multicast rate.

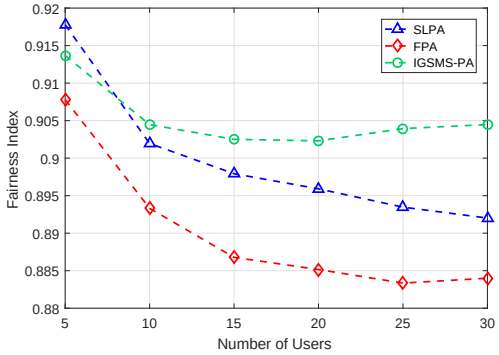


Fig. 11. Fairness index for $S = 4$ case against different number of users at $P_T = 46\,dBm$ with the target rates of $\Phi_{min}^{(1)} = 2.5$ Mbps, $\Phi_{min}^{(2)} = 5$ Mbps, $\Phi_{min}^{(3)} = 7.5$ Mbps, and $\Phi_{min}^{(4)} = 10$ Mbps.

Finally, the impact of the number of users on the fairness index of FPA, SLPA and IGSMS-PA for the 4-layer case is presented in Fig. 11. In this simulation, the layer streams are discarded if the target rate requirement of $\Phi_{min}^{(1)} = 2.5$ Mbps, $\Phi_{min}^{(2)} = 5$ Mbps, $\Phi_{min}^{(3)} = 7.5$ Mbps, and $\Phi_{min}^{(4)} = 10$ Mbps are not satisfied. Note also that, in layered coding technique, the decoding of a layer stream is highly dependent on the successful retrieval of lower layers. Hence, if a layer is in outage in this simulation, the layer and its upper layers are all discarded. Fig. 11 evaluates the performance of SLPA and FPA schemes in terms of Jain's fairness index which is defined in [41] as $J = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} R_k^2}$, where R_k is the transmission rate of each user k and the fairness index must be within the range $0 \leq J \leq 1$. From the Figure, it can be seen that SLPA achieves higher fairness compared to FPA due to the proportional allocation of resources to all the layers. Although IGSMS-PA achieves the highest user fairness, it has a much lower sum multicast rates because it only achieves the minimum rate for the lower layers. Note also that the fairness index decreases with the increase in number of users. This is because there are more possible combinations of subgrouping when the number of users are increased. Very often each subgroup will have a vastly different number of users, leading to reduced fairness. On the other hand, better fairness is achieved when considering less number of users since the users tend to be equally distributed among the subgroups.

VI. Conclusions

By exploiting the potential of NOMA and layered video coding, we investigated in this paper the joint power allocation and subgroup formation technique for multi-layer multicast video streaming. We developed two sub-optimal power allocation schemes for the 2-layer case considering arbitrary subgrouping: an iterative subgradient method and a closed-form M-ERPA technique. Furthermore, we modified the closed-form solution to derive a low-complexity SLPA scheme which successively allocates power to each layer in a general multi-layer case. This scheme was designed to maintain the property of achieving the minimum target rate and proportional fairness for the transmission rates of the base layer stream, which carries

the mandatory video content. The power allocation scheme was then embedded into three proposed subgrouping schemes in order to jointly optimize the power allocation and subgroup formation. For the 2-layer case with arbitrary subgrouping, simulation results show that both the subgradient method and M-ERPA offer superior performance over the existing schemes in terms of sum multicast rate and fairness towards the users with worst channel conditions. The subgradient method performs better than the closed-form M-ERPA, but with higher complexity. Simulation results for the multi-layer case have shown the significant performance gain achieved by optimizing the subgroup formation. Among all the proposed subgrouping techniques, Method 3 provides the best performance in terms of achieving higher sum multicast rates and offering lower computational complexity. Furthermore, in a general multi-layer case, the proposed SLPA scheme offers better performance compared to FPA and IGSMS-PA, and performs closely to the optimal solution at higher transmission power. In addition, SLPA offers robust delivery of mandatory video data to all users by maintaining substantial rates for the base layer stream. Therefore, the retrieval of basic quality video to all users are guaranteed. Since the quality of experience (QoE) of video delivery is highly influenced by the packet loss, the proposed solution will improve QoE through the robust delivery of the base layer (i.e., lower outage probability for the base layer). This also improves the QoE in terms of delay (e.g. buffering delay) as there is lower chance of retransmitting the lost packet. Moreover, the users with good channel conditions will be able to retrieve high quality video through the successful decoding of base and enhancement layers, and thus enhancing the QoE. The performance can be further enhanced by incorporating MISO system, which is our future work. In addition, investigating resource allocation problem in multi-cell systems, which deals with inter-cell interference issue, is an interesting future research.

APPENDIX A
CONCAVITY OF THE OBJECTIVE FUNCTION IN OPTIMIZATION PROBLEM (9)

Proof of Proposition 1: First, we consider the objective function for each RB which can be expressed as

$$f_n\left(P_n^{(L)},\, P_n^{(H)}\right) = K^{(L)} B \log_2\left(\left(P_n^{(L)}+P_n^{(H)}\right)\Gamma_n^{(L)}+1\right)+ K^{(H)} B \log_2\left(P_n^{(H)}\Gamma_n^{(H)}+1\right) - K^{(L)} B \log_2\left(P_n^{(H)}\Gamma_n^{(L)}+1\right) \tag{19}$$

where $\Gamma_n^{(L)} = \frac{\gamma_n^{(L)}}{BN_0}$ and $\Gamma_n^{(H)} = \frac{\gamma_n^{(H)}}{BN_0}$.

The Jacobian of $f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$ is calculated as

$$\nabla f_n\left(P_n^{(L)},\, P_n^{(H)}\right) = \begin{bmatrix} \frac{B}{\ln 2}\left(\frac{K^{(L)}\Gamma_n^{(L)}}{\left(P_n^{(L)}+P_n^{(H)}\right)\Gamma_n^{(L)}+1}\right) \\ \frac{B}{\ln 2}\left(\frac{K^{(L)}\Gamma_n^{(L)}}{\left(P_n^{(L)}+P_n^{(H)}\right)\Gamma_n^{(L)}+1} + \frac{K^{(H)}\Gamma_n^{(H)}}{P_n^{(H)}\Gamma_n^{(H)}+1} - \frac{K^{(L)}\Gamma_n^{(L)}}{P_n^{(H)}\Gamma_n^{(L)}+1}\right) \end{bmatrix}. \tag{20}$$

The Hessian of $f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$ is calculated as

$$\nabla^2 f_n\left(P_n^{(L)},\, P_n^{(H)}\right) = \begin{bmatrix} -\mathcal{A} & -\mathcal{A} \\ -\mathcal{A} & \mathcal{B}-\mathcal{A} \end{bmatrix} \tag{21}$$

where

$$\mathcal{A} = \frac{B}{\ln 2}\left(\frac{K^{(L)}\left(\Gamma_n^{(L)}\right)^2}{\left(\left(P_n^{(L)}+P_n^{(H)}\right)\Gamma_n^{(L)}+1\right)^2}\right) \tag{22}$$

$$\mathcal{B} = \frac{B}{\ln 2}\left(-\frac{K^{(H)}\left(\Gamma_n^{(H)}\right)^2}{\left(P_n^{(H)}\Gamma_n^{(H)}+1\right)^2} + \frac{K^{(L)}\left(\Gamma_n^{(L)}\right)^2}{\left(P_n^{(H)}\Gamma_n^{(L)}+1\right)^2}\right). \tag{23}$$

In order to prove that $f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$ is concave, the Hessian of $f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$ must be negative semidefinite [37], that is

$$\begin{bmatrix} P_n^{(L)} & P_n^{(H)} \end{bmatrix} \nabla^2 f_n\left(P_n^{(L)},\, P_n^{(H)}\right) \begin{bmatrix} P_n^{(L)} & P_n^{(H)} \end{bmatrix}^T \leq 0 \tag{24}$$

The condition (24) can be expressed as

$$-K^{(H)}\frac{\left(P_n^{(H)}\Gamma_n^{(H)}\right)^2}{\left(P_n^{(H)}\Gamma_n^{(H)}+1\right)^2} - K^{(L)}\left(P_n^{(L)}\Gamma_n^{(L)}\right)\times \left(\frac{\left(P_n^{(H)}\Gamma_n^{(L)}+2P_n^{(H)}\left(P_n^{(L)}+P_n^{(H)}\right)\left(\Gamma_n^{(L)}\right)^2+\left(P_n^{(L)}+P_n^{(H)}\right)\Gamma_n^{(L)}\right)}{\left(P_n^{(H)}\Gamma_n^{(L)}+1\right)^2\left(\left(P_n^{(L)}+P_n^{(H)}\right)\Gamma_n^{(L)}+1\right)^2}\right) \leq 0 \tag{25}$$

Since all the variables $K^{(L)}$, $K^{(H)}$, $\Gamma_n^{(L)}$, $\Gamma_n^{(H)}$, $P_n^{(L)}$, and $P_n^{(H)}$ are positive values, the condition in (25) always holds. This implies that $f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$ is concave. Note that the objective function in (9) is the summation of $f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$, i.e. $\sum_{n=1}^{N} f_n\left(P_n^{(L)},\, P_n^{(H)}\right)$. Since the concavity is retained for nonnegative weighted sum of concave functions [37], the objective function in (9) is concave.

APPENDIX B
DERIVATION OF POWER ALLOCATION EQUATIONS IN SUBGRADIENT METHOD

Based on the Lagrangian dual decomposition approach in [37], the Lagrangian function of the optimization problem (9), is expressed as

$$\begin{aligned}\mathcal{L}(P_n^{(L)}, P_n^{(H)}, \mu, \tau) =& K^{(L)}\sum_{n=1}^{N} B\log_2\left(1+\beta_n^{(L)}\right) \\ &+K^{(H)}\sum_{n=1}^{N} B\log_2\left(1+\beta_n^{(H)}\right) \\ &-\mu\left(\sum_{n=1}^{N}P_n^{(L)}+\sum_{n=1}^{N}P_n^{(H)}-P_T\right) \\ &-\tau\left(\frac{\sum_{n=1}^{N} B\log_2\left(1+\beta_n^{(L)}\right)}{\Phi_{min}^{(L)}} - \frac{\sum_{n=1}^{N} B\log_2\left(1+\beta_n^{(H)}\right)}{\Phi_{min}^{(H)}}\right)\end{aligned}$$

where μ and τ are the Lagrange multipliers, and $\beta_n^{(L)}$ and $\beta_n^{(H)}$ are respectively the SINR expressions of (7) and (8) which are defined as $\beta_n^{(L)} = \frac{P_n^{(L)}\gamma_n^{(L)}}{P_n^{(H)}\gamma_n^{(L)}+BN_0}$ and $\beta_n^{(H)} = \frac{P_n^{(H)}\gamma_n^{(H)}}{BN_0}$.

To solve the optimization problem (9), the KKT conditions are obtained as follows

$$\frac{d\mathcal{L}}{dP_n^{(L)}}=\sum_{n=1}^{N}\left[\left(K^{(L)}-\frac{\tau}{\Phi_{min}^{(L)}}\right)\frac{B}{\ln 2}\left(\frac{\beta_n^{(L)}}{P_n^{(L)}\left(1+\beta_n^{(L)}\right)}\right)\right]-\mu=0 \tag{26}$$

$$\frac{d\mathcal{L}}{dP_n^{(H)}}=\sum_{n=1}^{N}\left(K^{(L)}-\frac{\tau}{\Phi_{min}^{(L)}}\right)\frac{B}{\ln 2}\left(\frac{-\left(\beta_n^{(L)}\right)^2}{P_n^{(L)}\left(1+\beta_n^{(L)}\right)}\right)+\left(K^{(H)}+\frac{\tau}{\Phi_{min}^{(H)}}\right)\frac{B}{\ln 2}\left(\frac{\beta_n^{(H)}}{P_n^{(H)}\left(1+\beta_n^{(H)}\right)}\right)-\mu=0 \tag{27}$$

$$\mu\left(\sum_{n=1}^{N}P_n^{(L)}+\sum_{n=1}^{N}P_n^{(H)}-P_T\right)=0 \tag{28}$$

$$\tau\left(\frac{\sum_{n=1}^{N}B\log_2\left(1+\beta_n^{(L)}\right)}{\Phi_{min}^{(L)}}-\frac{\sum_{n=1}^{N}B\log_2\left(1+\beta_n^{(H)}\right)}{\Phi_{min}^{(H)}}\right)=0. \tag{29}$$

Solving the KKT conditions (26) and (27), we obtain the power allocated to the base layer and enhancement layer streams in terms of μ and τ as in (10) and (11) respectively. Then the problem (9) is transformed into a dual problem which is given by

$$\underset{\mu,\tau}{\text{maximize}} \quad D(\mu,\tau)=\inf_{P_n^{(L)},P_n^{(H)}}\mathcal{L}(P_n^{(L)},P_n^{(H)},\mu,\tau) \tag{30a}$$

$$\text{subject to} \quad \mu\geq 0. \tag{30b}$$

Finally, the Lagrange multipliers μ and τ are solved using the subgradient method.

Appendix C
Derivation of Multicast-based Equal RB Power Allocation (M-ERPA)

The M-ERPA scheme is derived from Lagrangian function and KKT conditions determined in Appendix B. By solving (26) and (27), the Lagrange variable μ is eliminated and $P_n^{(H)}$ can be expressed as

$$P_n^{(H)}=\frac{BN_0\left(K^{(H)}\Phi_{min}^{(H)}+\tau\right)\Phi_{min}^{(L)}}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(L)}}-\frac{BN_0\left(K^{(L)}\Phi_{min}^{(L)}-\tau\right)\Phi_{min}^{(H)}}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(H)}}. \tag{31}$$

It requires a high complexity numerical tool to solve $P_n^{(L)}$ using (28) and (31). The problem can be simplified by assuming that the power allocated to each RB are equal as represented in (14). By applying (14), the remaining power in each RB is allocated to the base layer stream and therefore, $P_n^{(L)}$ can be solved using (31) to give

$$P_n^{(L)}=P_{RB}-P_n^{(H)}=P_{RB}-\left(\frac{BN_0\left(K^{(H)}\Phi_{min}^{(H)}+\tau\right)\Phi_{min}^{(L)}}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(L)}}-\frac{BN_0\left(K^{(L)}\Phi_{min}^{(L)}-\tau\right)\Phi_{min}^{(H)}}{\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(K^{(L)}-K^{(H)}\right)-\tau\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\right)\gamma_n^{(H)}}\right). \tag{32}$$

The Lagrange variable τ in (31) and (32) can only be solved by substituting these equations into (29) using a high complexity iterative methods such as Newton Raphson method [38]. In order to obtain a closed form solution, (29) is transformed into a different form which is represented as

$$\tau\left(\sum_{n=1}^{N}B\frac{\Phi_{min}^{(H)}}{\Phi_{min}^{(L)}}+\sum_{n=1}^{N}B\log_2\left(1+\Gamma_n^{(L)}\right)-\sum_{n=1}^{N}B\log_2\left(1+\Gamma_n^{(H)}\right)\right)=0 \tag{33}$$

Note that, although (33) is not similar to (29), (33) still offers proportionality feature by adding the ratio between the minimum target rate of enhancement layer and that of base layer. By substituting (31) and (32) into (33), τ is obtained as

$$\tau=\frac{\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(2\left(K^{(L)}-K^{(H)}\right)\psi_1+\psi_2+BN_0\psi_3\right)}{2\psi_4} \tag{34}$$

where

$$\eta=2^{\left(\Phi_{min}^{(H)}/\Phi_{min}^{(L)}\right)} \tag{35}$$

$$\psi_1=\eta P_{RB}\left(\gamma_n^{(L)}\right)^2\gamma_n^{(H)}\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right) \tag{36}$$

$$\psi_2=\sqrt{BN_0}\left(\gamma_n^{(L)}-\gamma_n^{(H)}\right)\left(K^{(L)}\Phi_{min}^{(L)}+K^{(H)}\Phi_{min}^{(H)}\right)\times\sqrt{\frac{4\psi_1}{\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)}+BN_0\left(\left(\gamma_n^{(L)}-\gamma_n^{(H)}\right)^2+4\eta\gamma_n^{(L)}\gamma_n^{(H)}\right)} \tag{37}$$

$$\psi_3=\left(K^{(L)}\Phi_{min}^{(L)}-K^{(H)}\Phi_{min}^{(H)}\right)\left(\gamma_n^{(L)}-\gamma_n^{(H)}\right)^2+2\eta\gamma_n^{(L)}\gamma_n^{(H)}\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)\left(K^{(L)}-K^{(H)}\right) \tag{38}$$

$$\psi_4=\psi_1\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)+BN_0\left(\Phi_{min}^{(L)}\Phi_{min}^{(H)}\left(\gamma_n^{(L)}-\gamma_n^{(H)}\right)^2+\eta\gamma_n^{(L)}\gamma_n^{(H)}\left(\Phi_{min}^{(L)}+\Phi_{min}^{(H)}\right)^2\right) \tag{39}$$

Finally, the sub-optimal power for the base and enhancement layer streams can be solve by substituting (34) into (32) and (31) respectively to give the expressions in (15) and (16).

References

[1] M. N. Dani, Z. Q. Al-Abbasi, and D. K. C. So, "Power allocation for layered multicast video streaming in non-orthogonal multiple access system," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[2] M. Gruber and D. Zeller, "Multimedia broadcast multicast service: new transmission schemes and related challenges," *IEEE Communications Magazine*, vol. 49, no. 12, pp. 176–181, Dec. 2011.

[3] J. Montalban, P. Scopelliti, M. Fadda, E. Iradier, C. Desogus, P. Angueira, M. Murroni, and G. Araniti, "Multimedia multicast services in 5G networks: Subgrouping and non-orthogonal multiple access techniques," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 91–95, Mar. 2018.

[4] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 240–254, 2013.

[5] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Network*, vol. 31, no. 2, pp. 80–89, Mar. 2017.

[6] A. Benjebbour, A. Li, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 66–70.

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

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

[9] Z. Zhang, Z. Ma, X. Lei, M. Xiao, C. Wang, and P. Fan, "Power domain non-orthogonal transmission for cellular mobile broadcasting: Basic scheme, system design, and coverage performance," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 90–99, Apr. 2018.

[10] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2723–2735, Dec. 2017.

[11] R. H. Gau and H. T. Chiu, "Scalable NOMA multicast in cellular networks," in *Proc. IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sept. 2016, pp. 1–6.

[12] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[13] P. Henarejos, M. Shaat, and M. Navarro, "NOMA assisted joint broadcast and multicast transmission in 5G networks," in *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, Aug. 2017, pp. 420–425.

[14] S. R. Lee, S. H. Park, and I. Lee, "NOMA systems with content-centric multicast transmission for C-RAN," *IEEE Wireless Communications Letters*, pp. 828–831, Oct. 2018.

[15] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1794–1808, Aug. 2017.

[16] Z. Zhang, Z. Ma, M. Xiao, G. Liu, and P. Fan, "Modeling and analysis of non-orthogonal MBMS transmission in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2221–2237, Oct. 2017.

[17] G. Liu, Z. Wang, J. Hu, Z. Ding, and P. Fan, "Cooperative NOMA broadcasting/multicasting for low-latency and high-reliability 5G cellular V2X communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7828–7838, Oct. 2019.

[18] Y. Zhang, X. Wang, D. Wang, Y. Zhang, Q. Zhao, and Q. Deng, "NOMA-based cooperative opportunistic multicast transmission scheme for two multicast groups: Relay selection and performance analysis," *IEEE Access*, vol. 6, pp. 62 793–62 805, Oct. 2018.

[19] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 3151–3163, Jul. 2017.

[20] Z. Yang, J. A. Hussein, P. Xu, Z. Ding, and Y. Wu, "Power allocation study for non-orthogonal multiple access networks with multicast-unicast transmission," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3588–3599, Jun. 2018.

[21] F. Tan, P. Wu, Y. Wu, and M. Xia, "Energy-efficient non-orthogonal multicast and unicast transmission of cell-free massive MIMO systems with SWIPT," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2020.

[22] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[23] L. Yang, J. f, Q. Ni, J. Shi, and X. Xue, "NOMA-enabled cooperative unicast-multicast: Design and outage analysis," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7870–7889, Dec. 2017.

[24] Z. Q. Al-Abbasi and D. K. C. So, "Resource allocation in non-orthogonal and hybrid multiple access system with proportional rate constraint," *IEEE Transactions on Wireless Communication*, vol. 16, no. 10, pp. 6309–6320, Oct. 2017.

[25] H. Duan, Y. Zhang, and J. Song, "Block-level utility maximization for NOMA-based layered broadcasting," *IEEE Transactions on Broadcasting*, vol. 66, no. 1, pp. 21–33, Mar. 2020.

[26] J. Wu, B. Tan, J. Wu, and M. Wang, "Video multicast: Integrating scalability of soft video delivery systems into NOMA," *IEEE Wireless Communications Letters*, vol. 8, no. 6, pp. 1722–1726, Dec. 2019.

[27] H. Zhu, Y. Cao, T. Jiang, and Q. Zhang, "Scalable NOMA multicast for SVC streams in cellular networks," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6339–6352, Dec. 2018.

[28] M. Chen and S. Li, "Power allocation for NOMA based layered multicast transmission," in *Proc. IEEE 4th International Conference on Computer and Communications (ICCC)*, Dec. 2018, pp. 678–682.

[29] M. Zhang, H. Lu, F. Wu, and C. W. Chen, "NOMA-based scalable video multicast in mobile networks with statistical channels," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[30] C. Guo, Y. Cui, D. W. K. Ng, and Z. Liu, "Power-efficient multi-quality multicast beamforming based on SVC and superposition coding," in *Proc. IEEE Globecom*, Dec 2017, pp. 1–7.

[31] ——, "Multi-quality multicast beamforming based on scalable video coding," in *arXiv:1610.09530*, 2016.

[32] T. Li, H. Zhang, X. Zhou, and D. Yuan, "NOMA-enabled layered video multicast in wireless-powered relay systems," *IEEE Communications Letters*, vol. 23, no. 11, pp. 2118–2121, Nov. 2019.

[33] Y. Liu, Y. Liu, and G. Ma, "Hybrid decode-forward amplify-forward relaying with opportunistic layered multicast," in *Proc. 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct 2018, pp. 1–7.

[34] L. Yang, Q. Ni, L. Lv, J. Chen, X. Xue, H. Zhang, and H. Jiang, "Cooperative non-orthogonal layered multicast multiple access for heterogeneous networks," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1148–1165, Feb. 2019.

[35] Y. Huo, C. Hellge, T. Wiegand, and L. Hanzo, "A tutorial and review on inter-layer FEC coded layered video streaming," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 1166–1207, 2015.

[36] J. Wang, Q. Peng, Y. Huang, H. Wang, and X. You, "Convexity of weighted sum rate maximization in NOMA systems," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1323–1327, Sep. 2017.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[38] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[39] W. Yu, R. Lui, and R. Cendrillon, "Dual optimization methods for multiuser orthogonal frequency division multiplex systems," in *Proc. IEEE Globecom*, Nov. 2004, pp. 225–229.

[40] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[41] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in wireless networks:issues, measures and challenges," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 5–24, 2014.