# Computation Efficiency Optimization for Millimeter-Wave Mobile Edge Computing Networks with NOMA

Xiangbin Yu, Fangcheng Xu, Jiali Cai, Xiao-yu Dang, Kezhi Wang

***Abstract*—In this paper, the millimeter-wave (mmWave) communications and non-orthogonal multiple access (NOMA) are exploited for mobile edge computing (MEC) networks to improve the performance of task offloading. Aiming at improving the computation efficiency (CE) and ensuring the fairness among users, we study the CE optimization for mmWave-MEC with NOMA, where both the analog beamforming (ABF) and hybrid beamforming (HBF) architectures under the partial offloading mode are considered. Firstly, according to the max-min fairness criterion, the CE optimization problem is formulated to jointly optimize the ABF at the base station and the local resource allocation of each user in mmWave-MEC with ABF. An efficient algorithm based on the penalized successive convex approximation is proposed to solve this non-convex problem. Then, the max-min CE optimization problem in mmWave-MEC with HBF is studied, where the joint design of the HBF at the BS and the local resource allocation of each user is carried out. By using the penalty function and the inexact block coordinate descent method, a feasible optimization algorithm is developed to tackle this challenging problem. Simulation results verify the convergence of the proposed algorithms and show that the proposed resource allocation schemes can improve the system CE effectively, and the mmWave-MEC with HBF scheme can obtain higher CE than that with ABF scheme. Besides, the NOMA scheme exhibits superior performance over the conventional orthogonal multiple access scheme in terms of CE.**

***Index Terms*—Mobile edge computing, millimeter-wave communications, non-orthogonal multiple access, computation efficiency, resource allocation.**

## I. Introduction

RECENTLY, the technologies in mobile communications and Internet of Things have been developed rapidly and widely applied in various fields of our daily life. Most of new applications (such as augmented reality, virtual reality, etc.) require real-time processing and computation, which may be challenge for resource-constrained mobile devices to provide expected quality of service [1]. Mobile edge computing (MEC) has been considered as a promising technology to deal with these challenges. In MEC, the decentralized computing servers are deployed at the edge of wireless networks to provide computing services for nearby mobile devices.

In MEC networks, the computational tasks can be offloaded to the edge servers that have available computational resources. However, the offloading actions are limited by the available spectrum resources. Due to the rich spectrum resources, millimeter-wave (mmWave) communications have been recognized as an attractive solution for increasing the above-mentioned applications. In mmWave communications, since the number of radio frequency chains (RFCs) is usually much smaller than that of antennas, multiple access is still a crucial issue that affects the spectrum utilization and the number of accessing users. Therefore, it is a feasible solution to apply non-orthogonal multiple access (NOMA) in mmWave communications to form the so-called mmWave-NOMA communication system [2]. As a promising multiple access technology, NOMA can exhibit superior performance than the conventional orthogonal multiple access (OMA). By multiplexing the power-domain, NOMA can realize the efficient utilization of one orthogonal resource block to improve the number of accessing users, spectral efficiency and energy efficiency [3]. Moreover, the users in mmWave communications are highly correlated, which is conducive to the integration of NOMA. Owing to these unique advantages, mmWave-NOMA has great potential to support ultra-high bandwidth services and massive connectivity [2] [3].

With the development of MEC technology, the integration of MEC and other emerging mobile communication technologies has become one of research trends in recent years, such as MEC combines with mmWave communications [4], MEC combines with NOMA [5]. The potential of MEC technology has stimulated the extensive efforts of academia and industry in various fields [6]–[11]. Several literatures have comprehensively reviewed the standardization work in MEC, and discussed the existing problems, challenges and future research directions [6]–[8]. The previous contributions [9] and [10] respectively studied the latency optimization problems for the cases of single-user and multi-user in the mmWave-MEC system and exhibited the advantages of combining MEC with mmWave communications. Reference [11] analyzed the impact of NOMA on MEC and demonstrated that NOMA can reduce the latency of computation offloading and energy consumption by the theoretical and simulation results.

With the development of green communication, improving resource utilization is becoming more and more important for MEC [12]–[14]. In [12], computation efficiency (CE), i.e., the ratio of the computation bits (CBs) to the energy consumption, was proposed to evaluate the efficiency of computation and communication per joule in MEC. Moreover, [13] studied the maximization of the sum CE in MEC based on

X.Yu, F.Xu, J.Cai and X.Dang are with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, China. e-mail:(yxbxwy@gmail.com). K.Wang is with Department of Computer and Information Sciences, Northumbria University, United Kingdom

This work is supported in part by National Natural Science Foundation of China (61971220, 62031017, 61971221), and Open Research Fund of State Key Laboratory of Millimeter Waves of Southeast University (K202215).

orthogonal frequency division multiple access (FDMA), where the partial offloading mode and the binary offloading mode were considered and closed-form solutions of the optimal subchannel and power allocation scheme were derived for given Lagrange multipliers. Furthermore, [14] extended the CE maximization framework under the max-min fairness criterion to the wireless-powered MEC with time division multiple access (TDMA) and NOMA.

Nevertheless, the above contributions only focused on the traditional communication frequency band. Against this background, based on the mmWave communication, by exploiting the NOMA scheme, we study the CE optimization of mmWave-MEC for improving the system CE and guaranteeing the user fairness. Specifically, we consider the analog beamforming (ABF) and hybrid beamforming (HBF) architectures under the partial offloading mode with NOMA to design the corresponding resource allocation schemes. The main contributions of this paper are summarized as follows.

1) The computation-efficient resource allocation scheme in mmWave-MEC with ABF (mmWave-MEC-ABF) is firstly studied. For the purpose of improving CE and guaranteeing user fairness, the max-min CE optimization under the partial offloading mode with NOMA is formulated to maximize the minimum CE of the users subject to the minimum computation-bit rate and the maximum power consumption constraints, where the ABF at the base station (BS) and the local resource allocation of each user are jointly optimized. By introducing auxiliary variables, this non-convex problem is transformed into an equivalent form, which is easier to be addressed. Accordingly, an efficient CE optimization algorithm based on the penalized successive convex approximation (S-CA) method is proposed to obtain a local optimal solution and the corresponding computation-efficient resource allocation scheme.

2) Then, the computation-efficient resource allocation scheme in mmWave-MEC with HBF (mmWave-MEC-HBF) is further studied. According to the max-min fairness criterion, the CE optimization problem under the partial offloading mode is established to jointly optimize the HBF at the BS and the local resource allocation of each user. Considering that the optimization variables are highly coupled, this challenging problem is transformed into the penalty form by using the penalty function method. Since the problem has block structure, a feasible CE optimization algorithm is proposed to obtain the computation-efficient resource allocation scheme by means of the inexact block coordinate descent (IBCD) method combined with majorization-minimization (MM) and SCA algorithms.

3) Extensive simulation results are presented to evaluate the validity of the proposed CE optimization frameworks. The results first verify the convergence of the proposed algorithms, then demonstrate that the proposed computation-efficient resource allocation schemes can achieve good performance, and the system with HBF scheme can attain higher CE than that with ABF scheme. Moreover, NOMA scheme can boost CE significantly compared with the conventional OMA scheme.

Notations: The upper-case and lower-case boldface letters denote matrices and vectors, respectively. $(\cdot)^*$, $(\cdot)^{\rm T}$, $(\cdot)^{\rm H}$ and $(\cdot)^\dagger$ denote conjugation, transpose, Hermitian transpose, and pseudo inversion, respectively. $\|\cdot\|$, $\|\cdot\|_{\rm F}$, and $\|\cdot\|_\infty$ denote the 2-norm, Frobenius norm and infinite norm, respectively. $|\cdot|$ denotes the absolute value of a real scalar or the modulus of complex scalar. $[\cdot]_i$ is the $i$-th entry of a vector. $[\cdot]_{i,j}$ is the $i$-th row and $j$-th column entry of a matrix. $\lambda_{\max}(\mathbf{A})$ denotes the maximum eigenvalue of a Hermitian matrix $\mathbf{A}$. $\mathcal{CN}(a,b)$ denotes the complex Gaussian distribution with the mean $a$ and the variance $b$. $\mathrm{U}[a,b]$ denotes the uniform distribution in the range of $a$ to $b$. $\mathbb{C}^{A\times B}$ denotes the complex matrix with the size $A\times B$. $\mathrm{Re}\{\cdot\}$ denotes the real part of a complex number. $|\mathcal{A}|$ denotes the number of elements in the set $\mathcal{A}$. $\mathcal{A}\backslash\mathcal{B}$ denotes the different set of the sets $\mathcal{A}$ and $\mathcal{B}$.

## II. System Model of mmWave-MEC with ABF

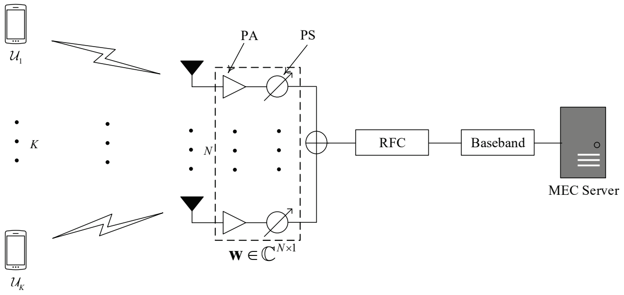


Fig. 1: System model of mmWave-MEC-ABF.

As shown in Fig. 1, the proposed mmWave-MEC-ABF system model consists of $N_{\rm U}$ single-antenna users and a mmWave BS connected to a high-performance MEC server, where the BS adopts the ABF architecture and is equipped with $N$ antennas, one RFC, $N$ power amplifiers (PAs) and $N$ phase shifters (PSs), in which each antenna is connected to the same RFC via the corresponding PA and PS. Accordingly, the ABF vector $\mathbf{w}\in\mathbb{C}^{N\times 1}$ of the BS is constricted by the constant modulus (CM) constraint, i.e., $|[\mathbf{w}]_n| = 1/\sqrt{N}, \forall n\in\mathcal{N}\triangleq\{1,\ldots,N\}$ [15] [16]. Each user can upload its computation task to the MEC server via the block-fading mmWave channel with the narrow bandwidth $B$ and the coherent time $T_{\rm C}$. $\mathbf{n}\in\mathbb{C}^{N\times 1}\sim\mathcal{CN}(\mathbf{0},\sigma^2\mathbf{I}_N)$ is the Gaussian white noise vector at the BS, in which $\sigma^2 = n_0 B$ and $n_0$ is the single-side power spectral density. Define the set $\mathcal{U}\triangleq\{1,\ldots,N_{\rm U}\}$, then the $k$-th user ($k\in\mathcal{U}$) is denoted by $\mathcal{U}_k$.

Assuming that the BS adopts the uniform linear array with half-wavelength spacing, the mmWave channel $\mathbf{h}_k\in\mathbb{C}^{N\times 1}$ between $\mathcal{U}_k$ and the BS can be modeled as [17]

$$\begin{aligned}&\mathbf{h}_k = \mathbf{h}_k^{\rm LOS} + \mathbf{h}_k^{\rm NLOS},\\&\mathbf{h}_k^{\rm LOS} = \vartheta^{\rm LOS}\sqrt{\beta_k^{\rm LOS}}\xi_k^{\rm LOS}\mathbf{a}(N,\theta_k^{\rm LOS}),\\&\mathbf{h}_k^{\rm NLOS} = \vartheta_k^{\rm NLOS}\sqrt{\beta_k^{\rm NLOS}}\sum_{l=1}^{L_k}\xi_{k,l}^{\rm NLOS}\mathbf{a}(N,\theta_{k,l}^{\rm NLOS}),\end{aligned}\tag{1}$$

where $\vartheta_k^{\rm LOS} = \sqrt{N}$, $\beta_k^{\rm LOS} = (c/(4\pi f_c))^2 d_k^{-\alpha^{\rm LOS}}$, $\alpha_k{}^{\rm LOS}$, $\xi_k^{\rm LOS}\sim\mathcal{CN}(0,1)$ and $\theta_k^{\rm LOS}\sim\mathrm{U}[0,2\pi]$ are the normalized coefficient, average path loss, path loss exponent, complex gain and angle of arrival (AoA) of the line-of-sight (LOS) path, respectively, $\vartheta_k^{\rm NLOS} = \sqrt{N/L_k}$, $\beta_k^{\rm NLOS} = (c/(4\pi f_c))^2 d_k^{-\alpha^{\rm NLOS}}$,

$\alpha_k{}^{\mathrm{NLOS}}$, $\xi_{k,l}^{\mathrm{NLOS}}\sim\mathcal{CN}(0,1)$ and $\theta_{k,l}\sim\mathrm{U}[0,2\pi]$ are the normalized coefficient, average path loss, path loss exponent, complex gain and AoA of the $l$-th non-line-of-sight (NLOS) path, respectively, $L_k$ is the number of the NLOS paths, $d_k$ is the distance between $\mathcal{U}_k$ and the BS, $c$ is the light speed, $f_c$ is the mmWave carrier frequency, and $\mathbf{a}(\cdot,\cdot)$ is the normalized steering vector function defined as

$$\mathbf{a}(N,\theta)=\tfrac{1}{\sqrt{N}}\left[1,e^{j\pi\sin\theta},\ldots,e^{j\pi(N-1)\sin\theta}\right]^{\mathrm{T}}. \quad (2)$$

Without loss of generality, it is supposed that $\|\mathbf{h}_1\|\geq\ldots\geq\|\mathbf{h}_K\|$.

The partial offloading mode is employed, i.e., the computational task of each user can be divided into two parts, in which one part is computed locally and another part is uploaded to the MEC server [12]–[14]. Specifically, the frame under this mode is composed of three stages and its duration $T$ is set as $T<T_C$. At the first stage, the uplink NOMA protocol is used for the transmission from the users to the BS. At the second stage, the BS decodes the signal of each user and the MEC server computes the uploaded tasks of users. At the final stage, the BS feedbacks the computed results to users. Following the assumption in [12]–[14], the time of the first stage can be approximated as $T$, since the MEC server has much stronger computation capability than users and the data size of computed results is relatively small. For the local computing of $\mathcal{U}_k$, the number of CBs $L_k^{\mathrm{loc}}$ and the corresponding energy consumption $E_k^{loc}$ can be expressed as [12]–[14]

$$L_k^{\mathrm{loc}}=Tf_k/C_k, \quad (3a)$$

$$E_k^{\mathrm{loc}}=T\xi_k f_k^3+TP_{k,c}, \quad (3b)$$

where $C_k$, $f_k$, $\xi_k$ and $P_{k,c}$ represent the CPU cycles per bit, CPU frequency, CPU chip coefficient and fixed circuit power consumption of $\mathcal{U}_k$, respectively.

With the uplink NOMA protocol, the received signal at the BS after the processing of ABF can be expressed as

$$y=\sum_{k=1}^{K}\mathbf{w}^{\mathrm{H}}\mathbf{h}_k\sqrt{p_k}x_k+\mathbf{w}^{\mathrm{H}}\mathbf{n}, \quad (4)$$

where $x_k\sim\mathcal{CN}(0,1)$ and $p_k$ denote the transmission signal and the transmission power of $\mathcal{U}_k$, respectively.

To deal with the superposed signal in (4), the BS adopts the successive interference cancellation (SIC) to decode the signal of each user. In general, the SIC decoding order of the uplink NOMA is the descending order of the effective channel gains of the users [3]. Therefore, when designing $\mathbf{w}$, the following SIC decoding constraint can be considered as

$$\left|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}\right|^2\geq\ldots\geq\left|\mathbf{h}_K^{\mathrm{H}}\mathbf{w}\right|^2, \quad (5)$$

where the effective channel gain of $\mathcal{U}_k$ is given by $\left|\mathbf{w}^{\mathrm{H}}\mathbf{h}_k\right|^2=\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}\right|^2$. According to the above decoding order and utilizing the SIC, the signal-to-interference-and-noise ratio (SINR) of $\mathcal{U}_k$ can be attained as

$$\mathrm{SINR}_k=\frac{p_k\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}\right|^2}{\sum\limits_{i=k+1}^{K}p_i\left|\mathbf{h}_i^{\mathrm{H}}\mathbf{w}\right|^2+\sigma^2}. \quad (6)$$

Accordingly, the achievable rate for the upload computing of $\mathcal{U}_k$ is given by $\bar{R}_k=B\log_2(1+\mathrm{SINR}_k)$. Based on this, the number of CBs and the corresponding energy consumption can be respectively expressed as

$$L_k^{\mathrm{off}}=T\bar{R}_k, \quad (7a)$$

$$E_k^{\mathrm{off}}=T\zeta_k p_k, \quad (7b)$$

where $\zeta_k$ is the PA coefficient of $\mathcal{U}_k$ [16].

Therefore, according to the definition of CE in [12]–[14], the CE of $\mathcal{U}_k$ in the mmWave-MEC-ABF is written as

$$\eta_k=\frac{L_k^{\mathrm{loc}}+L_k^{\mathrm{off}}}{E_k^{\mathrm{loc}}+E_k^{\mathrm{off}}}=\frac{B\log_2(1+\mathrm{SINR}_k)+f_k/C_k}{\zeta_k p_k+\xi_k f_k^3+P_{k,c}}. \quad (8)$$

## III. CE Optimization for mmWave-MEC with ABF

In this section, we study the CE optimization based on the max-min fairness criterion for mmWave-MEC-ABF and propose an efficient iterative algorithm to solve the non-convex CE optimization problem.

### A. Problem Formulation and Transformation

In order to improve the CE of the mmWave-MEC-ABF and ensure user fairness, the max-min fairness criterion can be used [14]. Following this criterion, the max-min CE optimization problem can be formulated as

$$\begin{aligned}
&(\mathrm{P1}):\\
&\max_{\{\mathbf{w},p_k,f_k\}}\quad \min_{k\in\mathcal{K}}\{\eta_k\}\\
&\quad\text{s.t.}\qquad C_1:\left|[\mathbf{w}]_n\right|=1/\sqrt{N},\forall n\in\mathcal{N},\\
&\qquad\qquad\ \ C_2:\left|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}\right|^2\geq\cdots\geq\left|\mathbf{h}_K^{\mathrm{H}}\mathbf{w}\right|^2,\\
&\qquad\qquad\ \ C_3:B\log_2(1+\mathrm{SINR}_k)+\tfrac{f_k}{C_k}\geq R_k^{\min},\forall k\in\mathcal{K},\\
&\qquad\qquad\ \ C_4:\zeta_k p_k+\xi_k f_k^3+P_{k,c}\leq P_k^{\max},\forall k\in\mathcal{K},\\
&\qquad\qquad\ \ C_5:0\leq f_k\leq f_k^{\max},\forall k\in\mathcal{K},\\
&\qquad\qquad\ \ C_6:P_k^{\min}\leq p_k\leq P_k^{\max},\forall k\in\mathcal{K},
\end{aligned} \quad (9)$$

where $C_1$ denotes the CM constraint, $C_2$ denotes the SIC decoding constraint, $C_3$ denotes the computation-bit rate constraint, $R_k^{\min}$ is the minimum computation-bit rate of $\mathcal{U}_k$, $C_4$ denotes the power consumption constraint, $P_k^{\max}$ is the maximum power consumption of $\mathcal{U}_k$, $C_5$ denotes the CPU frequency constraint, $f_k^{\max}$ is the maximum CPU frequency of $\mathcal{U}_k$, $C_6$ denotes the transmission power constraint, $P_k^{\min}$ is the minimum transmission power of $\mathcal{U}_k$, which is close to 0.

It can be found that the CE optimization problem (P1) is a non-smooth and non-convex fractional optimization problem, and the number of the real optimization variables is $2(N+K)$. Since $N$ is large at the BS, the complexity of directly searching the global optimal solution is extremely high. Therefore, we will design an efficient CE optimization algorithm with polynomial computation complexity to find the suboptimal solution of the problem (P1). To this end, an equivalent form of the CM constraint $C_1$ in the problem (P1) can be derived as

$$\begin{aligned}
&C_{1,a}:\left|[\mathbf{w}]_n\right|\leq 1/\sqrt{N},\forall n\in\mathcal{N},\\
&C_{1,b}:\left\|\mathbf{w}\right\|^2\geq 1.
\end{aligned} \quad (10)$$

The derivation is shown as follows: In the problem (P1), it is obvious that $C_1$ is equivalent to $C_{1,a}$ in (10) and $C_{1,c}$ :

$|[\mathbf{w}]_n| \geq 1/\sqrt{N}(\forall n \in \mathcal{N})$. Moreover, it can be derived from $C_{1,c}$ that $\|\mathbf{w}\|^2 = \sum_{n=1}^{N} |[\mathbf{w}]_n|^2 \geq 1$. Thus, the sufficient condition holds. Next, for $C_{1,a}$ and $C_{1,b}$ in (10), let us assume that there is at least one element whose modulus is less than $1/\sqrt{N}$ in the ABF vector $\mathbf{w}$, that is,

$$\begin{aligned} &|[\mathbf{w}]_i| < 1/\sqrt{N}, \forall i \in \mathcal{N}_1, \\ &|[\mathbf{w}]_j| = 1/\sqrt{N}, \forall j \in \mathcal{N}_2, \end{aligned} \tag{11}$$

where the sets $\mathcal{N}_1$ and $\mathcal{N}_2$ satisfy $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2, \mathcal{N}_1 \neq \emptyset, \mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$. Thus, we have

$$\begin{aligned} \sum_{n=1}^{N} |[\mathbf{w}]_i|^2 &= \sum_{i \in \mathcal{N}_1} |[\mathbf{w}]_i|^2 + \sum_{j \in \mathcal{N}_2} |[\mathbf{w}]_j|^2 \\ &< \tfrac{|\mathcal{N}_1|}{N} + \tfrac{|\mathcal{N}_2|}{N} = \tfrac{|\mathcal{N}|}{N} = 1. \end{aligned} \tag{12}$$

It can be found that (12) is conflict with $C_{1,b}$, which indicates that the modulus of each element in the ABF vector $\mathbf{w}$ is equal to $1/\sqrt{N}$, i.e., $C_1$ holds for $C_{1,a}$ and $C_{1,b}$. Thus, the necessary condition holds. In conclusion, the CM constraint $C_1$ is equal to (10). Using (10), the CE optimization problem $(\mathrm{P1})$ can be reformulated as

$$\begin{aligned} &(\bar{\mathrm{P}}1): \\ &\max_{\{\mathbf{w}, p_k, f_k\}} \quad \min_{k \in \mathcal{K}} \{\eta_k\} \\ &\quad \text{s.t.} \qquad C_{1,a}, C_{1,b}, C_2, C_3, C_4, C_5, C_6. \end{aligned} \tag{13}$$

**Theorem 1**: *By introducing auxiliary variables* $\{q_k, \gamma_k, z_k, R_k, P_k, \eta\}(\forall k \in \mathcal{K})$, *the problem* $(\bar{P}1)$ *can be transformed equivalently into the following problem:*

$$\begin{aligned} (\tilde{\mathrm{P}}1): \max_{\{\mathcal{V}, \eta\}} \quad & \eta \\ \text{s.t.} \quad & C_{1,a}, C_{1,b}, C_2, C_5, \\ & \tilde{C}_{3,a}: q_k^{-1} |\mathbf{h}_k^{\mathrm{H}} \mathbf{w}|^2 \geq \gamma_k z_k, \forall k \in \mathcal{K}, \\ & \tilde{C}_{3,b}: \sum_{i=k+1}^{K} q_k^{-1} |\mathbf{h}_i^{\mathrm{H}} \mathbf{w}|^2 + \sigma^2 \leq z_k, \forall k \in \mathcal{K}, \\ & \bar{C}_{3,c}: \log_2(1+\gamma_k) + \tfrac{f_k}{BC_k} \geq \tfrac{R_k}{B}, \forall k \in \mathcal{K}, \\ & \tilde{C}_{4,a}: \zeta_k q_k^{-1} + \xi_k f_k^3 + P_{k,c} \leq P_k, \forall k \in \mathcal{K}, \\ & \tilde{C}_{4,b}: P_k \leq P_k^{\max}, \forall k \in \mathcal{K}, \\ & \tilde{C}_6: 1/P_k^{\max} \leq q_k \leq 1/P_k^{\min}, \forall k \in \mathcal{K}, \\ & C_7: R_k \geq R_k^{\min}, \forall k \in \mathcal{K}, \\ & C_8: R_k \geq \eta P_k, \forall k \in \mathcal{K}, \end{aligned} \tag{14}$$

*where* $\mathcal{V} \triangleq \{\mathbf{w}, q_k, f_k, \gamma_k, z_k, R_k, P_k\}(\forall k \in \mathcal{K})$.

*Proof*: See Appendix A.

From Theorem 1, it is seen that $(\tilde{\mathrm{P}}1)$ is an equivalent form of the problem (P1). Thus, we can solve $(\tilde{\mathrm{P}}1)$ to address the original problem (P1).

### B. CE Optimization Algorithm Design

To design an efficient algorithm, we first analyze the structure of the problem $(\tilde{\mathrm{P}}1)$. It can be seen that the objective function of the problem $(\tilde{\mathrm{P}}1)$ is linear, and the constraints $\{C_{1,a}, \tilde{C}_{3,b}, \bar{C}_{3,c}, \tilde{C}_{4,a}, \tilde{C}_{4,b}, C_5, \tilde{C}_6, C_7\}$ are convex, but the constraints $\{C_{1,b}, C_2, \tilde{C}_{3,a}, C_8\}$ are non-convex. Therefore, we adopt SCA to approximately transform problem $(\tilde{\mathrm{P}}1)$ into a convex optimization. The core idea of SCA is to transform a non-convex optimization problem into a series of convex optimization problems, and at each iteration of SCA, the non-convex terms are replaced by appropriate convex terms [18].

Next, we introduce two Lemmas to find suitable convex sets to approximate the corresponding non-convex sets.

**Lemma 1:** *For a convex function* $f(\mathbf{x})$ *whose domain is* $\mathbf{x} \in \Omega$, *the following inequality holds:* [19]

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_0), \forall \mathbf{x}, \mathbf{x}_0 \in \Omega. \tag{15}$$

**Lemma 2:** *The following inequality holds:* [20]

$$xy \leq \tfrac{\tilde{y}}{2\tilde{x}} x^2 + \tfrac{\tilde{x}}{2\tilde{y}} y^2, \forall x, y, \tilde{x}, \tilde{y} > 0. \tag{16}$$

Let $\mathcal{V}^{(r-1)} \triangleq \{\mathbf{w}^{(r-1)}, q_k^{(r-1)}, f_k^{(r-1)}, \gamma_k^{(r-1)}, z_k^{(r-1)}, R_k^{(r-1)}, P_k^{(r-1)}\}(\forall k \in \mathcal{K})$ and $\eta^{(r-1)}$ denote the values of the optimization variables $\mathcal{V}$ and $\eta$ in (14) at the $(r-1)$-th SCA iteration, respectively.

By applying Lemma 1 and Lemma 2, the non-convex constraints $\{C_{1,b}, C_2, \tilde{C}_{3,a}, C_8\}$ can be transformed approximately into the following convex constraints:

$$\begin{aligned} & \tilde{C}_{1,b}: 2\mathrm{Re}\{(\mathbf{w}^{(r-1)})^{\mathrm{H}} \mathbf{w}\} - \|\mathbf{w}^{(r-1)}\|^2 \geq 1, \\ & \tilde{C}_2: 2\mathrm{Re}\{(\mathbf{w}^{(r-1)})^{\mathrm{H}} \mathbf{h}_{\tilde{k}} \mathbf{h}_{\tilde{k}}^{\mathrm{H}} \mathbf{w}\} - |\mathbf{h}_{\tilde{k}}^{\mathrm{H}} \mathbf{w}^{(r-1)}|^2 \\ & \qquad \geq |\mathbf{h}_{\tilde{k}+1}^{\mathrm{H}} \mathbf{w}|^2, \forall \tilde{k} \in \mathcal{K} \backslash K, \\ & \bar{C}_{3,a}: \tfrac{2\mathrm{Re}\{(\mathbf{w}^{(r-1)})^{\mathrm{H}} \mathbf{h}_k \mathbf{h}_k^{\mathrm{H}} \mathbf{w}\}}{q_k^{(r-1)}} - \tfrac{|\mathbf{h}_k^{\mathrm{H}} \mathbf{w}^{(r-1)}|^2}{(q_k^{(r-1)})^2} q_k \\ & \qquad \geq \tfrac{z_k^{(r-1)}}{2\gamma_k^{(r-1)}} \gamma_k^2 + \tfrac{\gamma_k^{(r-1)}}{2 z_k^{(r-1)}} z_k^2, \forall k \in \mathcal{K}, \\ & \tilde{C}_8: R_k \geq \tfrac{\eta^{(r-1)}}{2P_k^{(r-1)}} P_k^2 + \tfrac{P_k^{(r-1)}}{2\eta^{(r-1)}} \eta^2, \forall k \in \mathcal{K}. \end{aligned} \tag{17}$$

Therefore, at the $r$-th SCA iteration, the problem $(\tilde{\mathrm{P}}1)$ can be approximated as the following convex optimization problem:

$$\begin{aligned} (\hat{\mathrm{P}}1): \max_{\{\mathcal{V}, \eta\}} \quad & \eta \\ \text{s.t.} \quad & C_{1,a}, \tilde{C}_{1,b}, \tilde{C}_2, \bar{C}_{3,a}, \tilde{C}_{3,b}, \bar{C}_{3,c}, \\ & \tilde{C}_{4,a}, \tilde{C}_{4,b}, C_5, \tilde{C}_6, C_7, \tilde{C}_8. \end{aligned} \tag{18}$$

However, if the problem in (18) has a feasible solution $\hat{\mathbf{w}} \neq \mathbf{w}^{(r-1)}$, then we have $\|\hat{\mathbf{w}}\|^2 > 1$ and $\|\hat{\mathbf{w}}\|^2 \leq 1$ according to the constraints $C_{1,a}$ and $\tilde{C}_{1,b}$, which violates the assumption that $\hat{\mathbf{w}}$ is a feasible solution. Thus, due to the constraints $C_{1,a}$ and $\tilde{C}_{1,b}$, the feasible region of the problem $(\hat{\mathrm{P}}1)$ only contains $\mathbf{w} = \mathbf{w}^{(r-1)}$, which leads to that $\mathbf{w}^{(r)}$ in the SCA algorithm always equals to its initial value, i.e., $\mathbf{w}^{(r)} = \mathbf{w}^{(0)}$. In order to overcome the disadvantage of the above SCA algorithm, we use the penalty method in [21] [22] to modify the problem $(\hat{\mathrm{P}}1)$. Specifically, we add slack variable $u \geq 0$ to transform the constraint $\tilde{C}_{1,b}$ into:

$$\bar{C}_{1,b}: 2\mathrm{Re}\{(\mathbf{w}^{(r-1)})^{\mathrm{H}} \mathbf{w}\} - \|\mathbf{w}^{(r-1)}\|^2 + u \geq 1. \tag{19}$$

Similarly, we add slack variables $\{s_{1,\tilde{k}}, s_{2,k}, s_{3,k} \geq 0, \forall \tilde{k} \in \mathcal{K} \backslash K, k \in \mathcal{K}\}$ to transform the constraints $\{\tilde{C}_2, \bar{C}_{3,a}, \tilde{C}_8\}$

respectively into:

$$\begin{aligned}
&\bar{C}_2: 2\mathrm{Re}\{(\mathbf{w}^{(r-1)})^{\mathrm{H}}\mathbf{h}_{\tilde{k}}\mathbf{h}_{\tilde{k}}^{\mathrm{H}}\mathbf{w}\} - |\mathbf{h}_{\tilde{k}}^{\mathrm{H}}\mathbf{w}^{(r-1)}|^2 \\
&\quad +s_{1,\tilde{k}} \geq |\mathbf{h}_{\tilde{k}+1}^{\mathrm{H}}\mathbf{w}|^2, \forall \tilde{k} \in \mathcal{K}\backslash K, \\
&\hat{C}_{3,a}: \frac{2\mathrm{Re}\{(\mathbf{w}^{(r-1)})^{\mathrm{H}}\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}\mathbf{w}\}}{q_k^{(r-1)}} - \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}^{(r-1)}|^2}{(q_k^{(r-1)})^2}q_k \\
&\quad +s_{2,k} \geq \frac{z_k^{(r-1)}}{2\gamma_k^{(r-1)}}\gamma_k^2 + \frac{\gamma_k^{(r-1)}}{2z_k^{(r-1)}}z_k^2, \forall k \in \mathcal{K}, \\
&\bar{C}_8: R_k + s_{3,k} \geq \frac{\eta^{(r-1)}}{2P_k^{(r-1)}}P_k^2 + \frac{P_k^{(r-1)}}{2\eta^{(r-1)}}\eta^2, \forall k \in \mathcal{K}.
\end{aligned} \tag{20}$$

With (19) and (20), the convex optimization problem at the $r$-th iteration of the penalized SCA (PSCA) algorithm can be formulated as

$$\begin{aligned}
(\check{\mathrm{P}}1): \max_{\{\mathcal{V},\mathcal{S},\eta,u\}} \quad & \eta - \tau_1^{(r-1)}u - \tau_2^{(r-1)}\sum_{s\in\mathcal{S}} s \\
\text{s.t.} \quad & C_{1,a}, \bar{C}_{1,b}, \bar{C}_2, \hat{C}_{3,a}, \tilde{C}_{3,b}, \bar{C}_{3,c}, \\
& \tilde{C}_{4,a}, \tilde{C}_{4,b}, C_5, \tilde{C}_6, C_7, \bar{C}_8, \\
& C_9: s \geq 0, \forall s \in \mathcal{S}, \\
& C_{10}: u \geq 0,
\end{aligned} \tag{21}$$

where $\mathcal{S} \triangleq \{s_{1,\tilde{k}}, s_{2,k}, s_{3,k}\}(\forall \tilde{k} \in \mathcal{K}\backslash K, k \in \mathcal{K})$, $\tau_i^{(r-1)}(i=1,2)$ denotes the penalty variable at the $(r-1)$-th iteration of the PSCA algorithm. $\tau_i^{(r-1)}(i=1,2)$ is updated by

$$\tau_i^{(r)} = \min\{\mu_i\tau_i^{(r-1)}, \tau_i^{\max}\}, i = 1,2, \tag{22}$$

where $\mu_i > 1$ and $\tau_i^{\max}$ denote the increasing coefficient and upper bound of $\tau_i^{(r-1)}$, respectively.

Therefore, the PSCA-based CE optimization algorithm for solving the problem $(\tilde{\mathrm{P}}1)$ can be summarized as Algorithm 1. For the convenience of expression, "MaxMinCE" and "MaxMinCB" refer to maximizing the minimum CE of the users and maximizing the minimum CB of the users, respectively. Although Algorithm 1 is designed for the MaxMinCE scheme, it can be also applied to the MaxMinCB scheme. For the MaxMinCB scheme, the optimization objective is $\max\min\{R_k\}$, which needs to introduce auxiliary variable $\tilde{R}$ (such that $\min\{R_k\} \geq \tilde{R}$) so as to apply the framework of Algorithm 1.

### C. Convergence and Complexity Analysis

The convergence and complexity of Algorithm 1 are analyzed here. On one hand, from the constraints of problem $(\tilde{\mathrm{P}}1)$, it can be concluded that: $0 < R_k < B\log_2(1 + p_k^{\max}\|\mathbf{h}_k\|^2) + f_k^{\max}/C_k, P_{k,c} < P_k \leq P_k^{\max}, \forall k \in \mathcal{K}$. On the other hand, we have: $\eta \leq \min\{R_k/P_k\}$. Thus, the objective function of problem $(\tilde{\mathrm{P}}1)$ has an upper bound. Based on the above analysis, if the feasible region of problem $(\tilde{\mathrm{P}}1)$ is not empty, then Algorithm 1 can converge to a local optimal solution of problem $(\tilde{\mathrm{P}}1)$ [21] [22]. Next, the complexity of Algorithm 1 mainly comes from solving the convex optimization problem $(\check{\mathrm{P}}1)$. The standard convex optimization tools based on the interior point method (such as CVX [23]) can be used to obtain the optimal solution of the problem $(\check{\mathrm{P}}1)$. Since the number of real optimization variables in the problem $(\check{\mathrm{P}}1)$ is $2N + 9K + 1$, the complexity of the interior point method can be expressed as $\mathcal{O}((2N + 9K + 1)^{3.5}\ln(1/\delta))$, where $\delta$ is the solution accuracy [24]. Hence, the complexity of Algorithm 1 is given as $\mathcal{O}(I_1(2N + 9K + 1)^{3.5}\ln(1/\delta))$, where $I_1$ denotes the iterations of Algorithm 1.

**Algorithm 1** PSCA-based CE Optimization Algorithm for Solving the Problem $(\tilde{\mathrm{P}}1)$

```
Initialize: An initial point $\{\mathcal{V}^{(0)}, \mathcal{S}^{(0)}, \eta^{(0)}, u^{(0)}\}$ in the problem $(\check{\mathrm{P}}1)$, iteration index $r = 0$, maximum iteration number $r_{\max}$, iteration tolerance $\epsilon_i > 0 (i = 1,2)$, penalty parameters $\{\tau_i^{(0)}, \mu_i, \tau_i^{\max}\}(\forall i = 1,2)$, the sum of slack variables $\Gamma_1^{(0)} = 0$ and the objective value $\Gamma_2^{(0)} = 0$ in the problem $(\check{\mathrm{P}}1)$;
repeat
    $r = r + 1$;
    Compute the optimal solution $\{\mathcal{V}^{\mathrm{opt}}, \mathcal{S}^{\mathrm{opt}}, \eta^{\mathrm{opt}}, u^{\mathrm{opt}}\}$ of the problem $(\check{\mathrm{P}}1)$;
    Update $\mathcal{V}^{(r)} = \mathcal{V}^{\mathrm{opt}}, \mathcal{S}^{(r)} = \mathcal{S}^{\mathrm{opt}}, \eta^{(r)} = \eta^{\mathrm{opt}}, u^{(r)} = u^{\mathrm{opt}}$;
    Update $\Gamma_1^{(r)} = u^{(r)} + \sum_{s^{(r)}\in\mathcal{S}^{(r)}} s^{(r)}, \Gamma_2^{(r)} = \eta^{(r)} - \tau_1^{(r-1)}u^{(r)} - \tau_2^{(r-1)}\sum_{s^{(r)}\in\mathcal{S}^{(r)}} s^{(r)}$;
    if $|\Gamma_1^{(r)}| < \epsilon_1$ and $|\Gamma_2^{(r)} - \Gamma_2^{(r-1)}| < \epsilon_2$ then
        Set flag = 1;
    else:
        Update $\tau_i^{(r)}(i = 1,2)$ using (22);
    end if
until flag = 1 or $r > r_{\max}$
Output: A local optimal solution $\{\mathcal{V}^{(r)}, \eta^{(r)}\}$.
```

## IV. System model of mmWave-MEC with HBF

In the section above, the CE optimization of mmWave-MEC-ABF is presented, but the performance may be limited. This is because only the ABF is used and a single RFC is considered, where only one data stream can be processed at each time, which may result in the limited CE. Next, we give the CE optimization of mmWave-MEC-HBF. By jointly designing the digital BF and analog BF, the CE performance will be increased obviously.

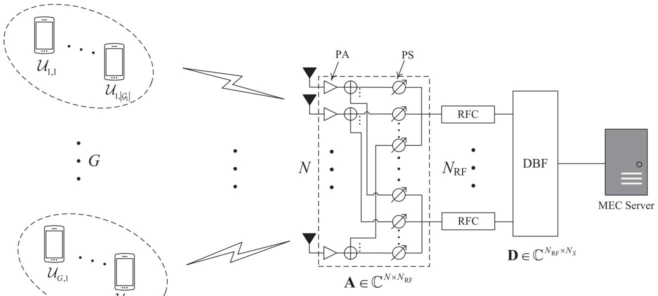

$\mathcal{U}_{1,1}$
$\mathcal{U}_{1,|\mathcal{G}_1|}$
$G$
$\mathcal{U}_{G,1}$
$\mathcal{U}_{G,|\mathcal{G}_G|}$
PA
PS
$N$
$\mathbf{A} \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$
RFC
RFC
$N_{\mathrm{RF}}$
DBF
$\mathbf{D} \in \mathbb{C}^{N_{\mathrm{RF}}\times N_S}$
MEC Server

Fig. 2: System model of mmWave-MEC-HBF.

The system model of mmWave-MEC-HBF is shown in Fig. 2, where the BS is equipped with $N$ antennas, $N_{\mathrm{RF}}$ RFCs, $N$ PAs and $N_{\mathrm{RF}}N$ PSs, in which each antenna is connected to all RFCs via one PA and $N_{\mathrm{RF}}$ PSs. The HBF of the BS is composed of the ABF matrix $\mathbf{A} \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$ and the DBF matrix $\mathbf{D} \in \mathbb{C}^{N_{\mathrm{RF}}\times N_S}$, where $N_S$ denotes the number of data streams. Accordingly, $\mathbf{A}$ needs to satisfy the CM constraint [25], i.e., $|[\mathbf{A}]_{i,j}| = 1/\sqrt{N}, \forall i \in \mathcal{N} \triangleq \{1, \ldots, N\}, j \in$

$\mathcal{N}_{\mathrm{RF}} \triangleq \{1, \ldots, N_{\mathrm{RF}}\}$. We suppose that $N_S = N_{\mathrm{RF}}$, while all $N_{\mathrm{U}}$ users are divided into $G$ groups and each group corresponds to an independent data stream, i.e., $G = N_S$ [25]. Hence, $\mathbf{D}$ can be rewritten as $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_G]$, where $\mathbf{d}_g \in \mathbb{C}^{N_{\mathrm{RF}} \times 1}, \forall g \in \mathcal{G} \triangleq \{1, \ldots, G\}$. Specifically, the users in the $g$-th group ($\forall g \in \mathcal{G}$) is represented by $\mathcal{G}_g$, and the number of the users in the $g$-th group is $|\mathcal{G}_g|$, where the $k$-th user in the $g$-th group is denoted by $\mathcal{U}_{g,k}(\forall g \in \mathcal{G}, k \in \mathcal{G}_g)$.

According to the uplink NOMA protocol, the received signal of the $g$-th group at the BS after the user grouping and the processing of HBF can be expressed as

$$y_g = \sum_{m=1}^{G} \sum_{n=1}^{|\mathcal{G}_m|} \mathbf{d}_g^{\mathrm{H}} \mathbf{A}^{\mathrm{H}} \mathbf{h}_{m,n} \sqrt{p_{m,n}} x_{m,n} + \mathbf{d}_g^{\mathrm{H}} \mathbf{A}^{\mathrm{H}} \mathbf{n}_g, \tag{23}$$

where $x_{i,j} \sim \mathcal{CN}(0, 1)$ and $p_{i,j}$ denote the transmission signal and power of $\mathcal{U}_{i,j}(\forall i \in \mathcal{G}, j \in \mathcal{G}_g)$, respectively. Besides, the mmWave channel between $\mathcal{U}_{g,k}$ and the BS $\mathbf{h}_{g,k} \in \mathbb{C}^{N \times 1}$, the number of CBs $L_{g,k}^{\mathrm{loc}}$, and the corresponding energy consumption $E_{g,k}^{\mathrm{loc}}$ are similar to (1) and (3), respectively, in which the subscripts $k$ of variables in (1) and (3) need to be revised as $\{g, k\}$ for consistency.

For the uplink NOMA, the decoding order of SIC is the descending order of the effective channel gains of the users in the group [26] generally. Similar to (5), the following SIC decoding constraint is necessary when designing $\mathbf{A}$:

$$|\mathbf{h}_{g,1}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2 \geq \cdots \geq |\mathbf{h}_{g,|\mathcal{G}_g|}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2, \forall g \in \mathcal{G}, \tag{24}$$

where the effective channel gain of $\mathcal{U}_{g,k}$ is defined as $|\mathbf{d}_g^{\mathrm{H}} \mathbf{A}^{\mathrm{H}} \mathbf{h}_{g,k}|^2 = |\mathbf{h}_{g,k}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2$. With the decoding order above, and using the SIC, the SINR of $\mathcal{U}_{g,k}$ can be expressed as

$$\mathrm{SINR}_{g,k} = \frac{p_{g,k} |\mathbf{h}_{g,k}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2}{\sum\limits_{\tilde{k}=k+1}^{|\mathcal{G}_g|} p_{g,\tilde{k}} |\mathbf{h}_{g,\tilde{k}}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2 + \sum\limits_{m \neq g}^{G} \sum\limits_{n=1}^{|\mathcal{G}_m|} p_{m,n} |\mathbf{h}_{m,n}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2 + I_{g,k}}, \tag{25}$$

where $I_{g,k} = \|\mathbf{A} \mathbf{d}_g\|^2 \sigma^2$. For the upload computing of $\mathcal{U}_{g,k}$, its achievable rate is given by $\bar{R}_{g,k} = B\log_2(1 + \mathrm{SINR}_{g,k})$. Then, the number of CBs $L_{g,k}^{\mathrm{off}}$ and the energy consumption $E_{g,k}^{\mathrm{off}}$ are given by

$$L_{g,k}^{\mathrm{off}} = T \bar{R}_{g,k}, \tag{26a}$$

$$E_{g,k}^{\mathrm{off}} = T \zeta_{g,k} p_{g,k}, \tag{26b}$$

where $\zeta_{g,k}$ denotes the PA coefficient of $\mathcal{U}_{g,k}$ [26].

Therefore, the CE of $\mathcal{U}_{g,k}$ in the mmWave-MEC-HBF can be defined as

$$\eta_{g,k} = \frac{B\log_2(1 + \mathrm{SINR}_{g,k}) + f_{g,k}/C_{g,k}}{\zeta_{g,k} p_{g,k} + \xi_{g,k} f_{g,k}^3 + P_{g,k,c}}. \tag{27}$$

## V. CE Optimization for mmWave-MEC with HBF

In this section, we will further study the CE optimization based on the max-min fairness criterion for mmWave-MEC-HBF and propose an efficient iterative algorithm to tackle this more complicate optimization problem.

### A. Problem Formulation and Transformation

Aiming at the joint design of the DBF matrix $\mathbf{D}$ and the ABF matrix $\mathbf{A}$ at the BS and the local resource allocation of each user, the max-min CE optimization problem for the mmWave-MEC-HBF can be formulated as

$$\begin{aligned} (\mathrm{P2}): \max_{\mathcal{V}_0} \quad & \min_{g \in \mathcal{G}, k \in \mathcal{G}_g} \{\eta_{g,k}\} \\ \text{s.t.} \quad & C_1: |[\mathbf{A}]_{i,j}| = 1/\sqrt{N}, \forall i \in \mathcal{N}, j \in \mathcal{N}_{\mathrm{RF}}, \\ & C_2: |\mathbf{h}_{g,1}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2 \geq \cdots \geq |\mathbf{h}_{g,|\mathcal{G}_g|}^{\mathrm{H}} \mathbf{A} \mathbf{d}_g|^2, \forall g \in \mathcal{G}, \\ & C_3: B\log_2(1 + \mathrm{SINR}_{g,k}) + f_{g,k}/C_{g,k} \\ & \quad \geq R_{g,k}^{\min}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g, \\ & C_4: \zeta_{g,k} p_{g,k} + \xi_{g,k} f_{g,k}^3 + P_{g,k,c} \\ & \quad \leq P_{g,k}^{\max}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g, \\ & C_5: 0 \leq f_{g,k} \leq f_{g,k}^{\max}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g, \\ & C_6: P_{g,k}^{\min} \leq p_{g,k} \leq P_{g,k}^{\max}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g, \end{aligned} \tag{28}$$

where $\mathcal{V}_0 \triangleq \{\mathbf{D}, \mathbf{A}, p_{g,k}, f_{g,k}\}(\forall g \in \mathcal{G}, k \in \mathcal{G}_g)$, $C_1$ denotes the CM constraint, $C_2$ denotes the SIC decoding constraint, $C_3$ denotes the computation-bit rate constraint, $R_{g,k}^{\min}$ is the minimum computation-bit rate of $\mathcal{U}_{g,k}$, $C_4$ denotes the power consumption constraint, $P_{g,k}^{\max}$ is the maximum power consumption of $\mathcal{U}_{g,k}$, $C_5$ denotes the CPU frequency constraint, $f_{g,k}^{\max}$ is the maximum CPU frequency of $\mathcal{U}_{g,k}$, $C_6$ denotes the transmission power constraint, $P_{g,k}^{\min}$ is the minimum transmission power of $\mathcal{U}_{g,k}$, which is close to 0.

Compared with the problem (P1), the number of real optimization variables in problem (P2) is increased to $2(N_{\mathrm{RF}}G + NN_{\mathrm{RF}} + N_{\mathrm{U}})$. Moreover, the optimization variables $\mathbf{D}$ and $\mathbf{A}$ in problem (P2) are coupled with each other. Hence, we aim to design an computation-efficient algorithm with polynomial computation complexity to find a suboptimal solution of problem (P2). To this end, we give an equivalent form of problem (P2) shown in Theorem 2.

**Theorem 2**: *By introducing auxiliary variables $\{\mathbf{w}_g, q_{g,k}, \gamma_{g,k}, z_{g,k}, R_{g,k}, P_{g,k}, \eta\}(\forall g \in \mathcal{G}, k \in \mathcal{G}_g)$, problem (P2) can be transformed equivalently into the following problem:*

$$\begin{aligned} & (\bar{\mathrm{P}}2): \\ & \max_{\mathcal{V}_1} \quad \eta \\ & \text{s.t.} \quad C_1, C_5 \\ & \quad \tilde{C}_2: |\mathbf{h}_{g,1}^{\mathrm{H}} \mathbf{w}_g|^2 \geq \cdots \geq |\mathbf{h}_{g,|\mathcal{G}_g|}^{\mathrm{H}} \mathbf{w}_g|^2, \forall g \in \mathcal{G} \\ & \quad \tilde{C}_{3,a}: q_{g,k}^{-1} |\mathbf{h}_{g,k}^{\mathrm{H}} \mathbf{w}_g|^2 \geq \gamma_{g,k} z_{g,k}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad \tilde{C}_{3,b}: \sum_{\tilde{k}=k+1}^{|\mathcal{G}_g|} q_{g,\tilde{k}}^{-1} |\mathbf{h}_{g,\tilde{k}}^{\mathrm{H}} \mathbf{w}_g|^2 + \sum_{m \neq g}^{G} \sum_{n=1}^{|\mathcal{G}_m|} q_{m,n}^{-1} |\mathbf{h}_{m,n}^{\mathrm{H}} \mathbf{w}_g|^2 \\ & \qquad + \|\mathbf{w}_g\|^2 \sigma^2 \leq z_{g,k}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad \bar{C}_{3,c}: \log_2(1 + \gamma_{g,k}) + \tfrac{f_{g,k}}{BC_{g,k}} \geq \tfrac{R_{g,k}}{B}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad \tilde{C}_{4,a}: \tfrac{\zeta_{g,k}}{q_{g,k}} + \xi_{g,k} f_{g,k}^3 + P_{g,k,c} \leq P_{g,k}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad \tilde{C}_{4,b}: P_{g,k} \leq P_{g,k}^{\max}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad \tilde{C}_6: 1/P_{g,k}^{\max} \leq q_{g,k} \leq 1/P_{g,k}^{\min}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad C_7: \mathbf{w}_g = \mathbf{A} \mathbf{d}_g, \forall g \in \mathcal{G} \\ & \quad C_8: R_{g,k} \geq R_{g,k}^{\min}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \\ & \quad C_9: R_{g,k} \geq \eta P_{g,k}, \forall g \in \mathcal{G}, k \in \mathcal{G}_g, \end{aligned} \tag{29}$$

*where $\mathcal{V} \triangleq \{\mathbf{w}_g, q_{g,k}, f_{g,k}, \gamma_{g,k}, z_{g,k}, R_{g,k}, P_{g,k}\}$* ($\forall g \in \mathcal{G}$, $k \in \mathcal{G}_g$) and $\mathcal{V}_1 \triangleq \{\mathbf{D}, \mathbf{A}, \mathcal{V}, \eta\}$.

*Proof*: See Appendix B.

In view of Theorem 2, we can solve the problem ($\bar{\text{P}}2$) to tackle the original problem (P2).

### *B. CE Optimization Algorithm Design*

Note that the optimization variables $\mathbf{D}$ and $\mathbf{A}$ only appear in $C_7$. In order to deal with the equality constraint, we use the penalty function method in [27] to add a quadratic penalty term into the objective function of the problem ($\bar{\text{P}}2$), which yields the following optimization:

$$\begin{aligned}(\tilde{\text{P}}2):\max_{\mathcal{V}_1}\quad & \eta-\tfrac{1}{2}\varrho^{(l-1)}\sum_{g=1}^{G}\|\mathbf{w}_g-\mathbf{A}\mathbf{d}_g\|^2\\ \text{s.t.}\quad & C_1,\tilde{C}_2,\tilde{C}_{3,a},\tilde{C}_{3,b},\bar{C}_{3,c},\tilde{C}_{4,a},\tilde{C}_{4,b},\\ & C_5,\tilde{C}_6,C_8,C_9,\end{aligned}\tag{30}$$

where $\varrho^{(l-1)}$ is the penalty variable of the $(l-1)$ iteration, which is updated by $\varrho^{(l)}=\varpi\varrho^{(l-1)}(\varpi>1)$.

By analyzing the structure of the problem ($\tilde{\text{P}}2$), we can find that the optimization variables in the problem ($\tilde{\text{P}}2$) have block structure. Thus, the problem ($\tilde{\text{P}}2$) can be solved efficiently by using the IBCD algorithm [28].

Let $\{\mathbf{D}^{(r-1)},\mathbf{A}^{(r-1)},\mathcal{V}^{(r-1)},\eta^{(r-1)}\}$ denote the values of the optimization variables $\{\mathbf{D},\mathbf{A},\mathcal{V},\eta\}$ in the problem ($\tilde{\text{P}}2$) at the $(r-1)$-th iteration of the IBCD algorithm, where $\mathcal{V}^{(r-1)}\triangleq\{\mathbf{w}_g^{(r-1)},q_{g,k}^{(r-1)},f_{g,k}^{(r-1)},\gamma_{g,k}^{(r-1)},z_{g,k}^{(r-1)},R_{g,k}^{(r-1)},P_{g,k}^{(r-1)}\}(\forall g\in\mathcal{G},k\in\mathcal{G}_g)$. Then, the following steps need to be performed at the $r$-th iteration of the IBCD algorithm:

1) Solving $\mathbf{D}$ for fixed $\{\mathbf{A},\mathcal{V},\eta\}$

When $\{\mathbf{A},\mathcal{V},\eta\}$ is fixed, the subproblem for solving $\mathbf{D}$ is

$$(\text{P}2.1):\min_{\mathbf{D}}\quad\sum_{g=1}^{G}\|\mathbf{w}_g-\mathbf{A}\mathbf{d}_g\|^2.\tag{31}$$

The above problem can be decomposed into $G$ subproblem, in which the $g$-th ($\forall g\in\mathcal{G}$) subproblem is

$$(\text{P}2.1.1):\min_{\mathbf{d}_g}\quad\|\mathbf{w}_g-\mathbf{A}\mathbf{d}_g\|^2.\tag{32}$$

Since the closed-form optimal solution of the problem (P2.1.1) can be derived as $\mathbf{d}_g^{\text{opt}}=(\mathbf{A}^{\text{H}}\mathbf{A})^{\dagger}\mathbf{A}^{\text{H}}\mathbf{w}_g$, the $\mathbf{D}^{(r)}$ is updated by

$$\mathbf{D}^{(r)}=[\mathbf{d}_1^{\text{opt}},...,\mathbf{d}_G^{\text{opt}}].\tag{33}$$

2) Solving $\mathbf{A}$ for fixed $\{\mathbf{D},\mathcal{V},\eta\}$

When $\{\mathbf{D},\mathcal{V},\eta\}$ is fixed, the subproblem for solving $\mathbf{A}$ is

$$\begin{aligned}(\text{P}2.2):\min_{\mathbf{A}}\quad & \sum_{g=1}^{G}\|\mathbf{w}_g-\mathbf{A}\mathbf{d}_g\|^2\\ \text{s.t.}\quad & |[\mathbf{A}]_{i,j}|=1/\sqrt{N},\forall i\in\mathcal{N},j\in\mathcal{N}_{\text{RF}}.\end{aligned}\tag{34}$$

Next, we apply the MM algorithm in [30] to tackle this problem. Firstly, letting $\mathbf{A}=[\mathbf{a}_1,...,\mathbf{a}_N]^{\text{H}}$, $\mathbf{a}_i\in\mathbb{C}^{N_{\text{RF}}\times 1}(\forall i\in\mathcal{N})$, the problem (P2.2) can be rewritten as

$$\begin{aligned}&(\bar{\text{P}}2.2):\\ &\min_{\{\mathbf{a}_i\}}\quad \sum_{g=1}^{G}\sum_{i=1}^{N}|[\mathbf{w}_g]_i-\mathbf{a}_i^{\text{H}}\mathbf{d}_g|^2=\sum_{i=1}^{N}\sum_{g=1}^{G}|\mathbf{d}_g^{\text{H}}\mathbf{a}_i-[\mathbf{w}_g]_i^*|^2\\ &\text{s.t.}\quad |[\mathbf{a}_i]_j|=1/\sqrt{N},\forall i\in\mathcal{N},j\in\mathcal{N}_{\text{RF}}.\end{aligned}\tag{35}$$

Moreover, the problem ($\bar{\text{P}}2.2$) can be decomposed into $N$ subproblem, in which the $i$-th ($i\in\mathcal{N}$) subproblem is

$$\begin{aligned}(\text{P}2.2.1):\min_{\mathbf{a}_i}\quad & \mathbf{a}_i^{\text{H}}\tilde{\mathbf{D}}\mathbf{a}_i-2\text{Re}\{\mathbf{a}_i^{\text{H}}\tilde{\mathbf{d}}_i\}+\sum_{g=1}^{G}|[\mathbf{w}_g]_i^*|^2\\ \text{s.t.}\quad & |[\mathbf{a}_i]_j|=1/\sqrt{N},\forall i\in\mathcal{N},j\in\mathcal{N}_{\text{RF}},\end{aligned}\tag{36}$$

where $\tilde{\mathbf{D}}=\sum_{g=1}^{G}\mathbf{d}_g\mathbf{d}_g^{\text{H}}$ and $\tilde{\mathbf{d}}_i=\sum_{g=1}^{G}\mathbf{d}_g[\mathbf{w}_g]_i^*$.

Let $\tilde{\mathbf{A}}^{(t-1)}=[\tilde{\mathbf{a}}_1^{(t-1)},...,\tilde{\mathbf{a}}_N^{(t-1)}]^{\text{H}}$ denote the value of $\mathbf{A}=[\mathbf{a}_1,...,\mathbf{a}_N]^{\text{H}}$ at the $(t-1)$ iteration of the MM algorithm. A tight upper bound of $\mathbf{a}_i^{\text{H}}\tilde{\mathbf{D}}\mathbf{a}_i$ at the $t$-th iteration of the MM algorithm can be expressed as [29]

$$\mathbf{a}_i^{\text{H}}\tilde{\mathbf{D}}\mathbf{a}_i\le\mathbf{a}_i^{\text{H}}\hat{\mathbf{D}}\mathbf{a}_i-2\text{R}e\{\mathbf{a}_i^{\text{H}}\mathbf{b}_i^{(t-1)}\}+(\tilde{\mathbf{a}}_i^{(t-1)})^{\text{H}}\mathbf{b}_i^{(t-1)},\tag{37}$$

where $\mathbf{b}_i^{(t-1)}=(\hat{\mathbf{D}}-\tilde{\mathbf{D}})\mathbf{a}_i^{(t-1)}$, $\hat{\mathbf{D}}=\lambda_{\max}(\tilde{\mathbf{D}})\mathbf{I}_{N_{\text{RF}}}$, and $\lambda_{\max}(\tilde{\mathbf{D}})$ represents the maximum eigenvalue of $\tilde{\mathbf{D}}$.

Note that $\mathbf{a}_i^{\text{H}}\hat{\mathbf{D}}\mathbf{a}_i$ is equal to a constant $\lambda_{\max}(\tilde{\mathbf{D}})$ for any feasible solution $\mathbf{a}_i$ in the problem (P2.2.1). Therefore, at the $t$-th iteration of the MM algorithm, using the problem (P2.2.1) and discarding constant terms, the problem (P2.2.1) can be approximated as

$$\begin{aligned}(\bar{\text{P}}2.2.1):\max_{\mathbf{a}_i}\quad & \text{Re}\{\mathbf{a}_i^{\text{H}}\tilde{\mathbf{b}}_i^{(t-1)}\}\\ \text{s.t.}\quad & |[\mathbf{a}_i]_j|=1/\sqrt{N},\forall i\in\mathcal{N},j\in\mathcal{N}_{\text{RF}},\end{aligned}\tag{38}$$

where $\tilde{\mathbf{b}}_i^{(t-1)}=\mathbf{b}_i^{(t-1)}+\tilde{\mathbf{d}}_i$.

Then, we decompose the problem ($\bar{\text{P}}2.2.1$) into $N_{\text{RF}}$ subproblems, in which the $j$-th ($\forall j\in\mathcal{N}_{\text{RF}}$) subproblem is

$$\begin{aligned}(\tilde{\text{P}}2.2.1):\max_{\psi_{i,j}}\quad & \cos(\varphi_{i,j}^{(t-1)}-\psi_{i,j})\\ \text{s.t.}\quad & 0\le\psi_{i,j}\le 2\pi,\end{aligned}\tag{39}$$

where $\varphi_{i,j}^{(t-1)}$ and $\psi_{i,j}$ denote the angle of $[\tilde{\mathbf{b}}_i^{(t-1)}]_j$ and $[\mathbf{a}_i]_j$, respectively. Obviously, the optimal solution of the problem ($\tilde{\text{P}}2.2.1$) is $\psi_{i,j}^{\text{opt}}=\varphi_{i,j}^{(t-1)}$, so the optimal solution of the problem ($\bar{\text{P}}2.2.1$) can be expressed as

$$\tilde{\mathbf{a}}_i^{\text{opt}}=\tfrac{1}{\sqrt{N}}[e^{j\psi_{i,1}^{\text{opt}}},\ldots,e^{\quad{}^{\text{opt}}_{i,N_{\text{RF}}}}]^{\text{T}}.\tag{40}$$

Based on the above analysis, the MM algorithm for solving (34) can be summarized as Algorithm 2.

**Algorithm 2** MM Algorithm for Solving the Problem (P2.2)

```
1: Initialize: iteration index $t = 0$, maximum iteration number $t_{\max}$, iteration tolerance $\varepsilon > 0$, initial point $\tilde{\mathbf{A}}^{(t)} = \mathbf{A}^{(r-1)}$;
2: repeat
3:     $t = t + 1$;
4:     for $i = 1 : N$
           Update $\tilde{\mathbf{a}}_i^{(t)}$ according to (40);
       end for
5:     Update $\tilde{\mathbf{A}}^{(t)} = [\tilde{\mathbf{a}}_1^{(t)}, ..., \tilde{\mathbf{a}}_N^{(t)}]^{\text{H}}$;
6: until $\|\tilde{\mathbf{A}}^{(t)} - \tilde{\mathbf{A}}^{(t-1)}\|_{\text{F}} < \varepsilon$ or $t > t_{\max}$
7: Output: $\mathbf{A}^{(r)} = \tilde{\mathbf{A}}^{(t)}$.
```

3) Solving $\{\mathcal{V},\eta\}$ for fixed $\{\mathbf{D},\mathbf{A}\}$

When $\{\mathbf{D}, \mathbf{A}\}$ is fixed, the subproblem for solving $\{\mathcal{V}, \eta\}$ is

$$\begin{aligned} &(\text{P2.3}): \\ &\max_{\{\mathcal{V},\eta\}} \quad \eta - \tfrac{1}{2}\varrho^{(l-1)} \sum_{g=1}^{G} \|\mathbf{w}_g - \mathbf{A}\mathbf{d}_g\|^2 \\ &\text{s.t.} \quad \tilde{C}_2, \tilde{C}_{3,a}, \tilde{C}_{3,b}, \bar{C}_{3,c}, \tilde{C}_{4,a}, \tilde{C}_{4,b}, \\ &\qquad C_5, \tilde{C}_6, C_8, C_9. \end{aligned} \tag{41}$$

The optimization problem in the problem (P2.3) can be solved approximately by the SCA algorithm. Firstly, we analyze the structure of the problem (P2.3) and know that the objective function of the problem (P2.3) is concave, and the constraints $\{\tilde{C}_{3,b}, \bar{C}_{3,c}, \tilde{C}_{4,a}, \tilde{C}_{4,b}, C_5, \tilde{C}_6, C_7, C_8\}$ are convex, but the constraints $\{\tilde{C}_2, \tilde{C}_{3,a}, C_9\}$ are non-convex.

Then, using Lemma 1 and Lemma 2, the non-convex constraints $\{\tilde{C}_2, \tilde{C}_{3,a}, C_9\}$ can be approximated as

$$\begin{aligned} &\bar{C}_2 : 2\text{Re}\{(\mathbf{w}_g^{(r-1)})^{\text{H}} \mathbf{h}_{g,k}\mathbf{h}_{g,k}^{\text{H}}\mathbf{w}_g\} - |\mathbf{h}_{g,k}^{\text{H}}\mathbf{w}_g^{(r-1)}|^2 \\ &\quad \geq |\mathbf{h}_{g,k+1}^{\text{H}}\mathbf{w}_g|^2, \forall g \in \mathcal{G}, k \in \mathcal{G}_g \backslash |\mathcal{G}_g|, \\ &\bar{C}_{3,a} : \frac{2\text{Re}\{(\mathbf{w}_g^{(r-1)})^{\text{H}}\mathbf{h}_{g,k}\mathbf{h}_{g,k}^{\text{H}}\mathbf{w}_g\}}{q_{g,k}^{(r-1)}} - \frac{|\mathbf{h}_{g,k}^{\text{H}}\mathbf{w}_g^{(r-1)}|^2}{(q_{g,k}^{(r-1)})^2} q_{g,k} \\ &\quad \geq \frac{z_{g,k}^{(r-1)}}{2\gamma_{g,k}^{(r-1)}}\gamma_{g,k}^2 + \frac{\gamma_{g,k}^{(r-1)}}{2z_{g,k}^{(r-1)}} z_{g,k}^2, \forall g \in \mathcal{G}, k \in \mathcal{G}_g, \\ &\tilde{C}_9 : R_{g,k} \geq \frac{\eta^{(r-1)}}{2P_{g,k}^{(r-1)}} P_{g,k}^2 + \frac{P_{g,k}^{(r-1)}}{2\eta^{(r-1)}}\eta^2, \forall g \in \mathcal{G}, k \in \mathcal{G}_g. \end{aligned} \tag{42}$$

Therefore, the problem (P2.3) can be approximated as the following convex optimization problem:

$$\begin{aligned} &(\bar{\text{P}}2.3): \\ &\max_{\{\mathcal{V},\eta\}} \quad \eta - \tfrac{1}{2}\varrho^{(l-1)} \sum_{g=1}^{G} \|\mathbf{w}_g - \mathbf{A}\mathbf{d}_g\|^2 \\ &\text{s.t.} \quad \bar{C}_2, \bar{C}_{3,a}, \tilde{C}_{3,b}, \bar{C}_{3,c}, \tilde{C}_{4,a}, \tilde{C}_{4,b}, \\ &\qquad C_5, \tilde{C}_6, C_8, \tilde{C}_9. \end{aligned} \tag{43}$$

Accordingly, the $\{\mathcal{V}^{(r)}, \eta^{(r)}\}$ is updated by

$$\mathcal{V}^{(r)} = \mathcal{V}^{\text{opt}}, \eta^{(r)} = \eta^{\text{opt}}, \tag{44}$$

where $\{\mathcal{V}^{\text{opt}}, \eta^{\text{opt}}\}$ represent the optimal solution of the problem ($\bar{\text{P}}$2.3).

With the analysis above, the CE optimization algorithm based on the penalty function method and the IBCD method for solving the problem ($\bar{\text{P}}$2), namely the penalized IBCD (PIBCD) algorithm is proposed, and it is summarized as Algorithm 3. Moreover, Algorithm 3 can be also applied to the MaxMinCB scheme. For the MaxMinCB scheme, the optimization objective is $\max\min\{R_{g,k}\}$, which needs to introduce auxiliary variable $\tilde{R}$ (such that $\min\{R_{g,k}\} \geq \tilde{R}$) so as to apply the framework of Algorithm 3.

**Algorithm 3** PIBCD-based CE Optimization Algorithm for Solving the Problem ($\bar{\text{P}}$2)

```
**Initialize**: outer iteration index $l = 0$, outer maximum iteration number $l_{\max}$, outer iteration tolerance $\epsilon_1 > 0$, inner iteration index $r = 0$, inner maximum iteration number $r_{\max}$, inner iteration tolerance $\epsilon_2 > 0$, penalty parameters $\{\rho^{(0)} > 0, \varpi > 1\}$, constraint violation $\Xi^{(0)}$;
**repeat**
    $l = l + 1$;
    Set $r = 0$, the initial point $\{\mathbf{D}^{(0)}, \mathbf{A}^{(0)}, \mathcal{V}^{(0)}, \eta^{(0)}\}$ and the objective value $\Gamma^{(0)} = 0$ in the problem ($\tilde{\text{P}}$2);
    **repeat**
        $r = r + 1$;
        For given $\{\mathbf{A}^{(r-1)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}\}$, update $\mathbf{D}^{(r)}$ using (33);
        For given $\{\mathbf{D}^{(r)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}\}$, update $\mathbf{A}^{(r)}$ using Algorithm 2;
        For given $\{\mathbf{A}^{(r)}, \mathbf{D}^{(r)}\}$, update $\{\mathcal{V}^{(r)}, \eta^{(r)}\}$ using (44);
        $\Gamma^{(r)} = \eta^{(r)} - \frac{1}{2}\varrho^{(l-1)} \sum_{g=1}^{G} \|\mathbf{w}_g^{(r)} - \mathbf{A}_g^{(r)}\mathbf{d}_g^{(r)}\|^2$;
    **until** $|\Gamma^{(r)} - \Gamma^{(r-1)}| < \epsilon_2$ or $r > r_{\max}$
    $\Xi^{(l)} = \max_{g \in \mathcal{G}} \{\|\mathbf{w}_g^{(r)} - \mathbf{A}_g^{(r)}\mathbf{d}_g^{(r)}\|_\infty\}$;
    **if** $\Xi^{(l)} < \epsilon_1$ **then**
        Set flag = 1;
    **else**:
        Update $\varrho^{(l)} = \varpi\varrho^{(l-1)}$;
    **end if**
**until** flag = 1 or $l > l_{\max}$
**Output**: A suboptimal solution $\{\mathbf{D}^{(r)}, \mathbf{A}^{(r)}, \mathcal{V}^{(r)}, \eta^{(r)}\}$.
```

### *C. Convergence and Complexity Analysis*

Here, we analyze the convergence and complexity of Algorithm 2 and Algorithm 3. For Algorithm 2, its convergence comes from the convergence of the MM algorithm [30]. Besides, the complexity of Algorithm 2 is given as $\mathcal{O}\left(N_{\text{RF}}^3 + I_1 N N_{\text{RF}}^2\right)$, where $I_1$ is the iterations of Algorithm 2 [30]. For Algorithm 3, we first show its convergence and then give the complexity. On one hand, similar to the problem ($\tilde{\text{P}}$1), the objective function of the problem ($\bar{\text{P}}$2) has an upper bound. On the other hand, for given $\rho^{(l-1)}$, we have:

$$\begin{aligned} &\Upsilon_1 : f(\mathbf{D}^{(r)}, \mathbf{A}^{(r-1)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}) \geq \\ &\qquad f(\mathbf{D}^{(r-1)}, \mathbf{A}^{(r-1)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}), \\ &\Upsilon_2 : f(\mathbf{D}^{(r)}, \mathbf{A}^{(r)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}) \geq \\ &\qquad f(\mathbf{D}^{(r)}, \mathbf{A}^{(r-1)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}), \\ &\Upsilon_3 : f(\mathbf{D}^{(r)}, \mathbf{A}^{(r)}, \mathcal{V}^{(r)}, \eta^{(r)}) \geq \\ &\qquad f(\mathbf{D}^{(r)}, \mathbf{A}^{(r)}, \mathcal{V}^{(r-1)}, \eta^{(r-1)}), \end{aligned} \tag{45}$$

where $\Upsilon_1$ holds because (33) is the optimal solution of the problem (P2.1), $\Upsilon_2$ holds due to the convergence of Algorithm 2, $\Upsilon_3$ holds due to the convergence of the SCA algorithm. Therefore, if the feasible region of the problem ($\bar{\text{P}}$2) is not empty, Algorithm 3 can converge to a local optimal solution of the problem ($\bar{\text{P}}$2) [27]. As for the complexity of Algorithm 3, it mainly comes from the matrix inversion in (33) whose complexity is $\mathcal{O}(GN_{\text{RF}}^3)$, the implementation of Algorithm 2 and the solving of the convex optimization problem in the problem ($\bar{\text{P}}$2.3). The standard convex optimization tools based on the interior point method (such as CVX [23]) can be used to obtain the optimal solution of the problem ($\bar{\text{P}}$2.3). Since the number of real optimization variables in the problem ($\bar{\text{P}}$2.3) is $2NG + 6N_{\text{U}} + 1$, the complexity of the interior point method can be expressed as $\mathcal{O}((2NG + 6N_{\text{U}} + 1)^{3.5} \ln(1/\delta))$, where $\delta$ is the solution accuracy [24]. Let $I_2$ and $I_3$ denote the outer and inner iterations of Algorithm 3, respectively, then the complexity of Algorithm 3 is given as $\mathcal{O}(I_2 I_3 (G N_{\rm RF}^3 + N_{\rm RF}^3 + I_1 N N_{\rm RF}^2 + (2NG + 6N_{\rm U} + 1)^{3.5} \ln(1/\delta)))$.

## VI. Simulation Results

In this section, we provide simulation results to evaluate the performance of NOMA-based mmWave-MEC systems in terms of CE and verify the convergence and effectiveness of the proposed CE optimization algorithms. In simulation, we consider that all users are uniformly distributed within the range of 10 m to 30 m from the BS, and the parameters of mmWave channel in (1) are given as [31]: mmWave carrier frequency $f_c = 28$ GHz, the number of LOS paths is 1, the path loss exponent of LOS is $\alpha^{\rm LOS} = 2$, the number of NLOS paths is 3, the path loss exponent of NLOS is $\alpha^{\rm NLOS} = 2.92$. Besides, the noise power spectral density is $n_0 = -174$ dBm. Unless otherwise specified, the default simulation parameters for the mmWave-MEC-ABF and the mmWave-MEC-HBF are listed in Table I and Table II, respectively [12]–[14].

TABLE I: Simulation parameters for mmWave-MEC with ABF

| Parameters | Default Values |
|---|---|
| Number of users | $K = 4$ |
| Number of antennas at the BS | $N = 16$ |
| System bandwidth | $B = 2$ MHz |
| Maximum power consumption | $P_k^{\max} = P_{\max}$ |
| Fixed power consumption | $P_{k,c} = 50$ mW |
| PA coefficient | $\zeta_k = 1/0.38$ |
| CPU chip coefficient | $\xi_k = 10^{-28}$ |
| CPU cycles per bit | $C_k = 10^3$ cycles/bit |
| Minimum computation-bit rate | $R_k^{\min} = 10^4$ bits/s |
| Maximum local CPU frequency | $f_k^{\max} = 1$ GHz |

TABLE II: Simulation parameters for mmWave-MEC with HBF

| Parameters | Default Values |
|---|---|
| Number of users | $K = 4$ |
| Number of antennas at the BS | $N = 32$ |
| Number of RFCs | $N_{\rm RF} = 2$ |
| Number of user groups | $G = 2$ |
| System bandwidth | $B = 2$ MHz |
| Maximum power consumption | $P_{g,k}^{\max} = P_{\max}$ |
| Fixed power consumption | $P_{g,k,c} = 50$ mW |
| PA coefficient | $\zeta_{g,k} = 1/0.38$ |
| CPU chip coefficient | $\xi_{g,k} = 10^{-28}$ |
| CPU cycles per bit | $C_{g,k} = 10^3$ cycles/bit |
| Minimum computation-bit rate | $R_{g,k}^{\min} = 10^4$ bits/s |
| Maximum local CPU frequency | $f_{g,k}^{\max} = 1$ GHz |

### A. Convergence behaviors of the algorithms

In this subsection, the convergence behaviors of the proposed two optimization algorithms (i.e., Algorithm 1 and Algorithm 3) are illustrated. Specifically, Fig. 3 and Fig. 4 show the convergence behaviors of the proposed Algorithm 1 and Algorithm 3, respectively, where $P_{\max} = 0.052$ W. It can be seen that the objective values of two algorithms can achieve convergence after a certain number of iterations. Meanwhile, the constraint violations of two algorithms tend to be smaller and finally decrease to a predefined acceptable level, e.g., $10^{-5}$, which indicates that the solution obtained by the algorithm is a feasible one of the problem. The above results verify the convergence of the proposed two CE optimization algorithms.

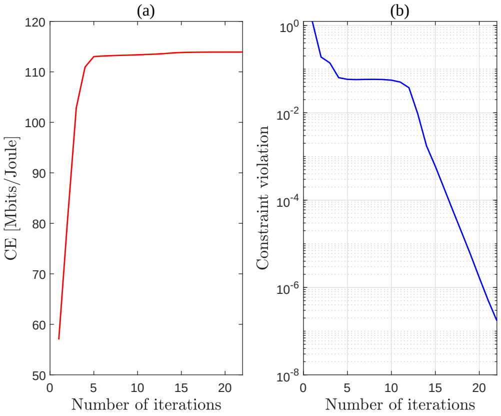


Fig. 3: Convergence behavior of Algorithm 1.

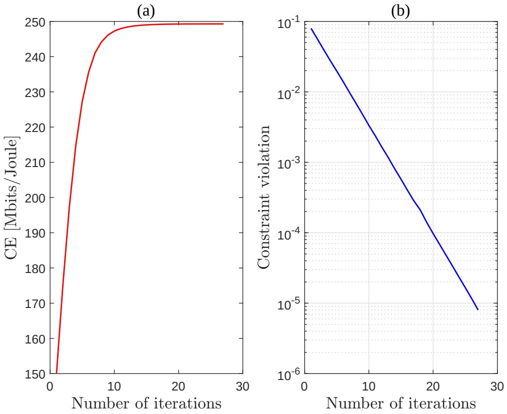


Fig. 4: Convergence behavior of Algorithm 3.

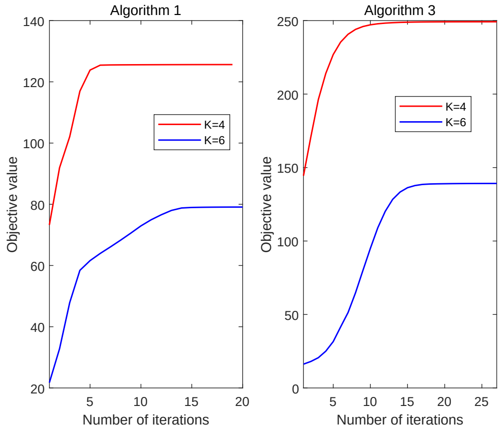


Fig. 5: Convergence behaviors of Algorithm 1 and Algorithm 3 under different numbers of users.

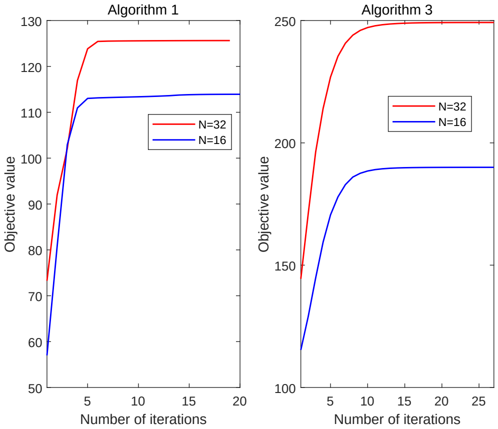
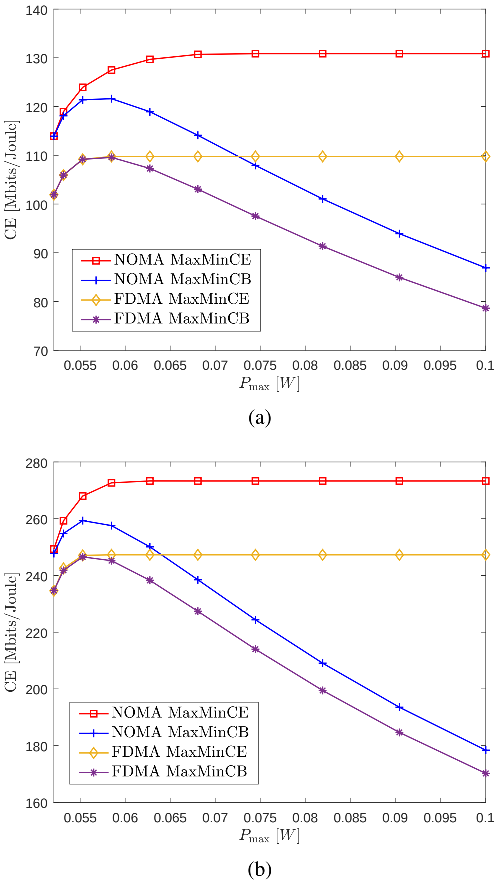


Fig. 6: Convergence behaviors of Algorithm 1 and Algorithm 3 under different numbers of antennas.

Fig. 5 illustrates the convergence behaviors of Algorithm 1 and Algorithm 3 with different number of users, where $P_{\max} = 0.052$W, $N = 32$, $K = 4, 6$. From Fig. 5, it is found that the CE is gradually increasing and finally saturated as the iteration increases. Namely, these two algorithms can converge to their respective stable points after some iterations. Thus, the convergence of Algorithm 1 and Algorithm 3 are guaranteed for different numbers of users. Besides, in Fig. 6, we give the convergence behaviors of Algorithm 1 and Algorithm 3 with different numbers of antennas, where $N = 16, 32$, and $P_{\max} = 0.052$W. From Fig. 6, we can observe the results similar to those in Fig. 5. Namely, for different numbers of antennas, Algorithm 1 and Algorithm 3 can still converge to their respective optimized values after a few number of iterations, and the CE with $N = 32$ is higher that with $N = 16$ after convergence, as expected. The results above further confirm that the proposed two algorithms can converge well under different system parameters.

### *B. CE performance with different multiple access schemes*

In this subsection, we provide the CE performances of the systems with different multiple access schemes, where NOMA, FDMA and TDMA are considered. In Fig. 7, the CE comparison between NOMA and FDMA with the MaxMinCE and MaxMinCB schemes under the partial offloading mode are shown, where the CEs of mmWave-MEC-ABF and mmWave-MEC-HBF are presented in Fig. 7(a) and Fig. 7(b), respectively. Specifically, each user is allocated by the same bandwidth for FDMA scheme in Fig. 7(a). While for FDMA scheme in Fig. 7(b), the users are divided into $G$ groups using the same user grouping strategy as NOMA, and users in the same group perform FDMA with equal bandwidth allocation [25]. On the one hand, it can be observed that the CE of NOMA scheme is significantly higher than that of FDMA scheme, which indicates the effectiveness of NOMA scheme in the mmWave-MEC system. The reason is that NOMA scheme allows the users to share the system bandwidth $B$ by multiplexing the power domain, which improves the achievable rates for the upload computing of the users, and thus it attains higher CE than FDMA scheme. On the other hand, the MaxMinCE scheme outperforms the MaxMinCB scheme in terms of CE. In particular, when $P_{\max}$ is small, due to the constraint of $P_{\max}$, the CE of the MaxMinCE scheme is slightly higher than that of the MaxMinCB scheme since they have similar resource allocation strategies. However, with the increase of $P_{\max}$, the CE of the MaxMinCE scheme rises to a stable level, but the CE of the MaxMinCB scheme first increases and then decreases. Different from the MaxMinCE scheme, the MaxMinCB scheme does not take into account the trade-off between the CBs and the power consumption. Thus, for the MaxMinCB scheme, the improvement of the CBs may be less than the improvement of the power consumption when $P_{\max}$ is large, which results in the decrease of CE.

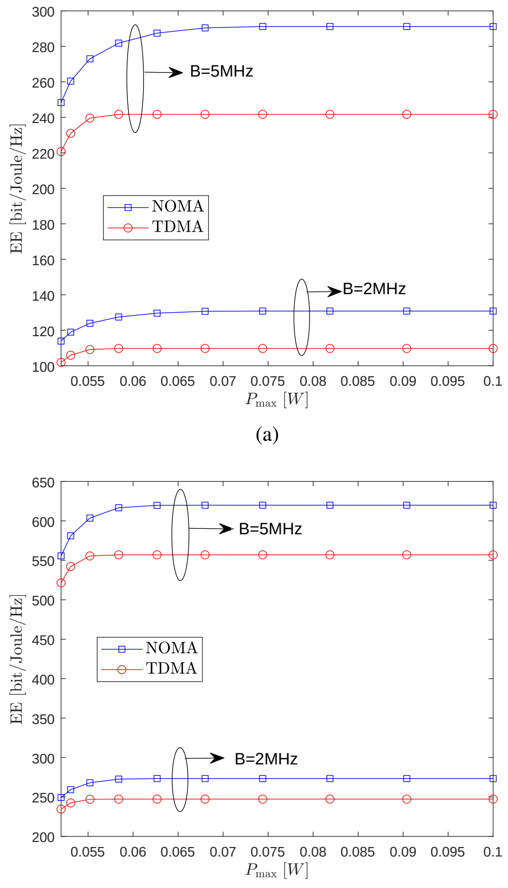


(a)

(b)

Fig. 7: (a) CE comparison between NOMA and FDMA in mmWave-MEC-ABF; (b) CE comparison between NOMA and FDMA in mmWave-MEC-HBF.

Fig. 8 gives the CE comparison between NOMA and TDMA schemes under the partial offloading mode with different bandwidths, where $B$=5MHz, 2MHz, the CEs of mmWave-MEC-ABF and mmWave-MEC-HBF are shown in Fig. 8(a) and Fig. 8(b), respectively. In particular, each user is allocated by the same time slot for TDMA scheme in Fig. 8(a). While for TDMA scheme in Fig. 8(b), the users are divided into $G$ groups with the same user grouping strategy as NOMA, and users in the same group perform TDMA with equal time slot allocation. As illustrated in Fig. 8, the NOMA-based mmWave-MEC system has better CE performance than



(a)

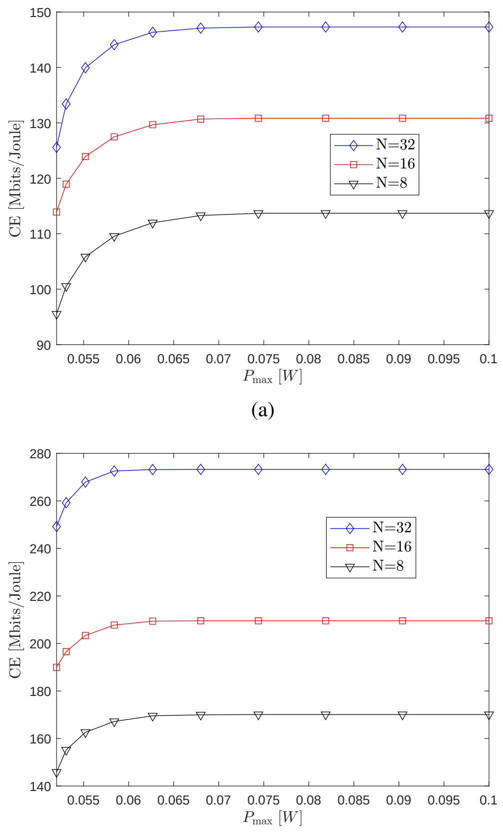
(b)

Fig. 8: (a) CE comparison between NOMA and TDMA in mmWave-MEC-ABF; (b) CE comparison between NOMA and TDMA in mmWave-MEC-HBF.

TDMA-based mmWave-MEC system. The reason is that NOMA scheme can allow the users to share the time resource of the system by multiplexing the power domain, i.e., higher multiplexing gain can be attained when NOMA is applied. Thus, it has higher CE than TDMA scheme. Besides, with the increase of bandwidth, the gap between the NOMA scheme and the TDMA scheme becomes larger. This is because NOMA scheme can upload more computational bits as the broadband increases under the same time resource block. Thus, the system CE is obviously increased.

## *C. CE performance with different system parameters and computing modes*

In this subsection, we give the CE performances of the systems with different system parameters and computing modes, where different antenna numbers, user numbers, computational capabilities, system bandwidths, RFC numbers, and computing modes are considered. In Fig. 9, we compare the CE performance of the systems with different numbers of antennas, where mmWave-MEC systems with ABF and HBF are respectively considered in Fig. 9(a) and Fig. 9(b), and the antenna number $N = 8, 16, 32$. From Fig. 9, it is found that the CE performance becomes better with the increase of the antenna number, as expected. This is because the higher diversity gain can be attained as the number of antennas increases. As a result, the CE performance is effectively improved. These results indicate that the proposed optimization schemes are effective for different antenna numbers.



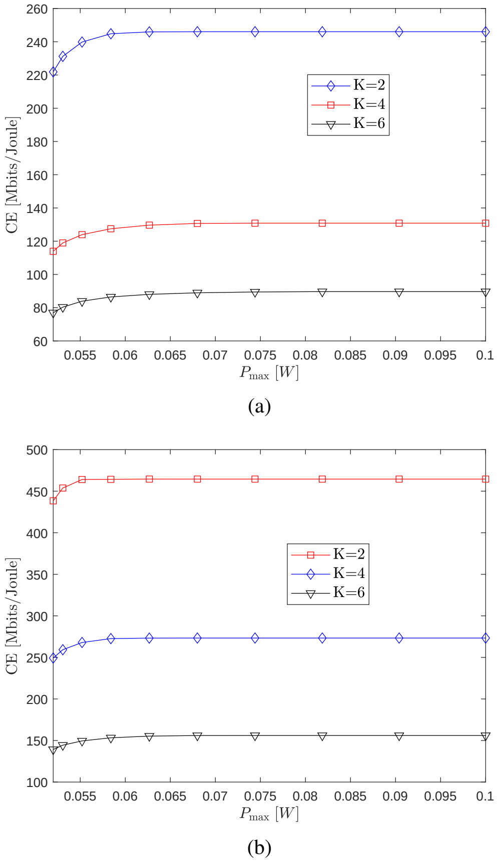
(a)

(b)

Fig. 9: (a) CE of mmWave-MEC-ABF with different antenna numbers; (b) CE of mmWave-MEC-HBF with different antenna numbers.

In Fig. 10(a) and Fig. 10(b), we plot the CE performances of mmWave-MEC-ABF and mmWave-MEC-HBF with different numbers of users, respectively, where $K = 2, 4, 6$. As shown in Fig. 10(a) and Fig. 10(b), with the increase of number of user $K$, the CE performance becomes worse. This is because the number of users with worse link increases when $K$ becomes larger, which results in the degradation of the CE performance. The results above show that the proposed optimization schemes are valid for different numbers of users.

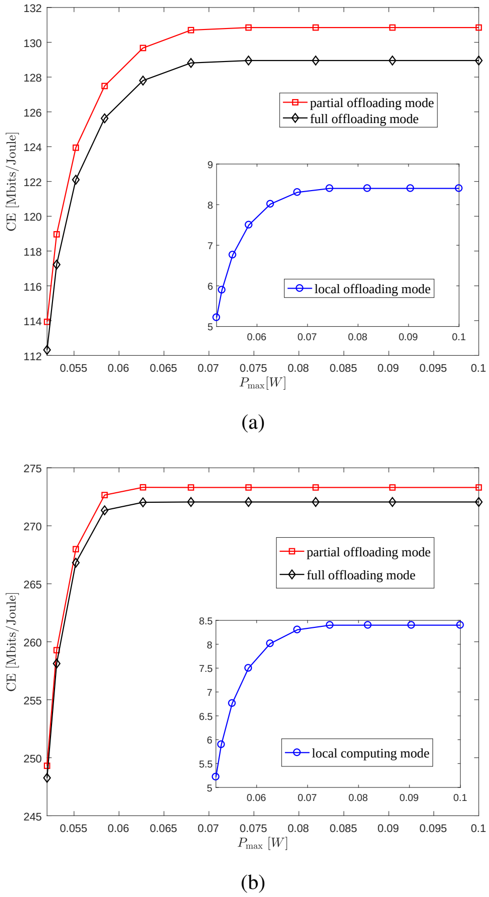
Fig. 11 compares different computing modes (i.e., the local computing mode, the partial offloading mode, and the full offloading mode) in terms of the system CE with ABF and HBF, where the resource allocation scheme under the local computing mode comes from [32], the resource allocation schemes under the partial offloading mode and the full offloading mode are obtained by using the framework of the proposed algorithms. It can be seen from Fig. 11 that the partial offloading mode has higher CE than the other two computing modes, especially it is much higher than the local computing mode. This is because the partial offloading mode



(a)

(b)

Fig. 10: (a) CE of mmWave-MEC-ABF with different user numbers; (b) CE of mmWave-MEC-HBF with different user numbers.



(a)

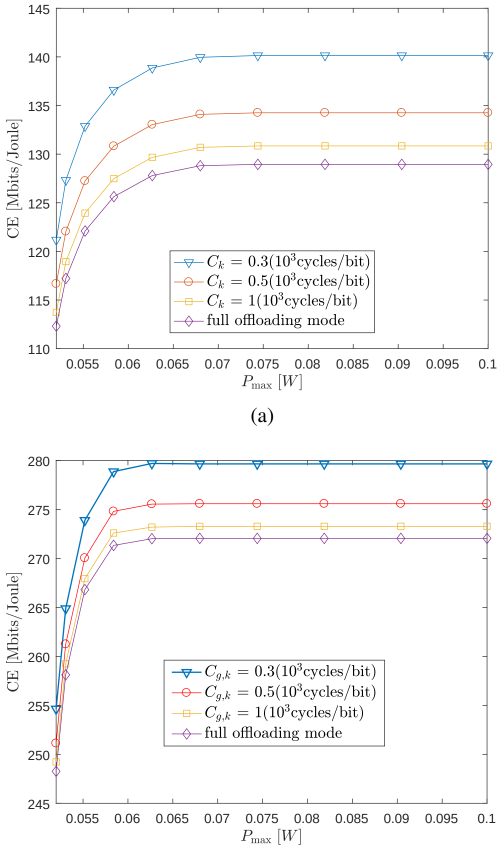
(b)

Fig. 11: (a) CE comparison among different computing modes in mmWave-MEC-ABF; (b) CE comparison among different computing modes in mmWave-MEC-HBF.

can dynamically adjust the resource allocation ratio of the local computing and upload computing. In fact, the partial offloading mode covers the local computing mode and the full offloading mode. Limited by the local computing capacities of the users, the CE of the local computing mode is the worst. Moreover, the CEs of the partial offloading mode and the full offloading mode are much higher than that of the local computing mode, which indicates that the application of mmWave-NOMA greatly improves the efficiency of the upload computing of the users. Therefore, it is beneficial to combine MEC with mmWave-NOMA.

In Fig. 12, we further compare the full offloading mode and partial offloading mode under different computational capabilities of users in terms of the system CE with ABF and HBF, where $C_k$= $\{0.3, 0.5, 1\} \times 10^3$ cycles/bit is considered in Fig. 12(a), and $C_{g,k}$= $\{0.3, 0.5, 1\} \times 10^3$ cycles/bit is considered in Fig. 12(b). From Fig. 12, it can be found that when $C_k$ and $C_{g,k}$ are small, the local processing capability will be strong, the computing task is preferred to be processed locally at the users since high time efficiency can be attained. Moreover, the computing tasks can be executed locally and on the MEC server simultaneously. Thus, the CE performance of partial offloading mode is improved greatly. While for full offloading mode, the energy consumption for data transmission is high and the corresponding savings accomplished by the computation offloading become low when the users have strong computational capabilities. Hence, the performance gap between partial offloading mode and full offloading mode increases with the decrease of $C_k$ and $C_{g,k}$.

Fig. 13 provides the CE comparison of full offloading mode and partial offloading mode with different system bandwidths, where mmWave-MEC-ABF and mmWave-MEC-HBF are respectively considered in in Fig. 13(a) and Fig. 13(b), the bandwidth $B$=3, 2, 0.5MHz. As shown in Fig. 13, the partial offloading mode still exhibits superior performance over full offloading mode, and has higher CE. This is because the computing task can be performed locally and on the MEC server simultaneously, which is energy-saving and then improves CE. With the increase of $B$, the CE performances under full offloading mode and partial offloading mode both improve, since large bandwidth can achieve high data rate, and thus increasing CE. Moreover, their CE gap becomes smaller due to relatively weaker computational capability of users when $B$ is larger. However, when $B$ is smaller, their CE gap will be larger because the achievable data rate becomes low for small bandwidth, which limits the performance of full offloading mode. Besides, from Fig. 11-Fig. 13, it is found that the full offloading mode can obtain the CE performance close



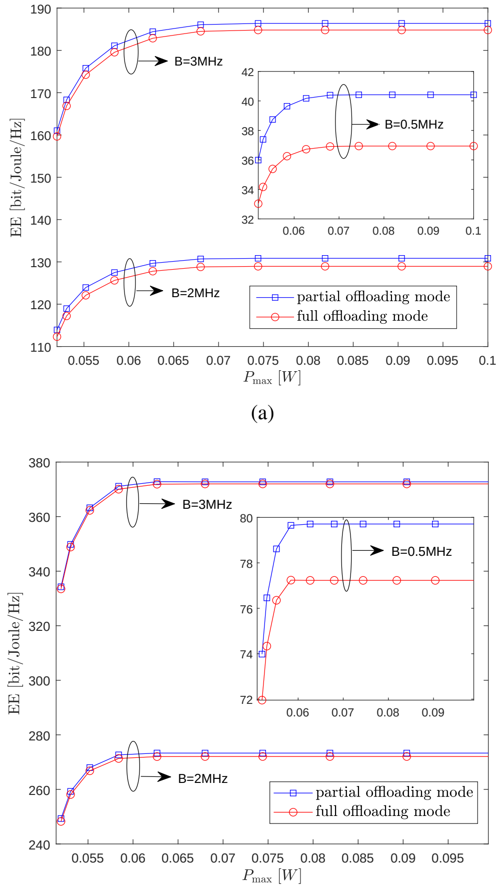
(a)

(b)

Fig. 12: (a) CE comparison of the ABF scheme under full offloading mode and partial offloading mode with different computational capabilities of user; (b) CE comparison of the HBF scheme under full offloading mode and partial offloading mode with different computational capabilities of user.

to that of partial offloading mode when the computational capabilities of users are weak and/or the system bandwidth is large. Conversely, its performance becomes worse.

Fig. 14(a) presents the CE comparison of the ABF scheme and the HBF scheme under the partial offloading mode, where the number of RFCs in the HBF scheme is $N_{\text{RF}} = 1$, and the parameters of the ABF scheme are consistent with those of the HBF scheme. When the number of RFCs is 1, the ABF matrix in the HBF scheme is actually degenerated into the ABF vector in the ABF scheme. However, the DBF matrix in the HBF scheme can provide additional performance gain and improve the CE of the HBF scheme. Thus, the CE of the HBF scheme is higher than that of the ABF scheme, which is verified by the simulation result in Fig. 14(a). Furthermore, Fig. 14(b) gives the CE of the partial offloading mode with different numbers of RFCs in the HBF scheme, where the number of RFCs is $N_{\text{RF}} \in \{1, 2, 3\}$. It can be seen from Fig. 14(b) that the CE is significantly improved with the increase of $N_{\text{RF}}$. This is due to that the larger $N_{\text{RF}}$ is, the more CBs can be transmitted, thereby improving the CE.



(a)

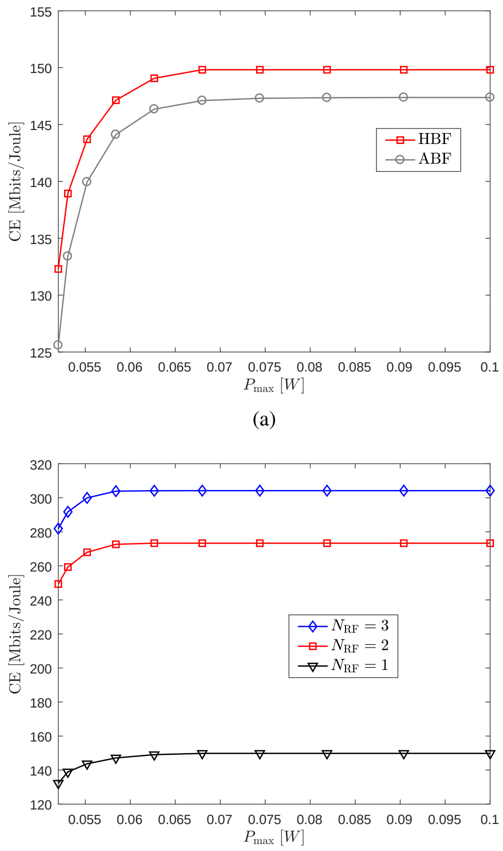
(b)

Fig. 13: (a) CE comparison of the ABF scheme under full offloading mode and partial offloading mode with different bandwidths; (b) CE comparison of the HBF scheme under full offloading mode and partial offloading mode with different bandwidths.

### D. CE performance with different optimization schemes

In this subsection, we provide the CE performances with different optimization schemes, where the proposed suboptimal scheme, the existing scheme, the upper bound of the suboptimal scheme are considered. In Fig. 15, we compare the CE performance between the presented NOMA-based mmWave-MEC network and the NOMA-based conventional MEC network without considering mmWave in [14], which are referred as “mmWave NOMA” and “non-mmWave NOMA” in Fig. 15, respectively, where different bandwidths (i.e., $B$=2, 3, 4MHz) are also considered. For the fairness, we set $N = 1$ and remove the ABF optimization for the proposed algorithm considering that the algorithm in [14] is based on single receive antenna. As illustrated in Fig. 15, the “mmWave NOMA” scheme has higher CE than the “non-mmWave NOMA” scheme due to the LoS existence of mmWave for the same bandwidth $B$=2MHz. Moreover, with the increase of the bandwidth, the “mmWave NOMA” scheme exhibits superior performance over the “non-mmWave NOMA” scheme, and their performance gap becomes larger as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/LCOMM.2023.3303091, IEEE Transactions on Mobile Computing



(a)

(b)

Fig. 14: (a) CE comparison between mmWave-MEC-ABF and mmWave-MEC-HBF; (b) CE versus the number of RFCs in mmWave-MEC-HBF.

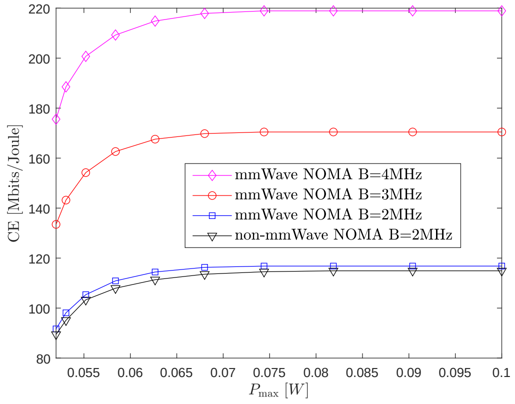


Fig. 15: CE comparison between NOMA-based mmWave-MEC system and NOMA-based MEC system.

well, especially when $B$=4MHz. This is due to the fact that the mmWave communication can possess larger bandwidth. Thus, much higher rate can be achieved, and corresponding higher CE is attained. The above results further demonstrate that the integration of the mmWave in the MEC system is beneficial.

To further evaluate the effectiveness of the proposed suboptimal optimization schemes, in Fig. 16, we give the upper

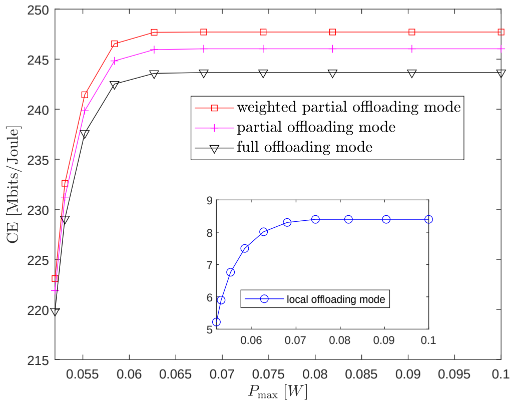


Fig. 16: CE comparison with different computing modes.

bound performance of the suboptimal scheme based on ABF considering the simplicity, where $K$=2. This upper bound performance can be attained by optimizing the following problem $\max_{\{\mathbf{w},p_k,f_k,e_k\}} \min_{k\in\mathcal{K}} \eta_k = \frac{e_k B\log_2(1+\mathrm{SINR}_k)+(1-e_k)f_k/C_k}{e_k\zeta_k p_k+(1-e_k)\left(\xi_k f_k^3+P_{k,c}\right)}$, where $e_k\in[0,1]$ is the weighted parameter and can be optimized to maximize the minimal CE of users. It is shown that the above problem includes the CE optimizations of three computing modes as special cases. For $e_k$=0 and $e_k$=1, the optimization is reduced to the one of local computing mode and the one of full offloading mode, respectively. When $e_k$=1/2, it is reduced to the optimization of our partial offloading mode. Since $\{e_k\}$ are generated by the optimization and not fixed, the obtained CE is higher than the above three modes. For the optimization of $\{e_k\}$, we can employ the multidimensional search method to find their values, and for each search, Algorithm 1 is used to achieve the suboptimal solution of other optimized parameters. Thus, the superior CE performance can be attained, but the complexity will be much higher since it needs to perform $K$ dimensional search. This weighted optimization scheme is referred as "weighted partial offloading mode" for ease of comparison with other modes. As seen from Fig. 16, the proposed scheme has the performance close to that of its upper bound, and is better than that under the full offloading mode and local computing mode. These results above further verify the effectiveness of the proposed scheme.

## VII. Conclusion

In this paper, we have studied the CE optimization in mmWave-MEC with NOMA, where both of the ABF and HBF architectures are considered. For mmWave-MEC with ABF, a CE optimization problem based on the max-min fairness criterion is formulated to jointly optimize the ABF vector at the BS and the local resource allocation of each user. A PSCA-based CE optimization algorithm is proposed to obtain a local optimal solution of this non-convex problem. Besides, the max-min CE optimization problem for mmWave-MEC with HBF is further studied by jointly optimizing the HBF at the BS and the local resource allocation of each user. A double-loop CE optimization algorithm based on the PIBCD algorithm is developed to tackle this challenging problem. Simulation results verify the convergence of the proposed algorithms and demonstrate the effectiveness of the proposed computation-efficient resource allocation schemes. As a result, the superior CE performance is achieved.

## Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments which improve the quality of this paper greatly.

## References

[1] H. Zhang, J. Li, B. Wen, Y. Xun, and J. Liu, “Connecting intelligent things in smart hospitals using NB-IoT,” *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1550–1560, 2018.

[2] A. N. Uwaechia and N. M. Mahyuddin, “A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges,” *IEEE Access*, vol. 8, pp. 62 367–62 414, 2020.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” *IEEE Commun. Sur. Tuts.*, vol. 19, no. 2, pp. 721–742, 2017.

[4] V. Frascolla *et al.*, “5G-MiEdge: Design, standardization and deployment of 5G phase II technologies: MEC and mmwaves joint development for Tokyo 2020 olympic games,” in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, 2017, pp. 54–59.

[5] Q. Pham *et al.*, “A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art,” *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[6] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, “Mobile edge computing: A survey,” *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, 2018.

[7] P. Mach and Z. Becvar, “Mobile edge computing: A survey on architecture and computation offloading,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 2017.

[8] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[9] C. Zhao, Y. Cai, M. Zhao, and Q. Shi, “Joint hybrid beamforming and offloading for mmWave mobile edge computing systems,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–6.

[10] C. Zhao, Y. Cai, A. Liu, M. Zhao, and L. Hanzo, “Mobile edge computing meets mmWave communications: Joint beamforming and resource allocation for system delay minimization,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2382–2396, 2020.

[11] Z. Ding, P. Fan, and H. V. Poor, “Impact of non-orthogonal multiple access on the offloading of mobile edge computing,” *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, 2019.

[12] H. Sun, F. Zhou, and R. Q. Hu, “Joint offloading and computation energy efficiency maximization in a mobile edge computing system,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3052–3056, 2019.

[13] Y. Wu, Y. Wang, F. Zhou, and R. Qingyang Hu, “Computation efficiency maximization in OFDMA-based mobile edge computing networks,” *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 159–163, 2020.

[14] F. Zhou and R. Q. Hu, “Computation efficiency maximization in wireless-powered mobile edge computing networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3170–3184, 2020.

[15] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, “Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6177–6189, 2018.

[16] X. Yu, F. Xu, K. Yu, and N. Li, “Joint energy-efficient power allocation and beamforming for uplink mmWave-NOMA system,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12291–12295, 2020.

[17] J. C. Guo, Q. Y. Yu, W. X. Meng, and W. Xiang, “Energy-efficient hybrid precoder with adaptive overlapped subarrays for large-array mmWave systems,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1484–1502, 2020.

[18] M. Razaviyayn, “Successive convex approximation: Analysis and applications,” 2014.

[19] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[20] W. Hao *et al.*, “Codebook-based max-min energy-efficient resource allocation for uplink mmWave MIMO-NOMA systems,” *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8303–8314, 2019.

[21] T. Lipp and S. Boyd, “Variations and extension of the convex-concave procedure,” *Optim. Eng.*, vol. 17, no. 2, pp. 263–287, 2016.

[22] X. Yu, K. Alhujaili, G. Cui, and V. Monga, “MIMO Radar waveform design in the presence of multiple targets and practical constraints,” *IEEE Trans. Signal Process.*, vol. 68, pp. 1974–1989, 2020.

[23] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” 2014.

[24] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[25] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, “Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, 2019.

[26] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, “Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3000–3004, 2019.

[27] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.

[28] Y. Yang, M. Pesavento, Z. Luo, and B. Ottersten, “Inexact block coordinate descent algorithms for nonsmooth nonconvex optimization,” *IEEE Trans. Signal Process.*, vol. 68, pp. 947–961, 2020.

[29] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconfigurable intelligent surfaces for energy efficiency in wireless communication,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, 2019.

[30] Y. Sun, P. Babu, and D. P. Palomar, “Majorization-minimization algorithms in signal processing, communications, and machine learning,” *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, 2017.

[31] M. R. Akdeniz *et al.*, “Millimeter wave channel modeling and cellular capacity evaluation,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, 2014.

[32] F. Xu, X. Yu, J. Cai, and G. Wang, “Computation efficiency optimization in UAV-enabled mobile edge computing system with multi-carrier non-orthogonal multiple access,” *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, pp. 1–22, 2020.