QoE-Aware Resource Allocation for Non-Orthogonal Multiple Access Enhanced HetNets

Liangyu Chen, Bo Hu, *Member, IEEE,* Shanzhi Chen, *Fellow, IEEE,* Jianpeng Xu, and Guixian Xu

***Abstract*—Non-orthogonal multiple access (NOMA) has drawn significant attention due to its high spectral efficiency. Invoking NOMA in heterogeneous networks (HetNets) can support ubiquitous connectivity and satisfy the growing demand for mobile data traffic. Existing studies on NOMA-enhanced HetNets mainly focus on network's quality of service (QoS) metrics such as delay, throughput, coverage, etc. However, these parameters are not sufficient for evaluating the quality of experience (QoE) perceived by users. To that end, we propose a QoE-aware resource allocation framework for NOMA-enhanced HetNets under web browsing and video services. Specifically, a unified QoE-aware joint subchannel and power allocation optimization problem is formulated to maximize the sum mean opinion scores (MOSs) of all users, while guaranteeing the QoE requirement of each user. However, this problem is mixed-integer, non-convex, and intractable. To solve it, a penalty-based iterative algorithm is proposed. In particular, binary constraints on subchannel assignment variables are equivalently transformed into equality constraints via penalty method. Then, subchannel assignment and power allocation are alternately optimized in each iteration by leveraging block coordinate descent method and sequential parametric convex approximation techniques. Extensive numerical results show that the proposed scheme could achieve competitive QoE performance compared to existing NOMA and orthogonal multiple access schemes.**

***Index Terms*—QoE, NOMA, HetNets, resource allocation.**

I. Introduction

THE widespread usage of smart devices and services has resulted in huge surges of wireless data traffic. Statistical report from Cisco predicts that the monthly global mobile traffic in 2022 will increase by almost 7-fold compared to 2017, reaching 77 EB (1 EB = 10^{18} B) per month by 2022 [1]. To meet such extremely high data traffic in wireless cellular networks, heterogeneous network (HetNet) [2], [3] is considered one of the most appealing solutions for designing future wireless networks. In HetNet, the macro base station (MBS) provides network coverage, and dense short-range and lower-power small base stations (SBSs) are deployed within the macro cell to enhance system capacity. Recently, non-orthogonal multiple access (NOMA) [4]–[6], as a promising enhanced technology to further boost spectral efficiency and connectivity density, has been introduced into HetNets [7]–[13]. The key characteristic of NOMA is to enable more than one user to simultaneously access the same channel via the multiplexing in power domain or code domain, which is essentially different from conventional orthogonal multiple access (OMA) strategies in which the time/frequency/code resource unit is occupied by only one user. To be specific, power domain NOMA (PD-NOMA) supports multi-user transmission within the same time/frequency/code resource via distinguishing them with different power levels, whereas code domain NOMA (CD-NOMA) enables multiple transmissions within the same time-frequency resource through allocating different spreading sequences to different users [4]–[6], [14]. In this paper, we concentrate on the PD-NOMA in a downlink HetNet, in which the macro cell users (MUEs) and small cell users (SUEs) in each small cell can share the same radio resource by adopting various power levels. Moreover, with the aid of successive interference cancellation (SIC) technique at each NOMA receiver, the multi-user signal can be detected and thus the inter-user interference is managed.

On the other hand, a notable trend observed from Cisco report is that video streaming will occupy 79 percent of global traffic by 2022 [1]. With the rising popularity of Internet of Things (IoT), video service has been developed in new paradigms, such as high-definition video streaming, video-embedded web browsing, etc. Motivated by the demands of high-quality video applications, network operators have turned their attention toward the user-oriented Quality of Experience (QoE). ITU-T defines QoE as the overall acceptability of the services subjectively by end-user [15]. In order to support the users with interactive and real-time services in NOMA-enhanced HetNets, it is essential for the network and service providers to provide a high QoE for each service user. However, the new challenges in terms of co-channel interference caused by frequency reuse between the macro cell and small cells in NOMA HetNets may lead to negative effect on fluency and quality of video streaming, and thus affect the users' perceived quality. In addition, owing to the diversity of video characteristics, the different level of QoE may be experienced by the users even when those users have the same data rate, which means that efficient QoE-aware resource allocation

This work was supported by the National Key R&D Program of China under Grant 2020YFB1807900 and the National Natural Science Foundation of China (NSFC) under Grant 61931005. (*Corresponding author* : *Bo Hu*.)

L. Chen, B. Hu are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: chenliangyu@bupt.edu.cn; hubo@bupt.edu.cn).

S. Chen is with the State Key Laboratory of Wireless Mobile Communications, China Academy of Telecommunication Technology, Beijing 100191, China (e-mail: chensz@cict.com).

J. Xu is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: 18111037@bjtu.edu.cn).

G. Xu is with the Department of Electrical Engineering, Tampere University, Tampere 33720, Finland (e-mail: guixian.xu@tuni.fi).

is critical to guarantee users with better experience in the resource-limited wireless networks. By exploiting the potential performance gain of NOMA scheme, this paper explores the QoE-based cross-tier resource allocation design in NOMA-enhanced HetNets.

A. Related Works

Recently, NOMA scheme has been applied to the HetNets to improve the system performance from different perspectives [7]–[13]. In [7], the authors investigated the fair energy efficient resource allocation problem in downlink NOMA HetNets with the consideration of the tradeoff among system sum rate, fairness and energy efficiency (EE), and developed the monotonic polyblock approach and a lower complexity successive convex approximation method to derive a global optimal and near-optimal solutions. In [8], the authors studied the impact of non-ideal SIC receiver on the performance of downlink NOMA HetNets. In [9], the authors investigated the EE maximization problem subject to the minimum rate requirement and the cross-tier interference constraint in NOMA HetNets. In [10], the authors proposed a joint subchannel and power allocation optimization scheme aiming at maximizing the EE of NOMA HetNet and sum rate of SUEs. Considering user fairness among the users, the authors in [11] proposed a novel time slotting (TS) technique in which time slot duration and power allocation coefficients are jointly optimized to maximize the throughput of weak users and the EE of their considered system. By considering both backhaul and access communication in HetNets, the authors in [12] investigated a cooperative transmission scheme to maximize the system achievable rate and EE, respectively. Similar to [12], the authors in [13] proposed a joint bandwidth and power allocation algorithm aiming at maximizing EE subject to the minimum data rate demands of the users and the maximum transmit power of the BSs. Different from the previous work in [7]–[13] concerning HetNet scenario, the authors in [16] considered a single-cell full-duplex multi-carrier NOMA network, and formulated a problem of joint subchannel allocation and power optimization which maximizes weighted sum system throughput. For both uplink and downlink single-cell NOMA scenario, the authors in [17] proposed a joint user clustering and power allocation scheme to maximize the sum-throughput. Considering a multi-cell in-band full duplex enabled NOMA system, the authors in [18] investigated the problem of user association, mode selection and power allocation.

In the aforementioned work, the QoS metrics (e.g., throughput, latency, coverage, etc.) have been widely investigated to optimize the network performance. Different from the QoS criteria which are primarily based on technical performance rather than perceived quality from the user perspective, QoE criteria combine technical parameters with human-related parameters [19]–[21]. The latest researches reveal that although the network-oriented QoS metrics are important, they are not perfect to evaluate user QoE. Therefore, to assure better user experience under limited wireless resources, some attempts focusing on QoE optimization through proper resource allocation have been carried out [22]–[29]. In [22], the authors investigated QoE optimization problem for device-to-device (D2D) communication in cognitive radio technology based HetNets, and proposed a joint resource block assignment, discrete power allocation and BS association design to maximize the average QoE of the D2D users. In [23], the authors considered QoE-aware joint beamforming and power optimization problem which maximizes the sum mean opinion scores (MOSs) of the users in a two-tier multiple-input multiple-output (MIMO) HetNets. In [24], the authors studied a cross-tier QoE design in a multiuser orthogonal frequency-division multiple access (OFDMA) network, with the consideration of subcarrier allocation and power optimization to improve the level of QoE among users. Considering a single-cell OMA system employing time-division multiple access (TDMA) and OFDMA, the authors in [25] proposed a joint resource block assignment and power allocation scheme aiming at maximizing the minimum MOS of the users under the minimum MOS constraint. Unlike the mentioned works mainly focusing on traditional OMA system, the QoE-oriented resource allocation optimization has also been considered in NOMA-enabled cellular network recently. In [26], the authors proposed a QoE-based power allocation scheme for wireless video service in a single-cell single-carrier NOMA network. By applying scalable video multicast transmission technique to single-cell NOMA scenario, the authors in [27] proposed a QoE-driven power allocation solution. Considering a multi-cell multi-carrier NOMA system, the authors in [28] proposed an optimal resource allocation scheme based on branch-and-bound method and a low complexity solution based on matching theory and successive convex approximation technique to address their considered QoE optimization problem optimally and sub-optimally. In [29], the authors investigated the combination of joint transmission (JT) and multi-carrier NOMA in a two-tier cognitive radio network consisting of one MBS for primary tier and multiple SBSs for secondary tier, and proposed a novel joint power control and scheduling scheme based on Augmented Lagrangian method to maximize sum MOSs of the secondary users. The summary of related works is shown in Table I.

B. Motivation and Contributions

To the best of our knowledge, the QoE-driven joint power and subchannel allocation for both macro cell users and small cell users in NOMA-enhanced HetNets, where NOMA policy is employed for both macro cell and small cell, has never been well studied. In contrast to the QoE optimization for multi-cell NOMA case [28], the NOMA-enhanced HetNet design presents new challenges in terms of interference management, because it brings additional cross-tier interference to the multi-cell network. As such, whether the combination of NOMA and HetNet is capable of enhancing the QoE of users still remains unknown. Motivated by the aforementioned issues, we investigate the QoE-aware joint power and subchannel allocation for NOMA HetNet, while taking into account the users' QoE requirement and the cross-tier interference mitigation. We consider that the NOMA HetNet provides its serving users with web browsing or video application.

TABLE I
COMPARISON BETWEEN OUR WORK AND THE EXISTING LITERATURE

Reference	Network Scenario	Multiple Access Scheme	Objective	QoE Constraints	Web Browsing	Video Streaming
[7]	heterogeneous	NOMA	EE and Fairness	×	×	×
[8]	heterogeneous	NOMA	Network Capacity	×	×	×
[9]	heterogeneous	NOMA	EE	×	×	×
[10]	heterogeneous	NOMA	EE and Network Capacity	×	×	×
[11]	heterogeneous	NOMA	Throughput, EE, and Fairness	×	√	×
[12]	heterogeneous	NOMA	Sum-Rate	×	×	×
[13]	heterogeneous	NOMA	EE	×	×	×
[16]	single-cell	NOMA	Throughput	×	×	×
[17]	single-cell	NOMA	Sum Throughput	×	×	×
[18]	multi-cell	NOMA	Time-averaged uplink and downlink rate	×	×	×
[22]	heterogeneous	OFDMA	Sum MOSs	√	√	√
[23]	heterogeneous	OMA	Sum MOSs	√	√	√
[24]	heterogeneous	OFDMA	Max-Min MOS	√	√	√
[25]	single-cell	OFDMA+TDMA	Max-Min MOS	√	√	×
[26]	single-cell	NOMA	Wireless Video QoE	×	×	×
[27]	single-cell	NOMA	Average QoE	×	×	√
[28]	multi-cell	NOMA	Sum MOSs	×	√	×
[29]	heterogeneous	OFDMA+NOMA	Sum MOSs	√	√	√
Our work	heterogeneous	NOMA	Sum MOSs	√	√	√

Such applications have become dominant application services in current wireless communication networks, and hence it is necessary to provide superior QoE for users using these services. It should be noted that there are main differences between our work and the previous work in [29], which are summarized as the following three aspects. Firstly, different from [29], where the QoE optimization is considered for the SUEs only, in this manuscript, we study the QoE-based resource allocation for both MUEs and SUEs. Secondly, in contrast to the considered scenario in [29], where OFDMA policy is employed in the macro cell, in our proposed system model, NOMA strategy is considered for both the macro cell and small cells. Thirdly, different from [29], where multi-user NOMA cluster is considered in small cells, in our system model, we assume that at most two users can reuse each subchannel of the BS (including MBS and SBSs) to reduce the error propagation and the decoding complexity at the receiver [30][1]. To summarize, the primary contributions of our work are as follows:

1) We propose a unified QoE-aware resource allocation framework for NOMA-enhanced HetNet to maximize the sum MOSs of users, which is considered the user-oriented QoE criteria rather than conventional network-oriented QoS criteria. Specifically, we formulate two QoE-based joint subchannel and power allocation optimization problems corresponding to two different services. Meanwhile, in order to ensure the minimum satisfaction of each user, we also consider the QoE constraint in the optimization problems, which is different from previous NOMA work only concerning QoS threshold. The considered optimization problems are presented as mixed-integer programming and non-convex problems.
2) We propose a penalty-based iterative algorithm, namely penalty block coordinate descent (abbreviated as P-BCD), to address the challenging optimization problems by applying the penalty method, block coordinate descent (BCD) method and sequential parametric convex approximation (SPCA) techniques. Specifically, the original non-convex problem is first converted into an equivalent and tractable one. Then, the proposed P-BCD based algorithm is developed to alternately optimize the subchannel assignment and the power allocation. Finally, we analyze the complexity of the proposed iterative optimization algorithm.
3) We quantify the benefits of our proposed QoE-based resource allocation schemes for NOMA-enhanced HetNet. Extensive numerical results demonstrate that proposed schemes can achieve competitive QoE performance compared to OMA scheme and two existing NOMA schemes in [10] and [29]. Specifically, when the number of MUEs is large, the proposed scheme can achieve better MOS performance with slightly high complexity than the reference scheme in [29].

[1]The case of multi-user NOMA cluster for the considered system model is set aside for our future work.

C. Organization

The rest of the paper is organized as follows. The considered system model and the formulated optimization problem are presented in Section II. The proposed joint optimization framework is introduced in Section III. Solutions to the QoE optimization problem for web browsing and video application are provided in Section IV and Section V, respectively. Our numerical results are presented in Section VI. The paper is concluded in Section VII.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a downlink two-tier HetNet that consists of a MBS and several SBSs as illustrated in Fig. 1. We assume that there are M MUEs, denoted by $\mathcal{M} = \{1, \cdots, M\}$, and S

TABLE II
TABLE OF NOTATIONS

Notations	Definition
$\mathcal{M}$	Set of M MUEs, $\mathcal{M} = \{1, \cdots, M\}$
$\mathcal{S}$	Set of S SBSs, $\mathcal{S} = \{1, \cdots, S\}$
$\mathcal{N}$	Set of N SCs, $\mathcal{N} = \{1, \cdots, N\}$
U_s	Set of SUEs served by SBS s
$\alpha^n_{s,u}$	SC assignment indicator for SUE u in SBS s
β^n_m	SC assignment indicator for MUE m
$h^n_{s,u}$	Channel gain from SBS s to SUE u over SC n
$h^n_{k,s,u}$	Channel gain from SBS k to SUE u in SBS s over SC n
$h^{M,S,n}_{s,u}$	Channel gain from MBS to SUE u in SBS s over SC n
g^n_m	Channel gain from MBS to MUE m over SC n
$g^{S,M,n}_{s,m}$	Channel gain from SBS s to MUE m over SC n
$p^n_{s,u}$	Transmit power from SBS s to SUE u over SC n
p^n_m	Transmit power from MBS to MUE m over SC n
$x^n_{s,u}$	Transmit signal from SBS s to SUE u over SC n
x^n_m	Transmit signal from MBS to MUE m over SC n
$\gamma^n_{s,u}, \gamma^n_m$	SINR of SUE u in SBS s and MUE m over SC n, respectively
$R_{s,u}, R_m$	Data rate of SUE u in SBS s and MUE m, respectively
$\text{PSNR}_{s,u}$	PSNR of SUE u in SBS s
PSNR_m	PSNR of MUE m
$\text{MOS}_{s,u}$	MOS of SUE u in SBS s
MOS_m	MOS of MUE m
$\text{MOS}_{\min}$	Minimum user satisfaction
$P^s_{\max}, P^M_{\max}$	Maximum transmit power of SBS s and MBS, respectively
$\mathbb{R}^{C\times 1}$	Set of all $C \times 1$ vectors with real entries
$\mathbb{Z}^{C\times 1}$	Set of all $C \times 1$ vectors with integer entries

small cells, denoted by $\mathcal{S} = \{1, \cdots, S\}$, are deployed within the macro cell area. We define U_s ($s \in \mathcal{S}$) as the set of SUEs in the small cell s. The system bandwidth W is equally divided into N subchannels (SCs), represented by $\mathcal{N} = \{1, \cdots, N\}$, and the bandwidth of each SC is $B_n = W/N$ ($n \in \mathcal{N}$). Let binary variable $\alpha^n_{s,u}$ be the subchannel assignment indicator for SUE, i.e., if SC n is allocated to SUE u in the small cell s, $\alpha^n_{s,u} = 1$; otherwise, $\alpha^n_{s,u} = 0$. Let binary variable β^n_m be the subchannel assignment indicator for MUE, i.e., if SC n is allocated to MUE m in the macro cell, $\beta^n_m = 1$; otherwise, $\beta^n_m = 0$. By applying PD-NOMA in the HetNet, superposition coding and SIC are performed at BSs and users, respectively. Since SIC performed at the receiver may lead to a considerable complexity, we consider a simple case in which at most two users can reuse each subchannel of the BS [30]. In this way, the error propagation and the decoding complexity at the receiver are reduced to a tolerable level [31]. A block fading channel is considered in this paper and channel gain does not vary on one SC. We assume perfect channel state information (CSI) is available at both MBS and SBSs [9], [30], [32]. Without loss of generality, we assume the channel gains of SUEs served by SBS s on SC n can be sorted as

$$\left|h^n_{s,1}\right| \leq \cdots \leq \left|h^n_{s,u}\right| \leq \cdots \leq \left|h^n_{s,U_s}\right|, \tag{1}$$

where $h^n_{s,u}$ denotes the channel gain of the SUE u served by SBS s on SC n. According to the NOMA policy, the SUE u can remove interference from the SUE v (for $|h^n_{s,u}| > |h^n_{s,v}|$) by performing SIC technique. Then, the remaining received signal at the SUE u served by SBS s on SC n can be expressed

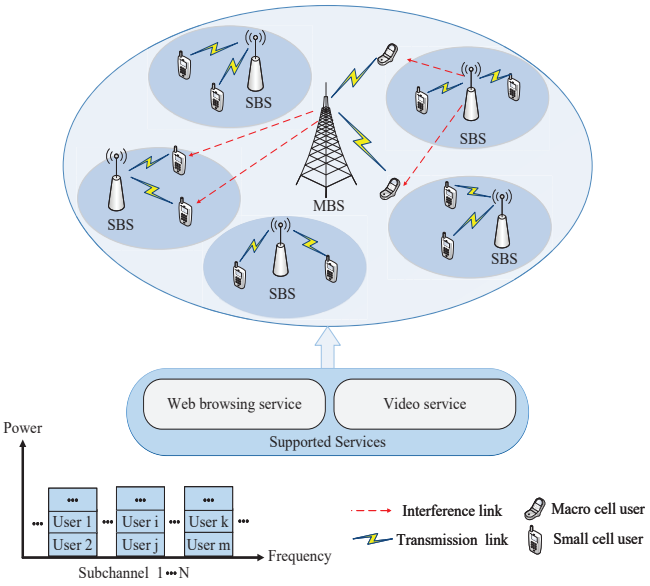


Fig. 1. A NOMA-enhanced HetNet model.

as:

$$y^n_{s,u} = \underbrace{h^n_{s,u}\sqrt{p^n_{s,u}}x^n_{s,u}}_{\text{desired signal}} + \underbrace{\sum\nolimits_{l\in U_s, |h^n_{s,l}|>|h^n_{s,u}|} h^n_{s,u}\sqrt{p^n_{s,l}}x^n_{s,l}}_{\text{intra-cell interference}} + \underbrace{\sum\nolimits_{k\in\mathcal{S}/s}\sum\nolimits_{t\in U_k} h^n_{k,s,u}\sqrt{p^n_{k,t}}x^n_{k,t}}_{\text{inter-cell interference}} + \underbrace{\sum\nolimits_{m\in\mathcal{M}} h^{M,S,n}_{s,u}\sqrt{p^n_m}x^n_m}_{\text{cross-tier interference}} + \underbrace{z^n_{s,u}}_{\text{noise}}, \tag{2}$$

where $x^n_{s,u}$ and x^n_m denote transmitted signals from SBS s to its SUE u, i.e., $SUE_{s,u}$, and from MBS to its MUE m on SC n, respectively. $p^n_{s,u}$, p^n_m are transmitted power from SBS s to $SUE_{s,u}$ and from MBS to MUE m on SC n, respectively. $h^n_{k,s,u}$, $h^{M,S,n}_{s,u}$ are channel gains from SBS k to $SUE_{s,u}$ served by SBS s, and that from MBS to $SUE_{s,u}$ served by SBS s over SC n, respectively. $z^n_{s,u}$ is noise term following the distribution $z^n_{s,u} \sim \mathcal{CN}(0, \delta^2)$. Based on (2), the signal-to-interference-plus-noise-ratio (SINR) at $SUE_{s,u}$ on SC n can be given by

$$\gamma^n_{s,u} = \frac{\left|h^n_{s,u}\right|^2 p^n_{s,u}}{I^n_{s,u,intra} + I^n_{s,u,inter} + I^n_{s,u,cross} + \delta^2}, \tag{3}$$

where $I^n_{s,u,intra} = \sum_{l\in U_s, |h^n_{s,l}|>|h^n_{s,u}|}\left|h^n_{s,u}\right|^2 p^n_{s,l}$, $I^n_{s,u,inter} = \sum_{k\in\mathcal{S}/s}\sum_{t\in U_k}\left|h^n_{k,s,u}\right|^2 p^n_{k,t}$, and $I^n_{s,u,cross} = \sum_{m\in\mathcal{M}}\left|h^{M,S,n}_{s,u}\right|^2 p^n_m$. Note that for the SUE u with the best channel quality over SC n in SBS s, it could remove the intra-cell interference of the user with weak channel quality by applying SIC, namely $I^n_{s,u,intra} = 0$ [33]. Thus, the data rate of $SUE_{s,u}$ on SC n is expressed as $R^n_{s,u} = B_n\alpha^n_{s,u}\log_2\left(1+\gamma^n_{s,u}\right)$. Then, the data rate of $SUE_{s,u}$ can be given by

$$R_{s,u} = \sum_{n\in N} R^n_{s,u} = \sum_{n\in N} B_n\alpha^n_{s,u}\log_2\left(1+\gamma^n_{s,u}\right). \tag{4}$$

In the macro cell, we assume the channel gains of MUEs on SC n are sorted as

$$\left|g_1^n\right| \leq \cdots \leq \left|g_m^n\right| \leq \cdots \leq \left|g_M^n\right|, \tag{5}$$

where g_m^n denotes the channel gain of MUE m on SC n. After applying the SIC technique at MUEs, the received signal at MUE m on SC n is expressed as

$$y_m^n = \underbrace{g_m^n \sqrt{p_m^n} x_m^n}_{\text{desired signal}} + \underbrace{\sum\nolimits_{l \in \mathcal{M}, |g_l^n| > |g_m^n|} g_m^n \sqrt{p_l^n} x_l^n}_{\text{intra-cell interference}} + \underbrace{\sum\nolimits_{s \in \mathcal{S}} \sum\nolimits_{u \in U_s} g_{s,m}^{S,M,n} \sqrt{p_{s,u}^n} x_{s,u}^n}_{\text{cross-tier interference}} + \underbrace{z_m^n}_{\text{noise}}, \tag{6}$$

where g_m^n and $g_{s,m}^{S,M,n}$ are channel gains from MBS to MUE m, and that from SBS s to MUE m over SC n, respectively. z_m^n is noise term following the distribution $z_m^n \sim \mathcal{CN}(0, \delta^2)$. According to (6), the SINR at MUE m on SC n can be expressed as

$$\gamma_m^n = \frac{\left|g_m^n\right|^2 p_m^n}{I_{m,intra}^n + I_{m,cross}^n + \delta^2}, \tag{7}$$

where $I_{m,intra}^n = \sum_{l \in \mathcal{M}, |g_l^n| > |g_m^n|} \left|g_m^n\right|^2 p_l^n$, $I_{m,cross}^n = \sum_{s \in \mathcal{S}} \sum_{u \in U_s} \left|g_{s,m}^{S,M,n}\right|^2 p_{s,u}^n$. Note that for the MUE m with the best channel quality over SC n, it would not perceive intra-cell interference after performing SIC, namely $I_{m,intra}^n = 0$ [33]. Thus, the data rate of MUE m on SC n is given by $R_m^n = B_n \beta_m^n \log_2\left(1 + \gamma_m^n\right)$. Then, the data rate of MUE m can be computed as

$$R_m = \sum_{n \in N} R_m^n = \sum_{n \in N} B_n \beta_m^n \log_2\left(1 + \gamma_m^n\right). \tag{8}$$

B. Necessary Conditions to Perform SIC

With the sorted channel gains $\left|h_{s,v}^n\right| < \left|h_{s,u}^n\right|$, $\forall s \in \mathcal{S}, n \in \mathcal{N}, (u, v) \in U_s$, SUE u can successfully decode and remove interference from the superposition signal of SUE v by SIC, if SUE u's received SINR for SUE v's signal is larger than or equal to the received SINR of SUE v for its own signal [31]–[33]. Thus, we have the following SIC decoding conditions:

$$\frac{\left|h_{s,u}^n\right|^2 p_{s,v}^n}{\sum_{l \in U_s, |h_{s,l}^n| > |h_{s,v}^n|} \left|h_{s,u}^n\right|^2 p_{s,l}^n + I_{s,u,inter}^n + I_{s,u,cross}^n + \delta^2} \geq \frac{\left|h_{s,v}^n\right|^2 p_{s,v}^n}{I_{s,v,intra}^n + I_{s,v,inter}^n + I_{s,v,cross}^n + \delta^2}, \tag{9}$$

where $I_{s,u,inter}^n = \sum_{k \in \mathcal{S}/s} \sum_{t \in U_k} \left|h_{k,s,u}^n\right|^2 p_{k,t}^n$, $I_{s,u,cross}^n = \sum_{m \in \mathcal{M}} \left|h_{s,u}^{M,S,n}\right|^2 p_m^n$, $I_{s,v,intra}^n = \sum_{l \in U_s, |h_{s,l}^n| > |h_{s,v}^n|} \left|h_{s,v}^n\right|^2 p_{s,l}^n$, $I_{s,v,inter}^n = \sum_{k \in \mathcal{S}/s} \sum_{t \in U_k} \left|h_{k,s,v}^n\right|^2 p_{k,t}^n$, and $I_{s,v,cross}^n = \sum_{m \in \mathcal{M}} \left|h_{s,v}^{M,S,n}\right|^2 p_m^n$.

Furthermore, the inequality in (9) can be equivalently rewritten as

$$\left|h_{s,u}^n\right|^2 \left(I_{s,v,inter}^n + I_{s,v,cross}^n + \delta^2\right) - \left|h_{s,v}^n\right|^2 \left(I_{s,u,inter}^n + I_{s,u,cross}^n + \delta^2\right) \geq 0. \tag{10}$$

Similarly, with the sorted channel gains $\left|g_z^n\right| < \left|g_m^n\right|$, $\forall n \in \mathcal{N}, (m, z) \in \mathcal{M}$, MUE m can correctly decode the signal from MUE z when the following inequality holds [31]–[33]:

$$\frac{\left|g_m^n\right|^2 p_z^n}{\sum_{l \in \mathcal{M}, |g_l^n| > |g_z^n|} \left|g_m^n\right|^2 p_l^n + I_{m,cross}^n + \delta^2} \geq \frac{\left|g_z^n\right|^2 p_z^n}{I_{z,intra}^n + I_{z,cross}^n + \delta^2}, \tag{11}$$

where $I_{m,cross}^n = \sum_{s \in \mathcal{S}} \sum_{u \in U_s} \left|g_{s,m}^{S,M,n}\right|^2 p_{s,u}^n$, $I_{z,intra}^n = \sum_{l \in \mathcal{M}, |g_l^n| > |g_z^n|} \left|g_z^n\right|^2 p_l^n$, $I_{z,cross}^n = \sum_{s \in \mathcal{S}} \sum_{u \in U_s} \left|g_{s,z}^{S,M,n}\right|^2 p_{s,u}^n$.

Furthermore, the inequality in (11) can be equivalently expressed as

$$\left|g_m^n\right|^2 \left(I_{z,cross}^n + \delta^2\right) - \left|g_z^n\right|^2 \left(I_{m,cross}^m + \delta^2\right) \geq 0. \tag{12}$$

C. MOS-Based QoE Evaluation Model

To evaluate user perceived quality for real-time and interactive services, the QoE model is needed to measure the users' experience. To this end, we adopt the application-oriented MOS, which is widely applied to transform objective technical parameters into the subjective user perceived quality. Different MOS models can be modelled for different services. As the two most popular applications in wireless networks [34], in this paper, we focus on web browsing as well as video streaming service.

1) Web Browsing Service: For web browsing service, its MOS model can be expressed as [24]

$$\text{MOS}_{web} = -G \ln\left(d\left(R\right)\right) + K, \tag{13}$$

where MOS_{web} value reflects the user QoE from a scale of 1 (bad) to 5 (excellent). $R\,[bit/s]$ represents the data rate. The constants G and K are determined via analyzing the experimental results of the web browsing service and are set to 1.1120 and 4.6746, respectively [24]. $d\left(R\right)$ represents the delay between the request for a web page and the reception of overall contents. $d\left(R\right)$ is related to the parameters including the round trip time, the web page size, and the employed protocols like Transfer Control Protocol (TCP) and Hypertext Transfer Protocol (HTTP). We apply these two protocols in our considered system. Thus, function $d\left(R\right)$ can be given by [34]

$$d\left(R\right) = 3RTT + \frac{FS}{R} + L\left(\frac{MSS}{R} + RTT\right) - \frac{2MSS\left(2^L - 1\right)}{R}, \tag{14}$$

where the parameters $RTT\,[s]$, $MSS\,[bit]$ and $FS\,[bit]$ represent the round trip time, the maximum segment size and the web page size, respectively. $L = \min\left[L_1, L_2\right]$ is the parameter for the packet-switching cycle from user to server in the process of downloading web pages [34]), where the parameters L_1 and L_2 are given by

$$L_1 = \log_2\left(\frac{R \cdot RTT}{MSS} + 1\right) - 1,\ L_2 = \log_2\left(\frac{FS}{2MSS} + 1\right) - 1. \tag{15}$$

Note that the MOS function corresponding to web browsing service shows a strong sensitivity with the data rate as well as FS, while the impact of RTT on MOS is less important

especially for the case of short range of *RTT* [24]. Moreover, as mentioned in the 3GPP technical specification of the LTE release 8, the *RTT* which is lower than 10 ms is expected to be achieved in the future wireless communication networks [35]. Therefore, we consider $RTT \approx 0$ ms [24] in this work. According to this assumption, the function in (14) is reformulated as $d(R_{web}) = (FS/R_{web})$. For the web user $SUE_{s,u}$ in small cell s, its MOS model can be denoted by

$$\mathrm{MOS}_{s,u}^{web} = G\ln(R_{s,u}) + M_{s,u}, \tag{16}$$

where $M_{s,u} = K - G\ln\left(FS_{s,u}\right)$ is a constant. Similarily, the MOS model of web user MUE m can be given by

$$\mathrm{MOS}_{m}^{web} = G\ln(R_{m}) + M_{m}, \tag{17}$$

where $M_m = K - G\ln\left(FS_m\right)$ is a constant.

2) Video Service: In this paper, we also apply the video streaming application with H.264/MPEG-4 Encoded Video. Furthermore, the MOS model of the video service can be expressed as [36]

$$\mathrm{MOS}_{video} = \rho\log\left(\mathrm{PSNR}\right) + \varphi, \tag{18}$$

where the parameters ρ and φ are set such that the MOS value in (18) remains the range from 1 to 4.5 [36]. For video service, the MOS value mainly depends on the peak SNR (PSNR), which can be written by

$$\mathrm{PSNR} = f + g\sqrt{\frac{R}{h}}\left(1 - \frac{h}{R}\right), \tag{19}$$

where f, g, and h is generally determined by a specific video sequence. In this work, we obtain these three parameters by using three MOS-Rate pairs. For the video user $SUE_{s,u}$ served by SBS s, its MOS model is expressed as

$$\mathrm{MOS}_{s,u}^{video} = \rho\log\left(\mathrm{PSNR}_{s,u}\right) + \varphi, \tag{20}$$

where $\mathrm{PSNR}_{s,u} = f + g\sqrt{\frac{R_{s,u}}{h}}\left(1 - \frac{h}{R_{s,u}}\right)$. Analogously, the MOS of MUE m can be given by

$$\mathrm{MOS}_{m}^{video} = \rho\log\left(\mathrm{PSNR}_{m}\right) + \varphi, \tag{21}$$

where $\mathrm{PSNR}_m = f + g\sqrt{\frac{R_m}{h}}\left(1 - \frac{h}{R_m}\right)$.

D. Problem Formulation

This paper aims to optimize the users' QoE in NOMA HetNet, by jointly designing subchannel assignment and power allocation. We consider two optimization problems corresponding to two types of applications used in the considered network. The objectives of these two problems are to maximize the sum MOSs of all users under the maximum transmitted power constraints of the BSs, and the subchannel assignment constraints. Additionally, in order to ensure the minimum satisfaction and fairness among NOMA users, we add the QoE constraints to the optimization problem. Therefore, a unified QoE optimization problem can be formulated as:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s} \mathrm{MOS}_{s,u} + \sum_{m\in\mathcal{M}} \mathrm{MOS}_m \tag{22a}$$

$$\text{s.t. } \mathrm{MOS}_{s,u} \geq \mathrm{MOS}_{\min}, \forall s,u;\ \mathrm{MOS}_m \geq \mathrm{MOS}_{\min}, \forall m, \tag{22b}$$

$$\sum_{u\in U_s}\sum_{n\in\mathcal{N}} p_{s,u}^n \leq P_{\max}^s, \forall s;\ \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} p_m^n \leq P_{\max}^M, \tag{22c}$$

$$p_{s,u}^n \geq 0, \forall s,u,\forall n;\ p_m^n \geq 0, \forall m,\forall n, \tag{22d}$$

$$\sum_{u\in U_s} \alpha_{s,u}^n \leq 2, \forall s,\forall n;\ \sum_{m\in\mathcal{M}} \beta_m^n \leq 2, \forall n, \tag{22e}$$

$$\alpha_{s,u}^n \in \{0,1\}, \forall s,u,\forall n;\ \beta_m^n \in \{0,1\}, \forall m,\forall n, \tag{22f}$$

$$(10),\ (12), \tag{22g}$$

where $\boldsymbol{\alpha} \in \mathbb{Z}^{NSU_s\times 1}$ and $\boldsymbol{\beta} \in \mathbb{Z}^{NM\times 1}$ are the collections of subchannel assignment variables $\alpha_{s,u}^n$ and β_m^n, respectively, and $\mathbf{P} \in \mathbb{R}^{(NM+NSU_s)\times 1}$ is the collections of power allocation variables $p_{s,u}^n$ and p_m^n. (22b) indicates that the minimum satisfaction of each user should be guaranteed, which is different from previous studies considering only QoS thresholds. (22c) restricts the transmitted power of BSs; (22d) demonstrates the transmitted power of each BS should be positive; (22e) and (22f) characterize the constraints on each subchannel, namely, at most two users can be multiplexed on each subchannel of the BS [30] to limit the decoding complexity at user receiver; (22g) guarantees successful SIC at user receiver. Note that (22) is a mixed-integer programming and non-convex problem due to the following three main reasons. First, the subchannel assignment variables (i.e., $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) are binary and thus (22b), (22e) and (22f) involve integer constraints. Second, objective function in (22a) is not jointly concave with respect to (w.r.t.) $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$. Third, both $\mathrm{MOS}_{s,u}$ and MOS_m in constraint (22b) are not jointly concave w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$, and thus the constraint (22b) constitutes non-convex feasible set. For solving such non-convex optimization problems, the existing methods such as polyblock outer approximation method [37] or branch-reduce-and-bound method [38] may derive the global optimal solutions. However, both methods have a worst-case complexity that could exponentially increase with the number of BSs and UEs. In the next section, we introduce a low-complexity and efficient iterative algorithm to address the formulated QoE-aware resource allocation optimization problem.

III. PROPOSED JOINT OPTIMIZATION FRAMEWORK

A. The P-BCD Optimization Framework

In this subsection, we give a brief introduction of the proposed P-BCD method, which can be employed to tackle non-convex optimization problem with coupling constraints. Consider the following problem:

$$\begin{aligned}(Q)\ \min_{z}\ & I(z)\\ \text{s.t. }& \boldsymbol{\omega}(z) \leq \mathbf{0},\\ & \boldsymbol{\mu}(z) = \mathbf{0},\\ & z \in \boldsymbol{Z},\end{aligned} \tag{23}$$

where $I(z)$ represents a scalar function with continuity and differentiability; $\boldsymbol{\mu}(z) \in \mathbb{R}^k$ denotes a vector consists of k functions with continuity and differentiability; $\boldsymbol{\omega}(z) \in \mathbb{R}^s$ denotes a vector consists of s functions with differentiability; $\boldsymbol{Z} \subseteq \mathbb{R}^n$ is a closed convex set.

To handle complicated equality constraints in problem Q, we resort to leveraging penalty method [39]. Therefore, we can formulate the following penalty problem:

$$\begin{aligned}(Q_\tau)\ \min_{z}\ & I_\tau(z) \triangleq I(z) + \frac{\tau}{2}\left\|\boldsymbol{\mu}(z)\right\|^2 \\ \text{s.t. }& \boldsymbol{\omega}(z) \le \mathbf{0}, \\ & z \in \boldsymbol{Z},\end{aligned} \tag{24}$$

where $\tau > 0$ acts as a scalar penalty parameter. Particularly, the solution of problem Q_τ is identical to that of problem Q when $\tau \to \infty$ [39]. However, problem Q_τ is still hard to be solved when $I(z)$, $\boldsymbol{\mu}(z)$ and $\boldsymbol{\omega}(z)$ are non-convex w.r.t. the variables z. To address Q_τ, we propose P-BCD method summarized in **Algorithm 1**. At each iteration, Q_τ can be approximately solved by adopting BCD [40] and SPCA techniques [41]. The basic idea behind BCD method integrated with SPCA for solving problem Q_τ is to successively minimize a locally tight upper of the objective of Q_τ, until a stationary point is obtained. The convergence analysis of P-BCD method is similar to that of [42].

Algorithm 1 P-BCD Algorithm for solving Q_τ

1: **Initialize** the maximum iterations $\triangle_1$, maximum tolerance ε_v, iteration index v, and parameter $c > 1$ for penalty.
2: **repeat**
3: Solve penalty problem Q_{τ_v} with given $\boldsymbol{\tau}_v$ by the proposed BCD based method, denote $z_{v+1} = BCD(Q_{\tau_v}, z_v)$ as obtained feasible solutions.
4: Update the penalty parameter $\boldsymbol{\tau}_v$ by $\boldsymbol{\tau}_{v+1} = c\boldsymbol{\tau}_v$.
5: Set $v = v + 1$ and compute $I_\tau(z_{v+1})$.
6: **until** $\frac{\left|I_\tau(z_{v+1}) - I_\tau(z_v)\right|}{\left|I_\tau(z_v)\right|} \le \varepsilon_v$ or meet the maximum iteration number $v = \triangle_1$.

As observed from **Algorithm 1**, the problem Q_τ can be decoupled into inner layer (step 3 in **Algorithm 1**) and outer layer. For a given penalty parameter τ_v in Q_{τ_v}, the inner-layer problem can be solved approximately via BCD method in combination with SPCA techniques. Then, according to Step 4 of **Algorithm 1**, the penalty parameter τ_v is updated iteratively at outer layer. The above steps are repeated until some termination criteria are satisfied.

The overview of the proposed P-BCD optimization framework is illustrated in Fig. 2, which is presented at the top of the next page. A unified QoE-aware joint subchannel and power optimization problem is formulated in (22). Furthermore, two problems corresponding to two adopted services are formulated in (25) and (40), respectively. To handle the complex binary constraints on subchannel allocation, the original problem is first reformulated into an equivalent penalty problem. Then, for a given penalty value, penalty problem can be solved via BCD method. Specifically, the variables of subchannel allocation and transmitted power are alternately optimized in the inner layer. When subchannel assignment subproblem or power allocation subproblem is non-convex, a series of transformations can be adopted to convert the non-convex problem into a convex form, which can be efficiently addressed by SPCA technique. The proposed method for solving the two QoE optimization problems will be discussed in detail in Section IV and V, respectively.

IV. Web Browsing Service Case

In this section, we sought to maximize aggregated MOSs of web users by carefully allocating the power and subchannel resource. Meanwhile, QoE constraint on each web user is considered in this QoE optimization problem to further guarantee minimum satisfaction for each web user. Since the considered QoE-based joint subchannel and power allocation optimization problem is complex to solve, we first transform this problem into an equivalent but more tractable one by invoking a series of auxiliary variables. Then, we apply the proposed P-BCD method to tackle the converted problem. Finally, we give complexity analysis of the proposed optimization method.

A. Problem Transformation

By integrating the MOS models of web browsing service (i.e., (16) and (17)) into (22), we can formulate the following joint subchannel and power allocation optimization problem

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P}}\ & \sum_{s\in\mathcal{S}}\sum_{u\in U_s} G\ln\Big(\sum_{n\in\mathcal{N}} B_n\alpha^n_{s,u}\log_2\left(1+\gamma^n_{s,u}\right)\Big) \\ & + \sum_{m\in\mathcal{M}} G\ln\Big(\sum_{n\in\mathcal{N}} B_n\beta^n_m\log_2\left(1+\gamma^n_m\right)\Big) + \vartheta \qquad (25a)\\ \text{s.t. }& (22b)-(22g). \qquad (25b)\end{aligned}$$

where $\vartheta = \sum_{s\in\mathcal{S}}\sum_{u\in U_s} M_{s,u} + \sum_{m\in\mathcal{M}} M_m$. Problem (25) is challenging to solve because the objective function in (25a) is non-concave w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$, and (22b) in the problem (25) is a non-convex constraint w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$. Meanwhile, the optimization variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are integer forms. Thus, we need to transform the problem (25) into a more tractable one. First, let us tackle the non-convex objective function (25a). Introducing auxiliary variables $\boldsymbol{\eta_s} = \{\eta_{s,u}\} \in \mathbb{R}^{SU_s\times 1}$ and $\boldsymbol{\eta_m} = \{\eta_m\} \in \mathbb{R}^{M\times 1}$, the problem (25) can be rewritten as

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P}\\ \boldsymbol{\eta_s},\boldsymbol{\eta_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) + \vartheta \tag{26a}$$

$$\text{s.t. } \sum_{n\in\mathcal{N}} B_n\alpha^n_{s,u}\log_2\left(1+\gamma^n_{s,u}\right) \ge \eta_{s,u}, \forall s,u;$$

$$\sum_{n\in\mathcal{N}} B_n\beta^n_m\log_2\left(1+\gamma^n_m\right) \ge \eta_m, \forall m, \tag{26b}$$

$$(22b)-(22g). \tag{26c}$$

Now, the objective function (26a) is a concave function w.r.t. $\boldsymbol{\eta_s}$ and $\boldsymbol{\eta_m}$. Subsequently, by introducing the auxiliary variables $\tilde{\boldsymbol{\alpha}} = \{\tilde{\alpha}^n_{s,u}\} \in \mathbb{Z}^{NSU_s\times 1}$ and $\tilde{\boldsymbol{\beta}} = \{\tilde{\beta}^n_m\} \in \mathbb{Z}^{NM\times 1}$, the integer constraint (22f) in the problem (26) can be equivalently transformed into the following forms [43]

$$\alpha^n_{s,u}(1-\tilde{\alpha}^n_{s,u}) = 0, \forall s,u,\forall n;\ \alpha^n_{s,u} = \tilde{\alpha}^n_{s,u}, \forall s,u,\forall n, \tag{27}$$

$$\beta^n_m(1-\tilde{\beta}^n_m) = 0, \forall m,\forall n;\ \beta^n_m = \tilde{\beta}^n_m, \forall m,\forall n. \tag{28}$$

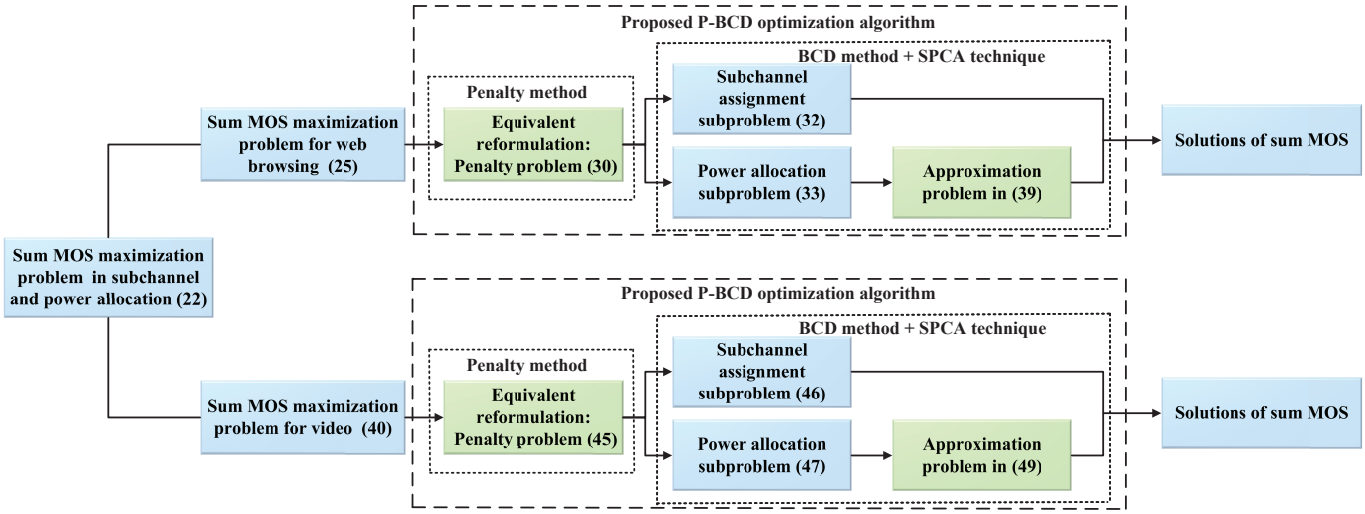


Fig. 2. The proposed P-BCD optimization framework for solving the sum MOSs maximization problem (22).

As a result, the integer constraint (22f) in the problem (26) is equivalently converted into corresponding equality constraints, i.e., (27)-(28). Then, (25) is equivalently reformulated as

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\\ \boldsymbol{\eta_s},\boldsymbol{\eta_m},\mathbf{P}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) + \vartheta \tag{29a}$$

$$\text{s.t. } (22b)-(22e),\ (22g),\ (26b),\ (27)-(28). \tag{29b}$$

B. The P-BCD for Solving (29)

In this subsection, we apply the P-BCD method described in Section III-A to address problem (29). To handle the converted equality constraints in (27)-(28), we first absorb the penalty terms into the objective function and formulate a penalized problem corresponding to problem (29). Then, for a given penalty parameter, we develop a BCD based inner-loop iterative algorithm to deal with the resultant optimization problem. Specifically, the subchannel allocation and transmitted power will be alternately optimized in the inner loop. Finally, we summarize the proposed P-BCD based algorithm for tackling the resultant resource allocation problem.

1) The Penalized Problem

The problem (29) is further complicated significantly due to the presence of equality constraints in (27)-(28). According to P-BCD method shown in Subsection III-A, we can get the following penalized problem

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\\ \boldsymbol{\eta_s},\boldsymbol{\eta_m},\mathbf{P}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) - \tau\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) + \vartheta \tag{30a}$$

$$\text{s.t.}\quad (22b)-(22e),\ (22g),\ (26b), \tag{30b}$$

where $\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) = \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\sum_{n\in\mathcal{N}}\left(\left|\alpha^n_{s,u}-\tilde{\alpha}^n_{s,u}\right|^2 + \left|\alpha^n_{s,u}(1-\tilde{\alpha}^n_{s,u})\right|^2\right) + \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\left(\left|\beta^n_m-\tilde{\beta}^n_m\right|^2 + \left|\beta^n_m(1-\tilde{\beta}^n_m)\right|^2\right)$. One can see that when $\tau\to\infty$, the solution to (30) is an approximate solution to (29) [39]. The problem (30) is still complex, since constraint (22b) in the problem (30) is a non-convex constraint w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$. Next, we need to address problem (30) with given penalty parameter τ.

2) The BCD Algorithm for Solving (30)

After constructing the penalty problem (30), we employ BCD method and SPCA technique to address the non-convex optimization problem (30) with given τ. According to BCD method [40], the optimization variables in inner-layer problem can be divided into three blocks for auxiliary variables $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$, subchannel assignment $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$, and power allocation $\{\mathbf{P}\}$. Then, we optimize these three blocks alternately. However, even with given solutions of auxiliary variables $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$ and subchannel assignment $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$, the resultant power allocation subproblem is still hard to handle due to its non-convexity. Thus, a series of transformations are further adopted via SPCA such that original inner-layer optimization problem is converted to a solvable and tractable one.

Step1: we optimize variables $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$ for any given subchannel allocation policy $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$. One can know that $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$ only lies in the objective function (30a). Similar to [43], we can obtain a closed-form solution for $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$, which can be written by

$$\tilde{\alpha}^n_{s,u} = \frac{\alpha^n_{s,u}+(\alpha^n_{s,u})^2}{1+(\alpha^n_{s,u})^2}, \forall s,u,\forall n;\quad \tilde{\beta}^n_m = \frac{\beta^n_m+(\beta^n_m)^2}{1+(\beta^n_m)^2}, \forall m,\forall n. \tag{31}$$

Step2: we optimize variables $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$ for any given auxiliary variables and power allocation policy $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\mathbf{P}\}$. Then subchannel assignment subproblem corresponding to (30) is given by

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta_s},\boldsymbol{\eta_m}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) - \tau\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) + \vartheta \tag{32a}$$

$$\text{s.t. } \sum_{n\in\mathcal{N}} B_n\alpha^n_{s,u}\log_2\left(1+\gamma^n_{s,u}\right) \ge \eta_{s,u},\forall s,u;$$
$$\sum_{n\in\mathcal{N}} B_n\beta^n_m\log_2\left(1+\gamma^n_m\right) \ge \eta_m,\forall m, \tag{32b}$$

$$(22b),\ (22e). \tag{32c}$$

The objective function in (32a) is a concave function w.r.t. $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta_s}$ and $\boldsymbol{\eta_m}$. (22b) and (22e) in the problem (32) are convex constraints w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In addition, (32b) is a

convex constraint w.r.t. $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta_s}$ and $\boldsymbol{\eta_m}$. Therefore, the convex problem (32) can be effectively addressed by employing advanced convex solvers, e.g., CVX [44].

Step3: we optimize variables $\{\mathbf{P}\}$ for any given auxiliary variables and subchannel assignment solution $\left\{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right\}$. Then power allocation subproblem corresponding to (30) can be given by

$$\max_{\mathbf{P}, \boldsymbol{\eta_s}, \boldsymbol{\eta_m}} \sum_{s\in\mathcal{S}} \sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) - \tau\Gamma(\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) + \vartheta \tag{33a}$$

$$\text{s.t.} \sum_{n\in\mathcal{N}} B_n \alpha_{s,u}^n \log_2\left(1+\gamma_{s,u}^n\right) \geq \eta_{s,u}, \forall s, u;$$

$$\sum_{n\in\mathcal{N}} B_n \beta_m^n \log_2\left(1+\gamma_m^n\right) \geq \eta_m, \forall m, \tag{33b}$$

$$(22b)-(22d),\ (22g). \tag{33c}$$

Since (22b) in the problem (33) is a non-convex constraint w.r.t. $\mathbf{P}$, and (33b) is also a non-convex constraint w.r.t $\mathbf{P}$, $\boldsymbol{\eta_s}$ and $\boldsymbol{\eta_m}$, the problem (33) is a non-convex optimization problem. By introducing slack variables $\mathbf{t_s} = \left\{t_{s,u}^n\right\} \in \mathbb{R}^{NSU_s\times 1}$ and $\mathbf{t_m} = \left\{t_m^n\right\} \in \mathbb{R}^{NM\times 1}$, the problem (33) can be reformulated as

$$\max_{\substack{\mathbf{P}, \boldsymbol{\eta_s}, \boldsymbol{\eta_m} \\ \mathbf{t_s}, \mathbf{t_m}}} \sum_{s\in\mathcal{S}} \sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) - \tau\Gamma(\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) + \vartheta \tag{34a}$$

$$\text{s.t.} \sum_{n\in\mathcal{N}} B_n \alpha_{s,u}^n \log_2\left(t_{s,u}^n\right) \geq \eta_{s,u}, \forall s, u;$$

$$\sum_{n\in\mathcal{N}} B_n \beta_m^n \log_2\left(t_m^n\right) \geq \eta_m, \forall m, \tag{34b}$$

$$\sum_{n\in\mathcal{N}} B_n \alpha_{s,u}^n \log_2\left(t_{s,u}^n\right) \geq \varpi_{s,u}, \forall s, u;$$

$$\sum_{n\in\mathcal{N}} B_n \beta_m^n \log_2\left(t_m^n\right) \geq \varpi_m, \forall m, \tag{34c}$$

$$\gamma_{s,u}^n \geq t_{s,u}^n - 1, \forall s, u, \forall n;\ \gamma_m^n \geq t_m^n - 1, \forall m, \forall n, \tag{34d}$$

$$t_{s,u}^n \geq 0, \forall s, u, \forall n;\ t_m^n \geq 0, \forall m, \forall n, \tag{34e}$$

$$(22c)-(22d),\ (22g). \tag{34f}$$

where $\varpi_{s,u} = \exp(\frac{\text{MOS}_{\min} - M_{s,u}}{G})$, $\varpi_m = \exp(\frac{\text{MOS}_{\min} - M_m}{G})$. Note that (34d) is a non-convex constraint w.r.t. $\mathbf{P}$, $\mathbf{t_m}$ and $\mathbf{t_s}$. By introducing the slack variables $\mathbf{s_s} = \left\{s_{s,u}^n\right\} \in \mathbb{R}^{NSU_s\times 1}$ and $\mathbf{s_m} = \left\{s_m^n\right\} \in \mathbb{R}^{NM\times 1}$, (34d) can be transformed into the following form

$$\sum_{l\in U_s, |h_{s,l}^n| > |h_{s,u}^n|} \left|h_{s,u}^n\right|^2 p_{s,l}^n + \sum_{k\in\mathcal{S}/s} \sum_{t\in U_k} \left|h_{k,s,u}^n\right|^2 p_{k,t}^n + \sum_{m\in\mathcal{M}} \left|h_{s,u}^{M,S,n}\right|^2 p_m^n + \delta_z^2 \leq s_{s,u}^n, \forall s, u, \forall n \tag{35}$$

$$\sum_{l\in\mathcal{M}, |g_l^n| > |g_m^n|} \left|g_m^n\right|^2 p_l^n + \sum_{s\in\mathcal{S}} \sum_{u\in U_s} \left|g_{s,m}^{S,M,n}\right|^2 p_{s,u}^n + \delta^2 \leq s_m^n, \forall m, \forall n \tag{36}$$

$$\left|h_{s,u}^n\right|^2 p_{s,u}^n \geq s_{s,u}^n \left(t_{s,u}^n - 1\right), \forall s, u, \forall n;$$

$$\left|g_m^n\right|^2 p_m^n \geq s_m^n \left(t_m^n - 1\right), \forall m, \forall n. \tag{37}$$

Note that the introduced constraints in (37) are non-convex, since $s_{s,u}^n t_{s,u}^n$ and $s_m^n t_m^n$ in (37) are quasi-concave functions. Hence, we need to transform these two functions into convex forms. Utilizing SPCA techniques [41], $s_{s,u}^n t_{s,u}^n$ and $s_m^n t_m^n$ can be converted into the following convex upper bounds

$$\frac{\lambda_{s,u}^n}{2}\left(t_{s,u}^n\right)^2 + \frac{1}{2\lambda_{s,u}^n}\left(s_{s,u}^n\right)^2 \geq t_{s,u}^n s_{s,u}^n, \forall s, u, \forall n;$$

$$\frac{\lambda_m^n}{2}\left(t_m^n\right)^2 + \frac{1}{2\lambda_m^n}\left(s_m^n\right)^2 \geq t_m^n s_m^n, \forall m, \forall n, \tag{38}$$

where $\boldsymbol{\lambda_s} = \left\{\lambda_{s,u}^n\right\} \in \mathbb{R}^{NSU_s\times 1}$ and $\boldsymbol{\lambda_m} = \left\{\lambda_m^n\right\} \in \mathbb{R}^{NM\times 1}$. The equalities in (38) are satisfied by $\lambda_{s,u}^n = s_{s,u}^n / t_{s,u}^n$ and $\lambda_m^n = s_m^n / t_m^n$. After a series of transformations based on (34)-(38), the problem (33) can be reformulated as

$$\max_{\substack{\mathbf{P}, \boldsymbol{\eta_s}, \boldsymbol{\eta_m} \\ \mathbf{t_s}, \mathbf{t_m}, \mathbf{s_s}, \mathbf{s_m}}} \sum_{s\in\mathcal{S}} \sum_{u\in U_s} G\ln(\eta_{s,u}) + \sum_{m\in\mathcal{M}} G\ln(\eta_m) - \tau\Gamma(\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) + \vartheta \tag{39a}$$

$$\text{s.t.} \left|h_{s,u}^n\right|^2 p_{s,u}^n \geq \frac{\lambda_{s,u}^n}{2}\left(t_{s,u}^n\right)^2 + \frac{1}{2\lambda_{s,u}^n}\left(s_{s,u}^n\right)^2 - s_{s,u}^n, \forall s, u, \forall n, \tag{39b}$$

$$\left|g_m^n\right|^2 p_m^n \geq \frac{\lambda_m^n}{2}\left(t_m^n\right)^2 + \frac{1}{2\lambda_m^n}\left(s_m^n\right)^2 - s_m^n, \forall m, \forall n, \tag{39c}$$

$$(34b)-(34c),\ (34e)-(34f),\ (35)-(36). \tag{39d}$$

Finally, the convex optimization problem (39) can be solved efficiently by using CVX [44].

Algorithm 2 BCD algorithm for solving (30)

1: **Initialize** the maximum iterations $\triangle_2$, maximum tolerance ε_l, iteration index l, feasible solution $\left\{\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \mathbf{P}_0\right\}$, auxiliary parameter $\boldsymbol{\lambda_s} > \mathbf{0}$ and $\boldsymbol{\lambda_m} > \mathbf{0}$.
2: **repeat**
3: For given $\left\{\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l\right\}$, obtain optimal solution $\left\{\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}\right\}$ according to (31).
4: For given $\left\{\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_l, \boldsymbol{\beta}_l, \mathbf{P}_l\right\}$, obtain optimal solution $\left\{\boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}\right\}$ by solving (32).
5: For given $\left\{\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}, \mathbf{P}_l\right\}$, obtain optimal solution $\mathbf{P}_{l+1}$ by solving (39).
6: Update $\boldsymbol{\lambda_s}_{,l+1}$ and $\boldsymbol{\lambda_m}_{,l+1}$, respectively, by using $\lambda_{s,u,l+1}^n = \frac{s_{m,u,l}^n}{t_{s,u,l}^n}$ and $\lambda_{m,l+1}^n = \frac{s_{m,l}^n}{t_{m,l}^n}$.
7: Update the iteration number: $l = l + 1$.
8: **until** $\frac{\left|\Psi_{\tau_v}(l+1) - \Psi_{\tau_v}(l)\right|}{\left|\Psi_{\tau_v}(l)\right|} \leq \varepsilon_l$ or meet the maximum iteration number $l = \triangle_2$.

The BCD algorithm is summarized in **Algorithm 2**. This algorithm can be exploited to solve (30) in the inner-layer iteration of the P-BCD framework (i.e., Step 3 in **Algorithm 1**). Specifically, the design variables in (30) are divided into three blocks, i.e., $\left\{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\right\}$, $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ and $\{\mathbf{P}\}$. Then, the auxiliary variables $\left\{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\right\}$, subchannel assignment $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$, and power allocation $\{\mathbf{P}\}$ are alternately optimized by solving subproblem (31), (32), and (39), respectively, while the other two blocks are fixed. At each iteration of **Algorithm 2**, the objective value of (30) is non-decreasing and bounded. The convergence of the BCD algorithm is proved in Appendix A.

After **Algorithm 2** converges, the penalty parameter τ_v is further updated by following Step 4 in **Algorithm 1**, i.e.,

$\tau_{v+1} = c\tau_v$. The above steps are iteratively repeated until **Algorithm 1** converges. The convergence analysis of the P-BCD algorithm is similar to that of [42].

C. Complexity Analysis

In the following, we probe into the complexity of the proposed P-BCD algorithm for solving the problem (29). **Table III** also summarizes the complexity.

i) Outer Layer: Suppose the iteration times of outer layer (Penalty approach) is $\triangle_1$.

ii) Inner Layer: Let $\triangle_2$ be iteration times of inner layer (BCD approach). At each iteration of inner loop, there are three subproblems, namely, (31), (32) and (39). Since the optimal solutions of (31) can be obtained by a closed-form solution, the complexity for (31) can be neglected. Therefore, solving the subchannel assignment subproblem (32) and the power allocation subproblem (39) dominate the overall complexity of the BCD algorithm. When CVX toolbox is adopted to solve these two convex optimization subproblems, i.e., (32) and (39), it exploits GP with the interior-point method (IPM). The complexity of subchannel assignment subproblem (32) is $O(\log(I_1/\xi\omega)/\log(\psi))$ [45], where $I_1 = 2(SU+M)+SN+N$ is the total number of constraints in (32). ξ denotes the initial point for approximating the accuracy of IPM, ω represents the stopping criterion for IPM, and ψ is employed to update the accuracy of IPM. Analogously, the complexity of power allocation subproblem (39) is $O(\log(I_2/\xi\omega)/\log(\psi))$, where $I_2 = 5(SUN+MN)+2(SU+M)+S+1$ is the total number of constraints in (39).

Therefore, the total computational complexity of P-BCD method for solving problem (29) is $O(\triangle_1 \times \triangle_2(\log(I_1/\xi\omega)/\log(\psi)+\log(I_2/\xi\omega)/\log(\psi)))$. Besides, the complexity of the algorithm proposed by [29] for web browsing is $O(\triangle_3 \times \triangle_4(2(SUN+MN)^2))$ [46], where $\triangle_3$ and $\triangle_4$ represent iteration times for outer layer and inner layer, respectively.

V. Video Service Case

In this section, we focus on the QoE-based joint subchannel and power allocation design under video service case. Different from web browsing application, the MOS function of video service is mainly determined by PSNR. As a result, a distinct different resource allocation policy would have to be explored due to the impact of different application parameter on the user QoE. Besides, to ensure minimum satisfaction and fairness between the video users, the satisfaction thresholds are introduced into the QoE optimization problem. In the following, we adopt the proposed P-BCD method to maximize the sum MOSs of video users by proper subchannel and power allocation.

A. Problem Transformation

By incorporating MOS function of video service (i.e., (20) and (21)) into (22), we can formulate the following joint subchannel and power allocation optimization problem

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(f+g\sqrt{\frac{R_{s,u}}{h}}\left(1-\frac{h}{R_{s,u}}\right)\right)+\varphi\right) + \sum_{m\in\mathcal{M}}\left(\rho\log\left(f+g\sqrt{\frac{R_m}{h}}\left(1-\frac{h}{R_m}\right)\right)+\varphi\right) \tag{40a}$$

$$\text{s.t. } (22b)-(22g). \tag{40b}$$

One can observe that the objective function in (40a) is a non-concave function w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$, and (22b) in the problem (40) is a non-convex constraint w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$. Meanwhile, (22f) in the problem (40) involves integer constraint. Thus, the problem (40) is challenging to solve. For this reason, we first convert the objective function into an equivalent form. Introducing auxiliary variables $\boldsymbol{\zeta_s} = \{\zeta_{s,u}\} \in \mathbb{R}^{SU_s\times 1}$ and $\boldsymbol{\zeta_m} = \{\zeta_m\} \in \mathbb{R}^{M\times 1}$, the problem (40) is equivalently reformulated as

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P}\\ \boldsymbol{\zeta_s},\boldsymbol{\zeta_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(\zeta_{s,u}\right)+\varphi\right)+\sum_{m\in\mathcal{M}}\left(\rho\log\left(\zeta_m\right)+\varphi\right) \tag{41a}$$

$$\text{s.t. } f+g\sqrt{\frac{R_{s,u}}{h}}\left(1-\frac{h}{R_{s,u}}\right)\geq\zeta_{s,u},\forall s,u;$$
$$f+g\sqrt{\frac{R_m}{h}}\left(1-\frac{h}{R_m}\right)\geq\zeta_m,\forall m, \tag{41b}$$
$$(22b)-(22g). \tag{41c}$$

Note that the added constraint (41b) is a non-convex constraint because the left-hand-side (LHS) of (41b), i.e., $f+g\sqrt{\frac{R_{s,u}}{h}}\left(1-\frac{h}{R_{s,u}}\right)\geq\zeta_{s,u}$ and $f+g\sqrt{\frac{R_m}{h}}\left(1-\frac{h}{R_m}\right)\geq\zeta_m$, are non-concave function w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{P}$. By introducing auxiliary variables $\boldsymbol{\eta_s} = \{\eta_{s,u}\} \in \mathbb{R}^{SU_s\times 1}$ and $\boldsymbol{\eta_m} = \{\eta_m\} \in \mathbb{R}^{M\times 1}$, (41b) can be converted into the following forms

$$f+g\sqrt{\frac{\eta_{s,u}}{h}}\left(1-\frac{h}{\eta_{s,u}}\right)\geq\zeta_{s,u},\forall s,u;$$
$$f+g\sqrt{\frac{\eta_m}{h}}\left(1-\frac{h}{\eta_m}\right)\geq\zeta_m,\forall m. \tag{42}$$
$$\sum_{n\in\mathcal{N}}B_n\alpha_{s,u}^n\log_2\left(1+\gamma_{s,u}^n\right)\geq\eta_{s,u},\forall s,u;$$
$$\sum_{n\in\mathcal{N}}B_n\beta_m^n\log_2\left(1+\gamma_m^n\right)\geq\eta_m,\forall m. \tag{43}$$

Then, by introducing the auxiliary variables $\tilde{\boldsymbol{\alpha}} = \{\tilde{\alpha}_{s,u}^n\} \in \mathbb{Z}^{NSU_s\times 1}$ and $\tilde{\boldsymbol{\beta}} = \{\tilde{\beta}_m^n\} \in \mathbb{Z}^{NM\times 1}$ and using (27)-(28), the discrete subchannel assignment constraint (22f) in the problem (40) is transformed into corresponding equality constraints. After the above transformations, the problem (40) is reformulated as

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\mathbf{P}\\ \boldsymbol{\zeta_s},\boldsymbol{\zeta_m},\boldsymbol{\eta_s},\boldsymbol{\eta_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(\zeta_{s,u}\right)+\varphi\right)+\sum_{m\in\mathcal{M}}\left(\rho\log\left(\zeta_m\right)+\varphi\right) \tag{44a}$$

$$\text{s.t. } (22b)-(22e),\ (22g),\ (27)-(28),\ (42)-(43). \tag{44b}$$

B. The P-BCD for Solving (44)

In this subsection, we apply the P-BCD technique described in Subsection III-A to tackle the problem (44). Similar to Subsection IV-B, we first formulate a penalized problem corresponding to (44) to handle the converted equality constraints in (27)-(28). Then, we apply the BCD-based inner-loop iterative approach to update the design variables alternately by solving the resulting optimization problem with given penalty parameter. After the inner-loop iterative algorithm converges, the penalty parameter will be updated iteratively at the outer loop.

1) The Penalized Problem

According to the P-BCD optimization framework shown in Subsection III-A, we can obtain the following penalized version of the problem (44)

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\mathbf{P}\\ \boldsymbol{\zeta_s},\boldsymbol{\zeta_m},\boldsymbol{\eta_s},\boldsymbol{\eta_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(\zeta_{s,u}\right)+\varphi\right)+\sum_{m\in\mathcal{M}}\left(\rho\log\left(\zeta_m\right)+\varphi\right) - \tau\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) \tag{45a}$$

$$\text{s.t. } (22b)-(22e),\ (22g),\ (42)-(43), \tag{45b}$$

where $\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) = \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\sum_{n\in\mathcal{N}}(\left|\alpha^n_{s,u}-\tilde{\alpha}^n_{s,u}\right|^2+\left|\alpha^n_{s,u}(1-\tilde{\alpha}^n_{s,u})\right|^2)+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}(\left|\beta^n_m-\tilde{\beta}^n_m\right|^2+\left|\beta^n_m(1-\tilde{\beta}^n_m)\right|^2)$.

2) The BCD Algorithm for Solving (45)

It is observed that the design variables in problem (45) are closely coupled with each other, which makes (45) intractable. In the following, we apply BCD method and SPCA techniques to solve (45) with given τ. Similar to Subsection IV-C, we divide variables of (45) into three blocks, i.e., auxiliary variables $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$, subchannel assignment $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$, and power allocation $\{\mathbf{P}\}$, respectively. These three blocks can be optimized alternately until the BCD algorithm converges.

Step1: we optimize variables $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$ for any given subchannel assignment policy $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$. By using (31), the optimal solution $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}\}$ can be obtained.

Step2: we optimize variables $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$ for any given auxiliary variables and power allocation policy $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\mathbf{P}\}$. Then subchannel assignment subproblem corresponding to (45) is expressed as

$$\max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\zeta_s}\\ \boldsymbol{\zeta_m},\boldsymbol{\eta_s},\boldsymbol{\eta_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(\zeta_{s,u}\right)+\varphi\right)+\sum_{m\in\mathcal{M}}\left(\rho\log\left(\zeta_m\right)+\varphi\right) - \tau\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) \tag{46a}$$

$$\text{s.t. } (22b),\ (22e),\ (42)-(43). \tag{46b}$$

The objective function in (46a) is a concave function w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\zeta_s}$ and $\boldsymbol{\zeta_m}$. Both (22b) and (22e) in the problem (46) are convex constraint w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. (42) in the problem (46) is a convex constraint w.r.t. $\boldsymbol{\eta_s}$, $\boldsymbol{\eta_m}$, $\boldsymbol{\zeta_s}$ and $\boldsymbol{\zeta_m}$, because the LHS of the constraint (42) in the problem (46), i.e., $f+g\sqrt{\frac{\eta_{s,u}}{h}}\left(1-\frac{h}{\eta_{s,u}}\right)$ and $f+g\sqrt{\frac{\eta_m}{h}}\left(1-\frac{h}{\eta_m}\right)$, are concave function w.r.t. $\boldsymbol{\eta_s}$ and $\boldsymbol{\eta_m}$. In addition, (43) in the problem (46) is a convex constraint w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\zeta_s}$ and $\boldsymbol{\zeta_m}$. Therefore, the convex optimization problem (46) can be effectively solved by advanced convex solvers, e.g., CVX [44].

Step3: we optimize variables $\{\mathbf{P}\}$ for any given auxiliary variables and subchannel assignment policy $\{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\beta}\}$. Then power allocation subproblem corresponding to (45) is formulated by

$$\max_{\substack{\mathbf{P},\boldsymbol{\zeta_s},\boldsymbol{\zeta_m}\\ \boldsymbol{\eta_s},\boldsymbol{\eta_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(\zeta_{s,u}\right)+\varphi\right)+\sum_{m\in\mathcal{M}}\left(\rho\log\left(\zeta_m\right)+\varphi\right) - \tau\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) \tag{47a}$$

$$\text{s.t. } (22b)-(22d),\ (22g),\ (42)-(43). \tag{47b}$$

Note that (22b) in the problem (47) is a non-convex constraint w.r.t. $\mathbf{P}$. In addition, since the LHS of the constraint (43) in the problem (47), i.e., $\sum_{n\in\mathcal{N}}B_n\alpha^n_{s,u}\log_2\left(1+\gamma^n_{s,u}\right)$ and $\sum_{n\in\mathcal{N}}B_n\beta^n_m\log_2\left(1+\gamma^n_m\right)$ are non-concave functions w.r.t. $\mathbf{P}$, the constraint (43) in the problem (47) constitutes non-convex feasible set. Thus, the problem (47) is a non-convex optimization problem. First, we handle the non-convex constraint (22b) in the problem (47). With the slack variables $\boldsymbol{\eta_s}=\{\eta_{s,u}\}\in\mathbb{R}^{SU_s\times 1}$ and $\boldsymbol{\eta_m}=\{\eta_m\}\in\mathbb{R}^{M\times 1}$, we introduce the following convex constraint:

$$\begin{aligned} &f+g\sqrt{\frac{\eta_{s,u}}{h}}\left(1-\frac{h}{\eta_{s,u}}\right)\geq 10^{\frac{\mathrm{MOS}_{\min}-\varphi}{\rho}},\forall s,u;\\ &f+g\sqrt{\frac{\eta_m}{h}}\left(1-\frac{h}{\eta_m}\right)\geq 10^{\frac{\mathrm{MOS}_{\min}-\varphi}{\rho}},\forall m. \end{aligned} \tag{48}$$

Therefore, the non-convex constraint (22b) in the problem (47) is converted into a convex constraint (48) as well as a non-convex constraint (43). Then, we focus on tackling the non-convex constraint (43). Following SPCA approach [41], we introduce auxiliary variables $\mathbf{t_s}=\left\{t^n_{s,u}\right\}\in\mathbb{R}^{NSU_s\times 1}$, $\mathbf{t_m}=\left\{t^n_m\right\}\in\mathbb{R}^{NM\times 1}$, $\mathbf{s_s}=\left\{s^n_{s,u}\right\}\in\mathbb{R}^{NSU_s\times 1}$ and $\mathbf{s_m}=\left\{s^n_m\right\}\in\mathbb{R}^{NM\times 1}$. Furthermore, the non-convex constraint (43) in the problem (47) can be converted into convex forms by utilizing (34b), (34e), (35)-(36), and (39b)-(39c).

Finally, after a series of transformations based on (34b), (34e), (35)-(36), (39b)-(39c) and (48), the power allocation subproblem (47) is transformed into the following convex optimization problem:

$$\max_{\substack{\mathbf{P},\boldsymbol{\zeta_s},\boldsymbol{\zeta_m}\boldsymbol{\eta_s},\\ \boldsymbol{\eta_m},\mathbf{t_s},\mathbf{t_m},\mathbf{s_s},\mathbf{s_m}}} \sum_{s\in\mathcal{S}}\sum_{u\in U_s}\left(\rho\log\left(\zeta_{s,u}\right)+\varphi\right)+\sum_{m\in\mathcal{M}}\left(\rho\log\left(\zeta_m\right)+\varphi\right) - \tau\Gamma(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) \tag{49a}$$

$$\text{s.t. } (22c)-(22d),\ (22g),\ (42),\ (34b),(34e),(35)-(36);\ (39b)-(39c),(48) \tag{49b}$$

which can be effectively solved by advanced convex solvers, e.g., CVX [44].

In **Algorithm 2**, the proposed BCD-based method to problem (49) can be shown which replaces (32) and (39) with (46) and (49). After executing the inner-layer optimization of the P-BCD framework (i.e., Step 3 in **Algorithm 1**), the penalty parameter τ_v is iteratively updated according to Step 4 in **Algorithm 1**. The process continues until the P-BCD algorithm converges.

C. Complexity Analysis

Similar to the complexity analysis discussed in Subsection IV-C, in the following, we give the complexity of the proposed P-BCD algorithm for solving problem (44). At each

TABLE III
COMPUTATIONAL COMPLEXITY OF PROPOSED AND REFERENCE SCHEMES FOR WEB BROWSING AND VIDEO SERVICES

Algorithms	Computational Complexity
The proposed scheme for Web Browsing	$O\left(\triangle_1 \times \triangle_2 \left(\frac{\log\left((2(SU+M)+SN+N)/\xi\omega\right)}{\log(\psi)} + \frac{\log\left((5(SUN+MN)+2(SU+M)+S+1)/\xi\omega\right)}{\log(\psi)}\right)\right)$
The proposed scheme for Video Service	$O\left(\triangle_1 \times \triangle_2 \left(\frac{\log\left((3(SU+M)+SN+N)/\xi\omega\right)}{\log(\psi)} + \frac{\log\left((5(SUN+MN)+3(SU+M)+S+1)/\xi\omega\right)}{\log(\psi)}\right)\right)$
The scheme in [29] for Web Browsing (or Video Service)	$O\left(\triangle_3 \times \triangle_4 \left((SUN+MN)^2 + (SUN+MN)^2\right)\right)$

iteration of inner loop, the complexity of BCD approach is $O(\log(I_3/\xi\omega)/\log(\psi) + \log(I_4/\xi\omega)/\log(\psi))$, where $I_3 = 3(SU+M)+SN+N$ and $I_4 = 5(SUN+MN)+3(SU+M)+S+1$. Hence, the total complexity of the proposed P-BCD method is $O(\triangle_1 \times \triangle_2(\log(I_3/\xi\omega)/\log(\psi)+\log(I_4/\xi\omega)/\log(\psi)))$, where $\triangle_1$ and $\triangle_2$ are iteration times for outer layer and inner layer, respectively. Besides, the complexity of the algorithm proposed by [29] for video service is $O(\triangle_3 \times \triangle_4(2(SUN+MN)^2))$ [46], where $\triangle_3$ and $\triangle_4$ represent iteration times for outer layer and inner layer, respectively. We summarize the complexities mentioned above in **Table III**.

VI. NUMERICAL RESULT

This section presents the numerical results to evaluate the performance of the proposed scheme. The primary simulation parameters are set as follows. The radii of macro cell and small cell are set to 500 m and 20 m, respectively [9]. The maximum transmit power at MBS and SBS are set to 43 dBm and 23 dBm, respectively [32]. The pathloss models at a distance R [km] from macro cell and small cell are 128.1 + 37.6 $\log_{10}(R)$ dB and 140.7 + 36.7 $\log_{10}(R)$ dB, respectively [13]. The Rayleigh fading is considered to model the small scale fading channels between BSs and UEs. The number of subchannels is $N = 10$. The bandwidth of each subchannel is $B_n = 75$ kHz [28]. The AWGN spectral density is -174 dBm/Hz [32]. For web browsing service, MSS is set to 1460 bytes [24]. For video service, $\text{PSNR}_{1.0}$ and $\text{PSNR}_{4.5}$ are set to 30 dB and 42 dB, respectively [36]. The tolerance parameters for the proposed P-BCD algorithm are set to $\varepsilon_v = \varepsilon_l = 0.01$. The penalty parameters c is given by $c = 2$ [43], and the initial value τ_0 for penalty is set to $\tau_0 = 0.001$ [43]. In addition, the number of MUEs is set to $M = 2$, the number of SUEs in each small cell is set to $U_s = 6$, the satisfaction threshold is set to $\text{MOS}_{\min} = 1$, and the web page size (FS) for web browsing is set to 320 KB [24], unless otherwise specified. Simulation results are obtained on a computer with the Intel Core i7 9700F and 16G RAM.

A. Impact of the Web Page Sizes and Application Types

This subsection presents the impact of different web page sizes (FS) and application types on the QoE perceived by users.

In Fig. 3, we plot the MOS value versus user data rate. The performance is compared at video application and web browsing application with different FS, where $FS = \{18, 30, 50, 100, 200\}$ KB [24]. From Fig. 3, we observe that user obtains different perceived qualities even when the data rates are same. This phenomenon illustrates that the network-oriented QoS criteria, such as data rate, are not sufficient for evaluating user QoE. In fact, user QoE is also affected by other factors, such as application type and FS. Such factors are not related to network QoS but do affect user QoE. Additionally, we see that the MOS value of user browsing the web with large FS is smaller than that of user browsing the web with small FS within a certain range of data rate. Also, we observe that the MOS value of video streamer is much smaller than that of web browsing user when those users have the same data rate. These results can be explained by the fact that compared to the user who browses the web with large FS, the user who browses the web with small FS can achieve a high QoE by the lower capacity. Moreover, a larger capacity is required for video streamer in order to obtain a high QoE.

In Fig. 4, we investigate the average MOS versus the number of SUEs per small cell under web browsing application. The performance is compared at different value of web page size $FS = 320, 400$ and 500 KB [24]. The number of SBSs is set to $S = 10$. The satisfaction threshold for web browsing user is set to $\text{MOS}_{\min}^{\text{web}} = 1$. The average MOS is defined by the ratio of sum MOSs to the total number of users. From Fig. 4, we observe that the average MOS decreases as FS increases. A similar explanation can be found as described in Fig. 3. Besides, for the results obtained from NOMA schemes, we can see that the average MOS first increases with the number of SUEs per small cell. The main reason is that as the number of SUEs per small cell increases, the distances between SBSs and SUEs become closer, which contributes to the less path loss as well as better services. As a result, the value of average MOS increases. On the other hand, as the number of users per small cell becomes large, the user interference becomes severe, which diminishes the performance of average MOS. For example, for web browsing application with $FS = 400$ KB, when the number of users in each small cell is 6, the average MOS of the proposed NOMA scheme shows 4% performance degradation compared with the case the number of users per small cell is 4.

B. Impact of Different Satisfaction Thresholds

This subsection shows the simulation results of average MOS against satisfaction thresholds under web browsing and video applications. For comparison, we also consider the following benchmark solutions: (i) 'NOMA-EPA': In this scheme, the transmitted power of BS is equally distributed to each user, and then the subchannel assignment is addressed via penalty method and SPCA technique; (ii) 'NOMA-RSA': In this scheme, the subchannels are randomly assigned to users, and then the power allocation is performed via SPCA technique. In addition, we assume the same minimum satisfaction thresholds for all users.

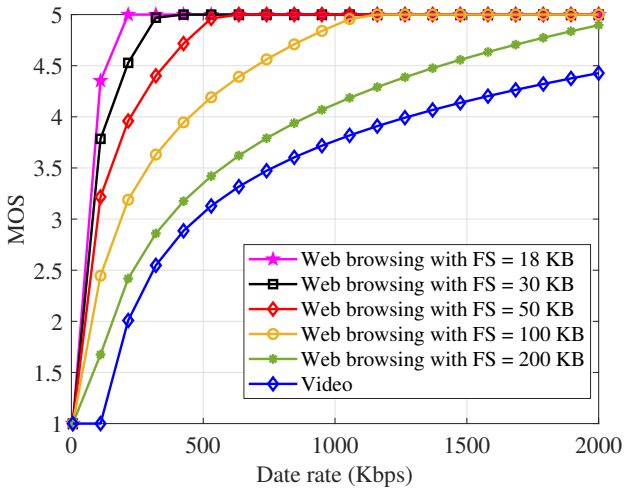


Fig. 3. MOS versus user data rate for video application and web browsing application with different FS.

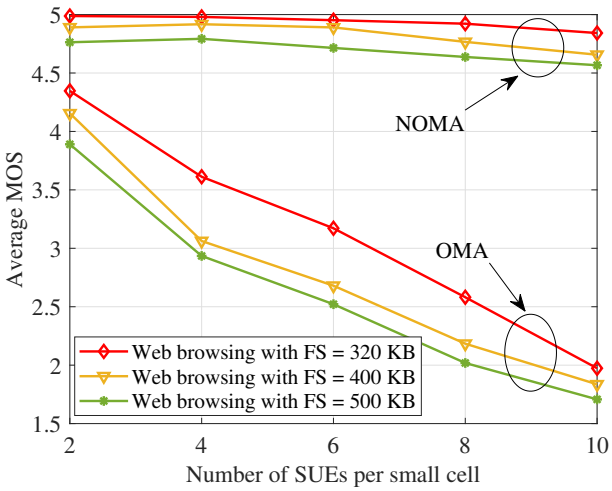


Fig. 4. Average MOS versus the number of SUEs per small cell under web browsing application with different FS.

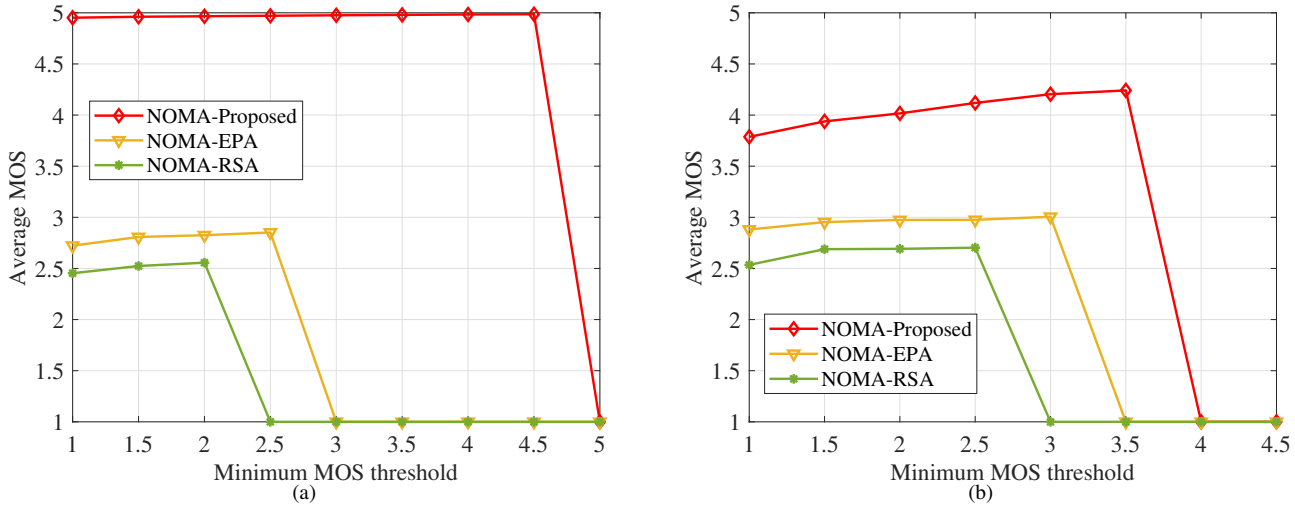


Fig. 5. Average MOS versus the minimum MOS threshold: a) web browsing application; b) video streaming application.

Figs. 5(a) and 5(b) show average MOS versus minimum MOS threshold under web browsing and video applications, respectively. The number of SBSs is set to $S = 10$. It should be noted that $\text{MOS}_{\text{min}}^{\text{web}} = 1$ or $\text{MOS}_{\text{min}}^{\text{video}} = 1$ means that there is not MOS constraint in our considered QoE optimization problem. As can be seen from Fig. 5(a), the average MOS of web users increases as the minimum MOS threshold, i.e., $\text{MOS}_{\text{min}}^{\text{web}}$, increases from 1 to 4.5. A similar phenomenon can also be observed from Fig. 5(b), where the average MOS of video users increases as $\text{MOS}_{\text{min}}^{\text{video}}$ increases from 1 to 3.5. These results imply that the proposed scheme is beneficial to improve the MOS of users. In addition, we also observe that by considering a certain level of satisfaction threshold, the average MOS can be enhanced in comparison with the cases in which there are not QoE constraints. For example, when $\text{MOS}_{\text{min}}^{\text{web}} = 3$, web users obtain about 1% improvement in average MOS compared with the case in which $\text{MOS}_{\text{min}}^{\text{web}} = 1$. When $\text{MOS}_{\text{min}}^{\text{video}} = 3$, video users also obtain about $13\% - 14\%$ improvement in average MOS compared with the case in which $\text{MOS}_{\text{min}}^{\text{video}} = 1$. The above results demonstrate that introducing a certain level of the minimum MOS threshold can lead to an improvement of the average MOS. However, from Figs. 5(a) and 5(b), we notice that as the minimum MOS threshold increases further, the achieved average MOS starts to decline and tends to 1 finally. It is pointed out that when $\text{MOS}_{\text{min}}^{\text{web}}$ or $\text{MOS}_{\text{min}}^{\text{video}}$ is large, there may exist some cases in which the minimum MOS threshold cannot be satisfied by each user. In such a case, average MOS is set to 1 for ensuring the fairness of the comparison among different approaches. These results can be explained by the fact that the more stringent the QoE demand is, the more difficult is to achieve high user satisfaction. On the other hand, the proposed method still outperforms other benchmark schemes due to its ability to derive more efficient power and subchannel allocation solution.

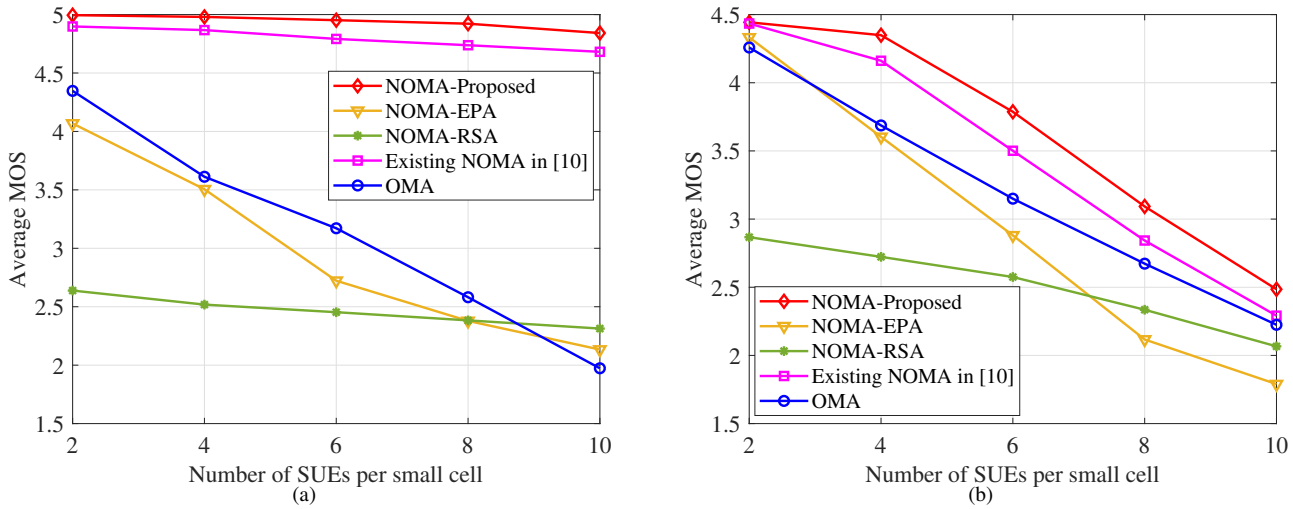


Fig. 6. Average MOS versus the number of SUEs per small cell: a) web browsing application; b) video streaming application.

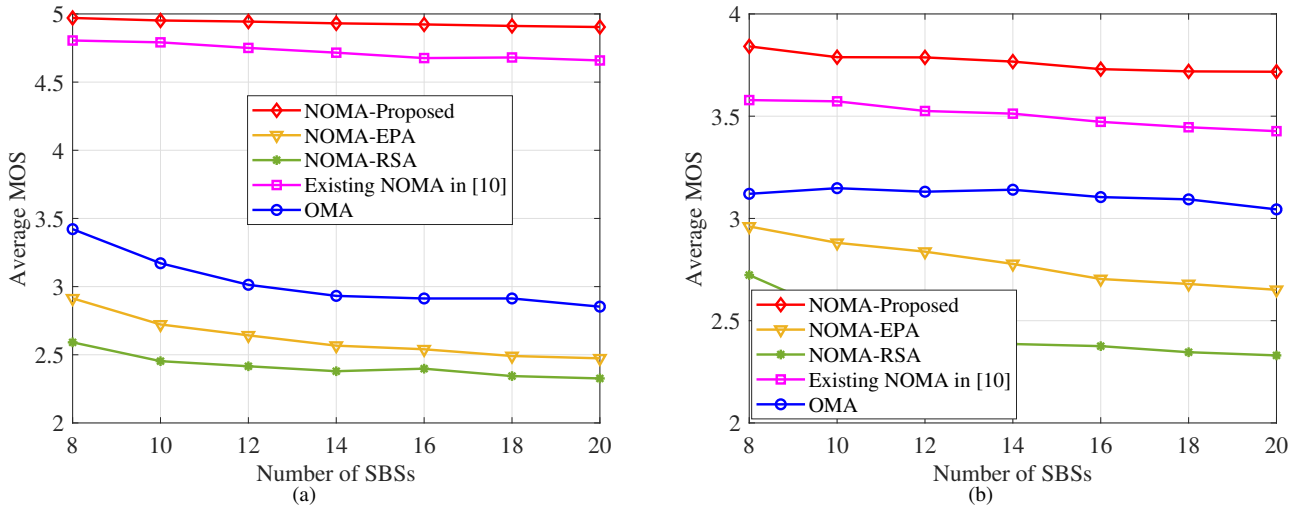


Fig. 7. Average MOS versus the number of SBSs: a) web browsing application; b) video streaming application.

C. Impact of the Number of SUEs per Small Cell

This subsection presents the numerical results of average MOS versus the number of SUEs per small cell under web browsing and video applications. To evaluate the effectiveness, we consider the following benchmark schemes: (i) 'NOMA-EPA'; (ii) 'NOMA-RSA'; (iii) Existing NOMA scheme in [10], where the subchannel and power allocation scheme is developed to maximize sum rate of all users in order to provide the QoS-aware NOMA benchmark solution; (iv) 'OMA' scheme [47], where the OMA strategy is further developed to maximize the sum MOSs of all users.

In Fig. 6(a), we investigate the web browsing service case, where the number of SBSs is $S = 10$ and the satisfaction threshold is $\text{MOS}_{\min}^{\text{web}} = 1$. As can be observed from this figure, average MOS obtained from some schemes almost remains constant at the beginning and then decreases as the number of SUEs per small cell increases, which shows a similar trend with Fig. 4. Moreover, we can find that MOS performance of the proposed scheme is always better than that of the existing QoS-aware NOMA scheme [10]. For example, when the number of SUEs per small cell is 6, the proposed scheme obtains about $3\% - 4\%$ improvement in average MOS compared with the existing scheme [10]. This is because existing QoS-based NOMA scheme does not consider other non-network-related factors affecting user QoE. As a result, conducting sum-rate maximization alone may not be able to ensure optimal user satisfaction. Additionally, average MOS of the proposed scheme significantly outperforms other two QoE-aware NOMA benchmark schemes, i.e., 'NOMA-EPA' and 'NOMA-RSA' schemes. This is because that 'NOMA-EPA' scheme does not effectively allocate transmitted power to users, and 'NOMA-RSA' scheme randomly assigns subchannel to users without taking into account CSI and other factors. As a result, average MOS value will decline.

In Fig. 6(b), we investigate the video application case, where the number of SBSs is $S = 10$ and the satisfaction threshold

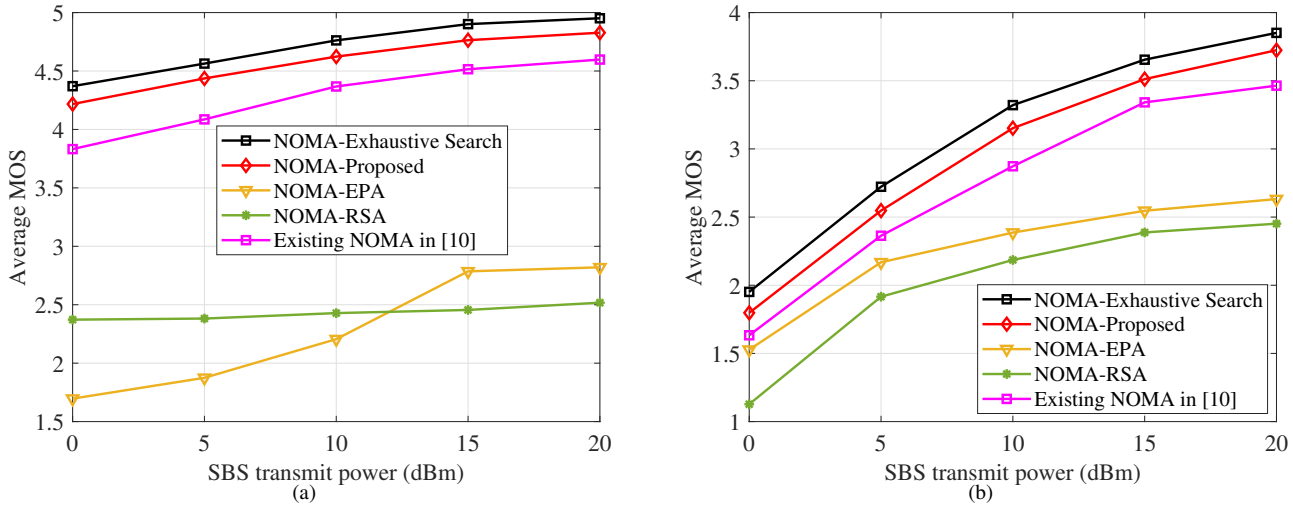


Fig. 8. Average MOS versus the SBS transmit power: a) web browsing application; b) video streaming application.

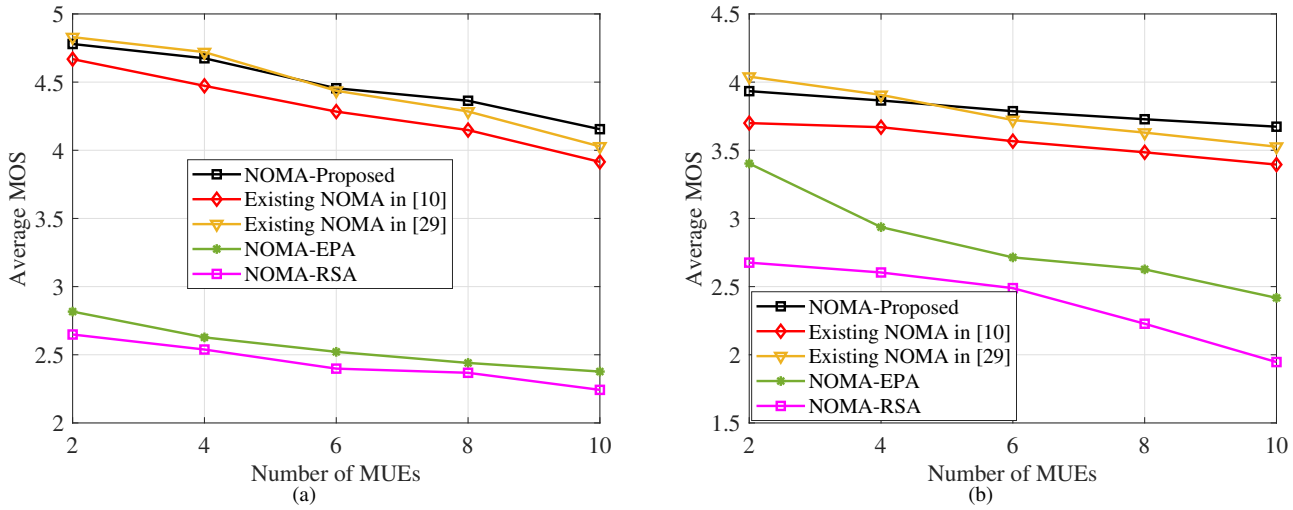


Fig. 9. Average MOS versus the number of MUEs: a) web browsing application; b) video streaming application.

is $\text{MOS}_{\text{min}}^{\text{video}} = 1$. We see that for all curves, average MOS decreases as the number of SUEs in each small cell increases. This implies that increasing the number of SUEs per small cell brings severe user interference, which limits average MOS performance of video users. However, average MOS of the proposed scheme is always better than benchmark schemes.

D. Impact of the Number of SBSs

Figs. 7(a) and 7(b) make the same comparison as Fig. 6(a) and Fig. 6(b), respectively, but from the number of SBSs perspectives. The performance of average MOS is compared via different benchmark schemes. From Fig. 7(a), we observe that when the number of SBSs increases, all curves tend to diminish. This can be explained by the effect of severe inter-cell interference between different small cells when more small cells are deployed within the macrocell. In addition, the proposed scheme is more effective than other benchmark schemes, the average MOS gap between them becomes larger as the number of SBSs increases. In Fig. 7(b), the average MOS of some schemes is almost constant and then degrades as the number of SBSs increases, which shows a similar curve trend with Fig. 7(a). Furthermore, the proposed scheme achieves substantial improvement of average MOS than OMA scheme as well as comparable NOMA schemes. This is because the proposed scheme can provide more freedom in subchannel assignment and more efficient power allocation policy, which further enhances the performance of average MOS.

E. Impact of the SBS Transmit Power

This subsection investigates the performance of the average MOS versus the maximum transmit power of each SBS. The performance is compared using the following NOMA benchmarks: (i) 'NOMA-Exhaustive Search' (over all possible choices of subchannels and transmit power). Herein, the optimal power allocation scheme with exhaustive search, would examine all possible power allocation combinations at a very small step, i.e., $\frac{P_{\text{max}}^{s}}{L}$ (or $\frac{P_{\text{max}}^{M}}{L}$), over $[0, P_{\text{max}}^{s}]$ (or $[0, P_{\text{max}}^{M}]$). We set $L = 1000$ for simulations [45]; (ii) 'NOMA-EPA'; (iii)

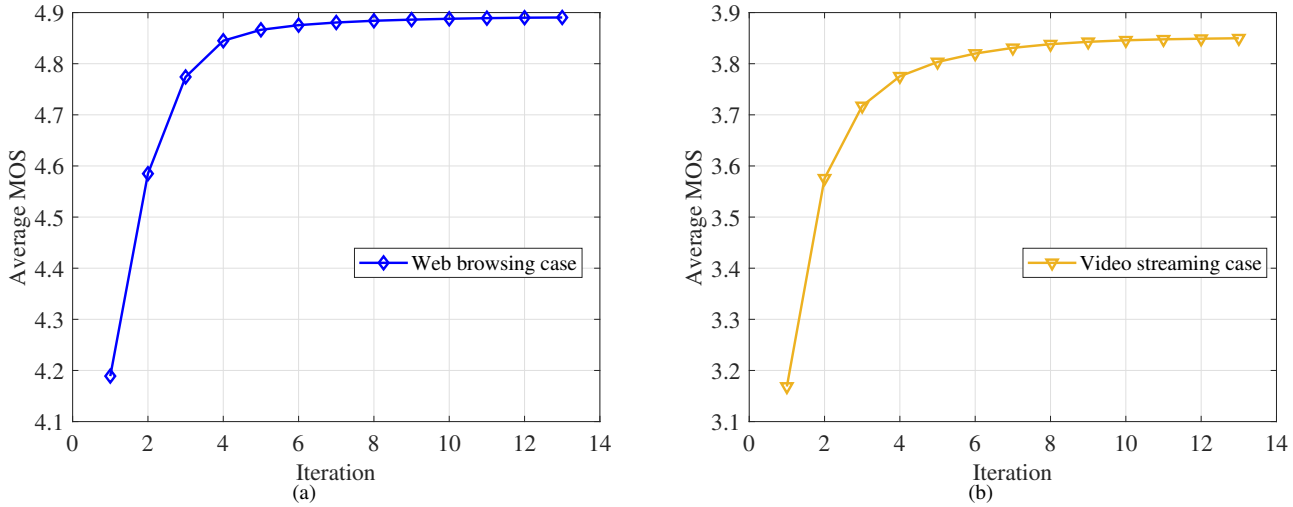


Fig. 10. Convergence behavior: a) web browsing application; b) video streaming application.

'NOMA-RSA'; (iv) Existing NOMA scheme in [10]. As can be seen from Figs. 8(a) and 8(b), the average MOS first increases from 0 dBm when SBS transmit power increases. This is because the achieved average MOS of users can be improved by efficiently allocating the transmit power and the subchannel resource through the proposed method. On the other hand, as the SBS transmit power increases further, the achieved average MOS takes on a slow growth trend. This is because the increase in SBS transmit power brings stronger interference, which degrades the average MOS performance. However, the proposed scheme still outperforms other baseline solutions in (ii)-(iv). Besides, it can be seen from Figs. 8(a) and 8(b), the proposed method can achieve a high MOS performance which is near to the optimal solution with lower complexity, compared to benchmark in (i) which can find the globally optimal solution but has exponential worst-case complexity with respect to the number of optimization variables. This result verifies the effectiveness of the proposed method.

F. Impact of the Number of MUEs

This subsection presents the numerical results of average MOS versus the number of MUEs. The performance is compared through the following NOMA benchmark schemes: (i) Existing NOMA scheme in [29], where a NOMA-based resource allocation scheme was proposed to maximize the sum MOSs of the JT-enabled SUEs, while OFDMA policy and QoS requirement were considered for the MUEs. It should be pointed out that the average MOS considered in this paper is defined by the ratio of sum MOSs of all users, including MUEs and SUEs, to the total number of users. In this context, the achievable MOSs of MUEs are also considered into the design solution of [29] for ensuring the fairness of the comparison among different schemes; (ii) Existing NOMA scheme in [10]; (iii) 'NOMA-EPA'; (iv) 'NOMA-RSA'. From Figs. 9(a) and 9(b), we can see that when the number of MUEs is small, the average MOS of [29] achieves better performance gain than that of the proposed scheme in this paper. However, as the number of MUEs becomes large, the proposed scheme achieves a higher average MOS than the scheme in [29], and meanwhile, the performance gap between the proposed scheme and the scheme in [29] tends to increase with growth of the number of MUEs. The reason is as follows. When the number of MUEs is small, the integration of JT-NOMA for SUEs is used in [29] such that the SUEs have opportunity to be served by multiple SBSs, which contributes to additional performance gain. On the other hand, when the number of MUEs becomes large, the proposed NOMA scheme in this paper achieves higher MOSs for MUEs in comparison with the results computed by the OFDMA-enabled scheme for MUEs in [29]. The performance gap tends to grow with the increase in the number of MUEs. For example, when the number of MUEs is $M = 10$, the proposed scheme takes about 1.4% (1.7%) extra computation time, but attains about 3.1% (4.2%) improvement in average MOS under web browsing case (video streaming case) compared with the reference scheme in [29]. From the above results, we can conclude that when the number of MUEs is large, the proposed scheme can achieve better average MOS with slightly high complexity than [29].

G. Convergence Behavior

This subsection presents the convergence behavior of the proposed algorithm under web browsing and video applications. The number of SBSs is set to $S = 10$. As observed from Fig. 10(a) and Fig. 10(b), the proposed iterative algorithm converges to its stable solution within few iterations, which validates the convergence properties of the proposed algorithm. This result also confirms that the proposed iterative algorithm has high practical value in NOMA HetNet.

VII. Conclusion

In this paper, we have presented the QoE-aware joint subchannel and power allocation framework for NOMA-enhanced HetNet. Specifically, two different QoE optimization problem corresponding to web browsing and video services have been formulated to maximize the sum MOSs of users.

In order to guarantee the satisfaction and fairness among NOMA users, QoE constraints on each NOMA user have also been considered on the resource allocation design. Since the formulated QoE optimization problems are complex to solve, we have proposed a P-BCD based optimization algorithm. The complexity of the proposed algorithm has also been discussed. Simulation results have demonstrated that the non-network-related factors, such as application type and web page size, have a significant impact on the user QoE. As a result, considering network-related QoS criteria alone is not sufficiently reliable to measure user QoE. In addition, the proposed QoE-aware resource allocation scheme can achieve competitive QoE performance compared to existing NOMA schemes. For the future work, the proposed QoE-aware resource allocation framework can be extended to other more applications by employing corresponding MOS models into the proposed method.

APPENDIX A
CONVERGENCE PROOF OF THE ALGORITHM 2

Define $\Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P})$ as the objective value of penalized problem (30) with fixed penalty value τ_v. Define $\Psi^{lb,l}_{pow,\tau_v}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P}) = \Psi^{l}_{pow,\tau_v}$, where Ψ^{l}_{pow,τ_v} is the objective value of (39) based on $\left\{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P}\right\}$. First, for given $\left\{\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l\right\}$, the optimal solution of (31) can be obtained by step 3 of **Algorithm 2**, and thus we have

$$\Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P}) \leq \Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P}) \tag{50}$$

Second, for given $\left\{\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}\right\}$, $\left\{\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l\right\}$, and $\mathbf{P}_l$ in step 4 of **Algorithm 2**, we can obtain

$$\Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P}) \leq \Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}, \mathbf{P}) \tag{51}$$

Third, for given $\left\{\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}\right\}$, $\left\{\boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}\right\}$, and $\mathbf{P}_l$ in step 5 of **Algorithm 2**, we can obtain

$$\begin{aligned}
&\Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}, \mathbf{P}) \\
&\quad \overset{(a)}{\leq} \Psi^{lb,l}_{pow,\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}, \mathbf{P}_{l+1}) \\
&\quad \overset{(b)}{\leq} \Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}, \mathbf{P}_{l+1})
\end{aligned} \tag{52}$$

where (a) holds since problem (33) can be optimally addressed with solution $\mathbf{P}_{l+1}$ for given $\left\{\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}\right\}$ and $\left\{\boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}\right\}$ in step 5 of **Algorithm 2**. (b) holds since problem (35) offers a lower bound to its original problem (33) at $\mathbf{P}_{l+1}$. In fact, the feasible set of problem (39) is always a subset of that of problem (33). The inequality in (52) indicates that objective value of (33) is still non-decreasing in each inner iteration, although only an approximate problem (39) is solved for obtaining power allocation solution. Furthermore, based on (50)-(52), one can know

$$\Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l}, \tilde{\boldsymbol{\beta}}_{l}, \boldsymbol{\alpha}_{l}, \boldsymbol{\beta}_{l}, \mathbf{P}_{l}) \leq \Psi_{\tau_v}(\tilde{\boldsymbol{\alpha}}_{l+1}, \tilde{\boldsymbol{\beta}}_{l+1}, \boldsymbol{\alpha}_{l+1}, \boldsymbol{\beta}_{l+1}, \mathbf{P}_{l+1}) \tag{53}$$

which implies that for each inner iteration of **Algorithm 2**, the objective value of (30) is non-decreasing. Due to the constraints existed in both total power and subchannel assignment, the objective value is thus bounded. Therefore, the convergence of **Algorithm 2** can be guaranteed.

REFERENCES

[1] Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022 white paper. 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html/

[2] F. Fang, G. Ye, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient joint user association and power allocation in a heterogeneous network," *IEEE Trans. Wireless Commun.*, 2020 (Early access).

[3] L. P. Qian, Y. Wu, B. Ji, and X. S. Shen, "Optimal ADMM-based spectrum and power allocation for heterogeneous small-cell networks with hybrid energy supplies," *IEEE Trans. Mobile Comput.*, 2019 (Early access).

[4] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple accesss–A novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, Apr. 2017.

[5] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. 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.

[6] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart. 2018.

[7] M. Moltafet, P. Azmi, N. Mokari, M. R. Javan, and A. Mokdad, "Optimal and fair energy efficient resource allocation for energy harvesting-enabled-PD-NOMA-based HetNets," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2054–2067, Mar. 2018.

[8] A. Celik, M. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M. Alouini, "Distributed cluster formation and power-bandwidth allocation for imperfect NOMA in DL-HetNets," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1677–1692, Feb. 2019.

[9] F. Fang, J. Cheng, and Z. Ding, "Joint energy efficient subchannel and power optimization for a downlink NOMA heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1351–1364, Feb. 2019.

[10] X. Chu, H. Zhang, W. Huangfu, W. Liu, Y. Ren, J. Dong, and K. Long, "Subchannel assignment and power optimization for energy-efficient NOMA heterogeneous network," in *IEEE GLOBECOM*, 2019, pp. 1–6.

[11] P. Swami, V. Bhatia, S. Vuppala, and T. Ratnarajah, "User fairness in NOMA-HetNet using optimized power allocation and time slotting," *IEEE Syst. J.*, pp. 1–10, 2020 (Early access).

[12] T. M. Nguyen, W. Ajib, and C. Assi, "A novel cooperative non-orthogonal multiple access (NOMA) in wireless backhaul two-tier HetNets," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4873–4887, Jul. 2018.

[13] A. J. Muhammed, Z. Ma, Z. Zhang, P. Fan, and E. G. Larsson, "Energy-efficient resource allocation for NOMA based small cell networks with wireless backhauls," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3766–3781, Mar. 2020.

[14] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 629–646, 1st Quart. 2018.

[15] Q. Wu, Z. Du, P. Yang, Y. Yao, and J. Wang, "Traffic-aware online network selection in heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 381–397, Jan. 2016.

[16] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[17] 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, no. 8, pp. 6325–6343, Aug. 2016.

[18] M. S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Resource optimization and power allocation in in-band full duplex-enabled non-orthogonal multiple access networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2860–2873, Dec. 2017.

[19] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1126–1165, 2nd Quart. 2015.

[20] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart. 2016.

[21] D. Yuan, M. Song, Y. Teng, D. Ma, X. Wang, and G. Lu, "QoE-oriented resource allocation for multiuser-multiservice femtocell networks," *China Communications*, vol. 12, no. 10, pp. 27–41, Oct. 2015.

[22] J. Chen, Y. Deng, J. Jia, M. Dohler, and A. Nallanathan, "Cross-Layer QoE Optimization for D2D Communication in CR-Enabled Heterogeneous Cellular Networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 719–734, Dec. 2018.

[23] H. Abarghouyi, S. M. Razavizadeh, and E. Björnson, "QoE-aware beamforming design for massive MIMO heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8315–8323, Sep. 2018.

[24] M. Rugelj, U. Sedlar, M. Volk, J. Sterle, M. Hajdinjak, and A. Kos, "Novel cross-layer QoE-aware radio resource allocation algorithms in multiuser OFDMA systems," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3196–3208, Sep. 2014.

[25] V. F. Monteiro, D. A. Sousa, T. F. Maciel, F. R. M. Lima, E. B. Rodrigues, and F. R. P. Cavalcanti, "Radio resource allocation framework for quality of experience optimization in wireless networks," *IEEE Netw.*, vol. 29, no. 6, pp. 33–39, Nov. 2015.

[26] S. He and W. Wang, "A QoE-optimized power allocation scheme for non-orthogonal multiple access wireless video services," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–6.

[27] M. Zhang, H. Lu, F. Wu, and C. W. Chen, "NOMA-based scalable video multicast in mobile networks with statistical channels," *IEEE Trans. Mobile Comput.*, 2020 (Early access).

[28] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "QoE-based resource allocation for multi-cell NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6160–6176, Sep. 2018.

[29] H. Zarini, A. Khalili, H. Tabassum, and M. Rasti, "Joint transmission in QoE-driven backhaul-aware MC-NOMA cognitive radio network," in *2020 IEEE Global Communications Conference (GLOBECOM)*, Aug. 2020, pp. 1–6.

[30] H. Zhang, M. Feng, K. Long, G. K. Karagiannidis, V. C. M. Leung, and H. V. Poor, "Energy efficient resource management in SWIPT enabled heterogeneous networks with NOMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 835–845, Nov. 2020.

[31] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[32] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in hetnets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825–5837, Sep. 2017.

[33] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[34] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, "QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems," *Comput. Commun.*, vol. 33, no. 5, pp. 571–582, 2010.

[35] 3GPP TR 36.912, "Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced) document v11.0.0," *3rd Gen. Partner. Proj., Sophia Antipolis, France, TS36*, Sep. 2012.

[36] L. U. Choi, M. T. Ivrlac, E. Steinbach, and J. A. Nossek, "Sequence-level models for distortion-rate behaviour of compressed video," in *IEEE Intern. Conf. on Image Processing*, vol. 2, 2005, pp. II – 486–9.

[37] A. Rubinov, H. Tuy, and H. Mays, "An alogrithm for monotonic global optimization problems," *Optimization*, vol. 49, no. 3, pp. 205–221, 2001.

[38] H. Tuy, F. Al-Khayyal, and P. T. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization*. Springer, 2005, pp. 39–78.

[39] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[40] M. Hong, M. Razaviyayn, Z. Luo, and J. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.

[41] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Opt.*, vol. 47, no. 1, pp. 29–51, 2010.

[42] Z. Lu, Y. Zhang, and X. Li, "Penalty decomposition methods for rank minimization," *Optimization Methods and Software*, vol. 30, no. 3, pp. 531–558, 2015.

[43] Y. Cai, F. Cui, Q. Shi, M. Zhao, and G. Y. Li, "Dual-UAV-enabled secure communications: Joint trajectory design and user scheduling," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1972–1985, Sep. 2018.

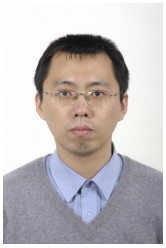
[44] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," Apr. 2011. [Online]. Available: http://cvxr.com/cvx

[45] N. Mokari, F. Alavi, S. Parsaeefard, and T. Le-Ngoc, "Limited-feedback resource allocation in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2509–2521, April. 2016.

[46] A. Khalili, S. Akhlaghi, H. Tabassum, and D. W. K. Ng, "Joint user association and resource allocation in the uplink of heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 804–808, Jun. 2020.

[47] L. Zhou, X. Hu, E. C. . Ngai, H. Zhao, S. Wang, J. Wei, and V. C. M. Leung, "A dynamic graph-based scheduling and interference coordination approach in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3735–3748, May. 2016.

Liangyu Chen received the M.S. degree from Inner Mongolia University, Hohhot, China. He is currently pursuing a Ph.D. degree with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China. His current research interests include non-orthogonal multiple access, deep reinforcement learning, resource allocation, and optimization.



Bo Hu (S'04–M'07) received his PhD degree of communications and information systems in 2006, from Beijing University of Posts and Telecommunications (BUPT), China. Currently, he is a professor in the State Key Laboratory of Networking and Switching Technology, BUPT. He has published over 80 research papers in international journals & conferences and co-authored 3 books. He is active in the ITU-T SG13 for international standards and focus on the mobility management and machine learning for IMT-2020(5G) & beyond. His research interests include: integrated satellite and terrestrial mobile communication system for B5G and 6G, network artificial intelligence and mobility management & control.

Shanzhi Chen [F'20] received the bachelor's degree from Xidian University in 1991 and the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 1997. He joined the Datang Telecom Technology and Industry Group and the China Academy of Telecommunication Technology (CATT) in 1994, and has been serving as the EVP of Research and Development since 2008. He is currently the Director of the State Key Laboratory of Wireless Mobile Communications, CATT, where he conducted research and standardization on 4G TD-LTE and 5G. He has authored and co-authored four books [including the well-known textbook Mobility Management: Principle, Technology and Applications (Springer Press)], 17 book chapters, more than 100 journal papers, 50 conference papers, and over 50 patents in these areas. He has contributed to the design, standardization, and development of 4G TD-LTE and 5G mobile communication systems. His current research interests include 5G mobile communications, network architectures, vehicular communication networks, and Internet of Things. He served as a member and a TPC Chair of many international conferences. His achievements have received multiple top awards and honors by China central government, especially the Grand Prize of the National Award for Scientific and Technological Progress, China, in 2016 (the highest Prize in China). He is the Area Editor of the IEEE INTERNET OF THINGS, the Editor of the IEEE NETWORK, and the Guest Editor of the IEEE WIRELESS COMMUNICATIONS, the IEEE Communications Magazine, and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Jianpeng Xu received the M.S. degree from Inner Mongolia University, Hohhot, China, in 2018. He is currently working towards the Ph.D. degree at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China. His current research interests include millimeter-wave communications, resource allocation and network architecture for high-mobility broadband wireless communications.

Guixian Xu received his Ph.D. degree in communications and information systems from Beijing University of Posts and Telecommunications (BUPT), China, in 2017. From June 2011 to August 2012, he was a TD-LTE Hardware Test Engineer with Datang Mobile, Beijing, China. Since October 2012, he has been an intern with CBC/XEV of Ericsson China and the State Key Laboratory of Wireless Mobile Communications, China Academy of Telecommunications Technology (CATT), Beijing. From 2015 to 2016, he was a visiting Ph.D. Student with the National Tsing Hua University, Hsinchu, Taiwan. From 2017 to 2018, he was a postdoctoral with connectivity section, Department of Electronic Systems, Aalborg University, Denmark. rrently, he is working at the Department of Engineering as a postdoctoral researcher, Aarhus University, Denmark. His research interests are in 5G beyond, Internet of Things (IoT), Machine learning for wireless communication networks.