A Max-Min Task Offloading Algorithm for Mobile Edge Computing Using Non-Orthogonal Multiple Access

Vaibhav Kumar , *Member, IEEE*, Muhammad Fainan Hanif ,
Markku Juntti , *Fellow, IEEE*,
and Le-Nam Tran , *Senior Member, IEEE*

Abstract—To mitigate the computational power gap between the network core and edges, mobile edge computing (MEC) is poised to play a fundamental role in future generations of wireless networks. In this correspondence, we consider a non-orthogonal multiple access (NOMA) transmission model to maximize the worst task to be offloaded among all users to the network edge server. A provably convergent and efficient algorithm is developed to solve the considered non-convex optimization problem for maximizing the minimum number of offloaded bits in a multi-user NOMA-MEC system. Compared to the approach of optimized orthogonal multiple access (OMA), for given MEC delay, power and energy limits, the NOMA-based system considerably outperforms its OMA-based counterpart in MEC settings. Numerical results demonstrate that the proposed algorithm for NOMA-based MEC is particularly useful for delay sensitive applications.

Index Terms—Non-orthogonal multiple access, mobile edge computing, task offloading, B5G, 6G.

I. INTRODUCTION

Non-orthogonal multiple access (NOMA) has gained tremendous popularity in the past few years as a potential scheme to facilitate massive connectivity, as well as to support low-latency communications. On the other hand, mobile edge computing (MEC) is being considered as a key enabler for the communication among devices with low data-processing capability, where such devices offload their computationally-intensive tasks to an edge server. The superiority of NOMA-based MEC has been proven over its orthogonal multiple access (OMA) based counterparts, in terms of energy and/or delay minimization.

In this context, the problem of energy consumption minimization in NOMA-MEC system was explored in [1], [2], [3], [4]. On the other hand, the authors in [5], [6], [7], [8], [9] considered the problem of delay/latency minimization in NOMA-MEC system. More interestingly, in [10], the optimization problem for a two-user NOMA-MEC system was formulated as a Stackelberg game where the users tend to minimize their energy consumption, whereas the MEC server attempts to minimize the task execution time. The problem of system cost minimization in an ultra-dense NOMA-MEC network was discussed in [11]. The problem of minimizing a weighted sum of energy consumption and

latency in a NOMA-MEC system was considered in [12] and an efficient solution was obtained using deep reinforcement learning (DRL) approach. The problem of minimizing the total energy consumption in a NOMA-MEC system using DRL in the context of industrial Internet of Things was presented in [13].

In the case of a NOMA-MEC system, the users often partition their computation tasks in two categories: low-complexity and high-complexity tasks. The low-complexity tasks are executed locally, while the high-complexity tasks are offloaded to the MEC server. Therefore, different from the existing literature, we consider the problem of maximizing the minimum number of bits (of the high-complexity tasks) offloaded by users to an MEC server using NOMA by constraining the total energy, transmission power and the offloading delay within given limits. In this context, Huang et al. studied a max-min computation efficiency problem in NOMA-MEC system [14], and a similar problem in a millimeter wave settings was explored in [15]. In order to improve the physical-layer security of offloaded data in a NOMA-MEC system, a max-min antieavesdropping poroblem was presented in [16].

Our system model is unique in a sense that it exploits NOMA to offload data of multiple users in a single time slot for mobile edge computations without impacting the performance of users. We solve the formulated max-min non-convex problem using the framework of successive convex approximation (SCA). A provably convergent algorithm is derived based on novel techniques to bound non-convex functions appearing in the main optimization problem. Precisely, our main contributions include: i) considering a novel model that integrates NOMA with MEC in a multi-user scenario, the main problem of maximizing the number of bits offloaded by the worst user to the MEC server using NOMA is formulated; ii) deriving a tractable and provably convergent algorithm using novel bounding techniques and based on successive convex optimization strategy to solve the formulated problem, and iii) numerical demonstration of the superiority of the proposed solution compared to the corresponding orthogonal multiple access (OMA) based approach.

II. SYSTEM MODEL AND PROBLEM STATEMENT

Consider a NOMA-MEC system consisting of a base station (BS) which is assumed to have an integrated MEC server, and a set of M users, denoted by $\mathcal{U} \triangleq \{U_1, U_2, \ldots, U_M\}$. It is assumed that the BS and all users are each equipped with a single antenna. Furthermore, we assume that the users have very limited data processing capability, and therefore, they offload their computationally-intensive tasks to the MEC server. The channel gain between U_m ($m \in \mathcal{M} \triangleq \{1, 2, \ldots, M\}$) and the BS is denoted by g_m. The deadline for U_m to offload its task to the MEC server is denoted by D_m, and without loss of generality, we consider $D_1 \leq D_2 \leq \cdots \leq D_M$. It is also assumed that a user stops transmitting immediately even if it completely offloads its data before the assigned deadline. In particular, after U_{m-1} finishes offloading its data, U_m, which also transmits in previous users' time slot, is allowed some additional time $\bar{D}_m$ to finish offloading its data until the hard deadline D_m, such that $\sum_{j=1}^{m} \bar{D}_j \leq D_m$.[1] As a result, users m to M simultaneously offload their data till $\bar{D}_m$ to the MEC server by means of NOMA. A timing diagram for an example of data offloading in a three-user NOMA-MEC system is shown in Fig. 1. For sum rate,

---

[1]To simplify the exposition, throughout the article $\bar{D}_1$ is understood as D_1, if not mentioned otherwise.
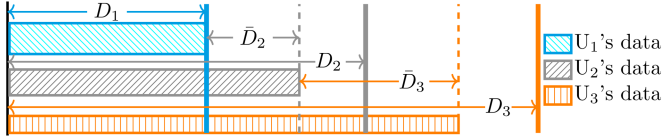
Fig. 1. Timing diagram of data offloading in a three user NOMA-MEC system. The duration shown by D_1, D_2 and D_3 denote the hard deadlines, and those shown by $\bar{D}_2$ and $\bar{D}_3$ denote the additional times.

an arbitrary decoding order can be used for successive interference cancellation (SIC) in uplink NOMA [9], [17]. Nonetheless, optimal decoding order in the NOMA uplink with fairness consideration is relatively unexplored [18]. For the purpose of this work, we have fixed user decoding order according to their delay sensitivity, i.e., we consider SIC with respect to users' delay ordering such that higher-indexed users are decoded first, similar to the arguments in [2], [5], [9] and [19]. In this way, the lower-indexed users (i.e., with stricter offloading deadlines) do not see interference from the higher indexed ones that transmit in their time slots. It should be emphasized here that NOMA is a natural candidate to integrate with MEC for the type of system model considered in our article. Particularly, since users also transmit in previous users' time slot, NOMA, with its interference cancellation and power adaptation capabilities, is required for the smooth operation of users impacted by such interference.

The transmit power of U_m during $\bar{D}_j$, $j \in \{1, 2, \ldots, m\}$, is denoted by P_{mj}. Also, by properly normalizing the channel gains, we assume that the noise at the BS is distributed as zero-mean unit-variance circularly symmetric complex Gaussian random variable, $\mathcal{CN}(0, 1)$. Therefore, letting B denote the total available bandwidth, the data offloaded (in nats) by U_m to the MEC server is given by

$$N_{\text{off},m} = B \sum_{j=1}^{m} \bar{D}_j \ln \left(1 + \frac{g_m P_{mj}}{\sum_{i=j}^{m-1} g_i P_{ij} + 1} \right), m \in \mathcal{M}. \quad (1)$$

Note that the expression for the rate of U_m (i.e., $\ln(1 + \frac{g_m P_{mj}}{\sum_{i=j}^{m-1} g_i P_{ij}+1})$) in (1) satisfies the condition for (deadline-dependent) SIC ordering as discussed in [2] and [5].

In this article, we are interested to find the power allocation P_{mj} and additional times $\bar{D}_m$ to maximize the worst task to be offloaded among all users, which is formulated as

$$\underset{\{P_{mj} \geq 0, \bar{D}_j \geq 0\}}{\text{maximize}} \quad \underset{m \in \mathcal{M}}{\min} N_{\text{off},m}, \quad (2a)$$

$$\text{subject to} \quad \sum_{m=1}^{M} \sum_{j=1}^{m} \bar{D}_j P_{mj} \leq E_{\text{th}}, \quad (2b)$$

$$\sum_{j=1}^{m} \bar{D}_j \leq D_m, m \in \mathcal{M} \setminus \{1\}, \quad (2c)$$

$$P_{mj} \leq P_m, j \in \mathcal{M}, m \in \{j, j+1, \ldots, M\}. \quad (2d)$$

It is noteworthy that (2) is different from the conventional max-min rate optimization problem due to the presence of $\bar{D}_j$ in (1), which is one of the optimization variables. The main motivation for considering this problem stems from promoting fairness among users to offload their data to the MEC server for computational purposes. If this consideration is ignored, the users that are more delay sensitive, and are unable to offload more data to meet delay constraints due to impeding channel

conditions, suffer the most.[2] We further remark that, to the best of our knowledge, (2) has not been studied previously. In addition, the formulation of (2) is different from that of classical uplink NOMA systems [20] both in terms of its objective and constraints. In (2b) we have constrained the total energy to E_{th}, (2c) ensures that U_m completes its data offloading before its assigned deadline D_m (with $\bar{D}_1 = D_1$, which is not an optimization variable), and in (2d) the transmit power of U_m during any time slot j with $j \in \mathcal{M}$, $m \in \{j, j+1, \ldots, M\}$ is constrained to P_m. It is also important to note that the constraint in (2c) does not prevent a user with a longer deadline from offloading its data before a user with a shorter one. Specifically, such a scenario will occur when the interference encountered by a higher-indexed user in time slots of lower indexed users is small enough so that it is able to offload its data even before the lower indexed users have completed offloading their tasks. If such a scenario does not prevail, the higher indexed users are allowed to continue transmitting their data till the given task offloading deadline. We remark that unlike downlink NOMA, in an uplink NOMA based system an explicit SIC constraint is not required as each user is to be decoded only once by the BS [21]. To keep the problem formulation more tractable, computation rates of user devices and MEC server have not been included as optimization variables in the problem formulation. Moreover, in line with the approximation used in existing works [2], [5], [7], [19] we have ignored the processing energy and time to send data back to users from the MEC server in our problem statement. Inclusion of the mentioned above parameters is a ripe direction for future research.

III. PROPOSED SOLUTION

In this section, we propose an efficient solution to problem (2) and prove the convergence of the proposed algorithm. Moreover, for a fair comparison, we also discuss a corresponding optimized OMA-MEC system. Since the problem in (2) is non-convex, to solve it we approximate the non-convex constraints with convex ones. Afterwards, we use the SCA framework to develop an iterative procedure to solve the original problem. One of novel features of this work lies in how we develop approximating functions that satisfy certain conditions within the SCA framework, as we see below.

For ease of description, we define $\mathbf{p}_{mj} \triangleq [P_{jj}, P_{(j+1)j}, \ldots, P_{mj}]^\mathsf{T}$ for $m \geq j$, which consists of the power allocation coefficients of U_j to U_m during $\bar{D}_j$, and $\mathbf{d} \triangleq [\bar{D}_2, \bar{D}_3, \ldots, \bar{D}_M]^\mathsf{T}$. It is obvious that the non-convexity of (2) is due to that of the objective in (2a), and the constraint in (2b). Let us consider an equivalent reformulation of (2) given by

$$\underset{\mathbf{p} \geq 0, \mathbf{d} \geq 0, \ell}{\text{maximize}} \quad \ell \quad (3a)$$

$$\text{subject to} \quad N_{\text{off},m} \geq \ell, m \in \mathcal{M}, \quad (3b)$$

$$\text{(2b), (2c), (2d)} \quad (3c)$$

where

$$\mathbf{p} = \left[\mathbf{p}_{MM}^\mathsf{T}, \ldots, \mathbf{p}_{M1}^\mathsf{T}, \mathbf{p}_{(M-1)(M-1)}^\mathsf{T}, \ldots, \mathbf{p}_{(M-1)1}^\mathsf{T}, \ldots, \mathbf{p}_{11}^\mathsf{T} \right]^\mathsf{T}.$$

It is clear that the non-convexity of (2) is carried over into (3b) and (2b). Let us focus on (3b) first. In light of SCA method, to deal with (3b) we need to find *a concave lower bound* of $N_{\text{off},m}$. In the related

---

[2]If resource allocation for all offloading users is to be considered on individual basis, the cost function of weighted sum of users' data to be offloaded may instead be considered. This problem is beyond the scope of current work and thus is left for future work.

literature, a popular method is to introduce some auxiliary variables to make (3b) more tractable, which eventually increases the complexity of the convex subproblems. Instead, to approximate (3b) we present the following theorem.

Theorem 1: For given $x_0 \geq 0$, $u_0 \geq 0$, and $v_0 > 0$, the inequality in (4) shown at the bottom of this page, holds for any $u \geq 0$ and $v > 0$.

Proof: See Appendix A. ∎

We also remark that the derivative of both sides of (4) with respect to each variable involved is identical, which is another requirement for the use of a lower bound in light of the SCA principles.

Let $\mathcal{I}_{mj} \triangleq \sum_{i=j}^{m-1} g_i P_{ij} + 1$, which is an affine function of $\{P_{ij}\}_{i=j}^{m-1}$. Then (1) is equivalent to

$$N_{\text{off},m} = B \sum_{j=1}^{m} \bar{D}_j \ln\left(1 + g_m P_{mj}/\mathcal{I}_{mj}\right). \tag{5}$$

Now, let $P_{mj}^{(n)}$ and $\bar{D}_j^{(n)}$ be the values of P_{mj} and $\bar{D}_j$ in iteration n of the iterative process presented in the discussion to follow, respectively. Then, applying Theorem 1 we can approximate (3b) as

$$B \sum_{j=1}^{m} f_{mj}\left(\mathbf{p}_{mj}, \bar{D}_j; \mathbf{p}_{mj}^{(n)}, \bar{D}_j^{(n)}\right) \geq \ell, m \in \mathcal{M}, \tag{6}$$

where $f_{m,j}\left(\mathbf{p}_{mj}, \bar{D}_j; \mathbf{p}_{mj}^{(n)}, \bar{D}_j^{(n)}\right)$ is given by (7) shown at the bottom of this page, $\mathcal{I}_{mj}^{(n)} \triangleq \sum_{i=j}^{m-1} P_{ij}^{(n)} g_i + 1$, $a_{mj}^{(n)} \triangleq 0.5(1 - \bar{D}_j^{(n)}) - 0.25(\bar{D}_j^{(n)} - \mathcal{I}_{mj}^{(n)})^2/\mathcal{I}_{mj}^{(n)}$, $b_{mj}^{(n)} \triangleq 2 + \ln(1 + g_m P_{mj}^{(n)}/\mathcal{I}_{mj}^{(n)}) + 0.5(\bar{D}_j^{(n)} - 1) + 0.5(-1 + \bar{D}_j^{(n)}/\mathcal{I}_{mj}^{(n)})$, $c_{mj}^{(n)} \triangleq 0.25(\bar{D}_j^{(n)} - 1)^2/(g_m P_{mj}^{(n)} + \mathcal{I}_{mj}^{(n)}) + 0.5(-1 + \bar{D}_j^{(n)}/\mathcal{I}_{mj}^{(n)}), d_{mj}^{(n)} \triangleq 0.25(\bar{D}_j^{(n)} - 1)^2 g_m/(g_m P_{mj}^{(n)} + \mathcal{I}_{mj}^{(n)})$, and $e_{mj}^{(n)} \triangleq 0.25(g_m P_{mj}^{(n)} + \mathcal{I}_{mj}^{(n)})$.

We remark that (6) can be expressed as a few simple second-order cone (SOC) constraints. To this end, auxiliary variables are introduced to obtain a system of linear and quadratic inequalities such that the quadratic inequalities are transformed to SOC constraints using simple manipulations [22]. This is accomplished automatically by a modeling tool such as CVX.

Our next step is to approximate (2b) which requires a convex upper bound of the term $\bar{D}_j P_{mj}$. There are several ways to achieve this. For example, the authors in [22] used $\bar{D}_j P_{mj} \leq \frac{\phi}{2} \bar{D}_j^2 + \frac{1}{2\phi} P_{mj}^2$, which holds good for any $\phi > 0$, and the equality is achieved when $\phi = P_{mj}/\bar{D}_j$. However, in this article $\bar{D}_j$ and/or P_{mj} *can approach zero*, which may lead to some numerical issues if the above convex upper bound is used. Thus, in this article we simply use the following convex upper bound

$$\bar{D}_j P_{mj} \leq 0.25\left(\bar{D}_j + P_{mj}\right)^2 + 0.25\left(\bar{D}_j^{(n)} - P_{mj}^{(n)}\right)^2 - 0.5$$

$$\times \left(\bar{D}_j^{(n)} - P_{mj}^{(n)}\right)\left(\bar{D}_j - P_{mj}\right) \triangleq q_{mj}\left(P_{mj}, \bar{D}_j; P_{mj}^{(n)}, \bar{D}_j^{(n)}\right), \tag{8}$$

Algorithm 1: Offloading Rate Maximization Algorithm.

Input: $P_{mj}^{(0)}$, $\bar{D}_j^{(0)}; \forall m, j \in \mathcal{M}$
Output: $P_{mj}^{\star}, D_j^{\star}; \forall m, j \in \mathcal{M}$

1 $n \leftarrow 0$;
2 **repeat**
3   Solve (10) and denote the obtained solution as $P_{mj}^{\star}$ and $D_j^{\star}$;
4   Update: $P_{mj}^{(n+1)} \leftarrow P_{mj}^{\star}$, $\bar{D}_j^{(n+1)} \leftarrow D_j^{\star}$;
5     $n \leftarrow n + 1$
6 **until** *convergence*;

---

which approximates (2b) as

$$\sum_{m=1}^{M} \sum_{j=1}^{m} q_{mj}(P_{mj}, \bar{D}_j; P_{mj}^{(n)}, \bar{D}_j^{(n)}) \leq E_{\text{th}}. \tag{9}$$

We remark that $q_{mj}(P_{mj}, \bar{D}_j; P_{mj}^{(n)}, \bar{D}_j^{(n)})$ is in fact symbolic representation of a convex quadratic function and thus (9) is an SOC constraint. From the convex approximation of (3b) and (2b) presented above, we arrive at the following second-order cone program (SOCP):

$$\underset{\mathbf{p} \geq 0, \mathbf{d} \geq 0, \ell}{\text{maximize}} \quad \ell \tag{10a}$$

$$\text{subject to} \quad (2c), (2d), (6), (9). \tag{10b}$$

Now we are in a position to present an algorithmic framework to solve our main problem in (2), using the approximation developed in (10). The procedure is outlined in Algorithm 1. We simply set $P_{mj}^{(0)} = 0$ and $\bar{D}_j^{(0)} = 0$ as a starting feasible point. It is quite insightful to observe that for each run of the proposed algorithm, we solve a convex problem whose feasible set is a subset of the original problem in (2). This occurs due to the fact that we approximate the non-convex function in (5) using a lower bound as given in (6). It not only ensures that solution in each iteration of Algorithm 1 is feasible to (2), but also results in convergence as we see in the next section.

A. Convergence of Algorithm 1

Let the feasible set of (3) be denoted by $\mathcal{F}$; $\mathcal{F}^{(n)} \triangleq \{(\mathbf{p}, \mathbf{d}, \ell)$ satisfying constraints in (10)$\}$, and $\ell^{(n)}$ denote the feasible set and the optimal objective at the nth iteration, respectively. Let the *non-convex constraints* in (3) be compactly denoted by $\psi_i(\mathbf{p}, \mathbf{d}, \ell)$, and the corresponding approximated convex constraints in (10) by $\Psi_i(\mathbf{p}, \mathbf{d}, \ell; \mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)})$. It is straightforward to note that due to the proposed bounds, the convex constraints used to approximate the non-convex ones satisfy the following three properties

$$\Psi_i\left(\mathbf{p}, \mathbf{d}, \ell; \mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right) \geq \psi_i\left(\mathbf{p}, \mathbf{d}, \ell\right), \forall \mathbf{p}, \mathbf{d}, \ell \tag{11a}$$

$$x \ln\left(1 + \frac{u}{v}\right) \geq \frac{1 - x_0}{2} - \frac{(x_0 - v_0)^2}{4v_0} + \left[2 + \ln\left(1 + \frac{u_0}{v_0}\right) + \frac{x_0 - 1}{2} + \frac{1}{2}\left(\frac{x_0}{v_0} - 1\right)\right] x - \left[\frac{(x_0 - 1)^2}{4(u_0 + v_0)} + \frac{1}{2}\left(\frac{x_0}{v_0} - 1\right)\right] v$$

$$- \frac{(x_0 - 1)^2}{4(u_0 + v_0)} u - \frac{(x + v)^2}{4v_0} - \frac{(u_0 + v_0)(x + 1)^2}{4(u + v)}. \tag{4}$$

$$f_{m,j}\left(\mathbf{p}_{mj}, \bar{D}_j; \mathbf{p}_{mj}^{(n)}, \bar{D}_j^{(n)}\right) = a_{mj}^{(n)} + b_{mj}^{(n)} \bar{D}_j - c_{mj}^{(n)} \mathcal{I}_{mj} - d_{mj}^{(n)} P_{mj} - \frac{(\bar{D}_j + \mathcal{I}_{mj})^2}{4\mathcal{I}_{mj}^{(n)}} - e_{mj}^{(n)} \frac{(\bar{D}_j + 1)^2}{(g_m P_{mj} + \mathcal{I}_{mj})}. \tag{7}$$

$$\Psi_i(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}; \mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}) = \psi_i\left(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right) \quad (11b)$$

$$\nabla\Psi_i(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}; \mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}) = \nabla\psi_i\left(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right). \quad (11c)$$

First of all we note that $\mathcal{F}^{(n)}$ is a subset of the feasible set of the original problem due to (11a). Further, as a consequence of (11b) and the update rule given in Algorithm 1, the sequence $\left(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right)$ is feasible to both $\mathcal{F}^{(n)}$ and $\mathcal{F}^{(n+1)}$. Since $\left(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right)$ is feasible to $\mathcal{F}^{(n+1)}$, the optimal variables corresponding to $\mathcal{F}^{(n+1)}$ should yield an improved objective value in the $(n+1)^{\text{th}}$ run of the algorithm i.e., $\ell^{(n)} \leq \ell^{(n+1)}$. It is easy to see that $\mathcal{F}$ is a compact set, and thus, the sequence $\{\ell^{(n)}\}$ converges.

We also remark that Slater's condition holds for the subproblem in all iterations. This can be shown by choosing sufficiently small $(\mathbf{p}, \mathbf{d}, \ell)$. Thus, the Karush-Kuhn-Tucker (KKT) conditions must be satisfied for (10) in all iterations. Let $\left(\mathbf{p}^*, \mathbf{d}^*, \ell^*\right)$ be the limit of the sequence $\left(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right)$.[3] Without loss of generality we assume that $\left(\mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}\right) \rightarrow \left(\mathbf{p}^*, \mathbf{d}^*, \ell^*\right)$. Otherwise we can restrict to a subsequence. Thus it holds that $\Psi_i(\mathbf{p}^{(n+1)}, \mathbf{d}^{(n+1)}, \ell^{(n+1)}; \mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}) \rightarrow \Psi_i(\mathbf{p}^*, \mathbf{d}^*, \ell^*; \mathbf{p}^*, \mathbf{d}^*, \ell^*) = \psi_i(\mathbf{p}^*, \mathbf{d}^*, \ell^*)$. Also, due to (11c), we have $\nabla\Psi_i(\mathbf{p}^{(n+1)}, \mathbf{d}^{(n+1)}, \ell^{(n+1)}; \mathbf{p}^{(n)}, \mathbf{d}^{(n)}, \ell^{(n)}) \rightarrow \nabla\Psi_i(\mathbf{p}^*, \mathbf{d}^*, \ell^*; \mathbf{p}^*, \mathbf{d}^*, \ell^*) = \nabla\psi_i(\mathbf{p}^*, \mathbf{d}^*, \ell^*)$. The same arguments also apply to the convex constraints in (2). Thus if $\left(\mathbf{p}^*, \mathbf{d}^*, \ell^*\right)$ is a regular point of (3), then we can show that $\left(\mathbf{p}^*, \mathbf{d}^*, \ell^*\right)$ is also a KKT point of (3). We skip the details here for the sake of brevity.

B. Complexity of Algorithm 1

As shown in Section IV, Algorithm 1 converges in few iterations and in each iteration the SOCP in (10) needs to be solved. The worst case complexity of generic interior-point methods for an SOCP depends on the number of optimization variables, number of SOC constraints and the dimension of SOC constraints [24]. In our case, (10) has $(M^2 + 2M)$ optimization variables, $O(M^2)$ SOC constraints and the total dimension of SOC constraints[4] is $O(M^2)$. Using the complexity estimate of [24], the accuracy of the duality gap can be improved by a constant factor in $O((M^2)^{1/2})$ iterations and the time complexity of each iteration is $O(M^4 \times M^2)$. This leads to an overall complexity of $O(M^7)$. This complexity bound is obtained by simply viewing each linear inequality constraint of the form $\mathbf{c}^{\mathsf{T}}\mathbf{x} + d \geq 0$ as a special case of the SOC constraint $\mathbf{c}^{\mathsf{T}}\mathbf{x} + d \geq \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2$, where $\mathbf{A} = 0$ and $\mathbf{b} = 0$. Consequently, the problem data of the resulting SOCP consists of many zeros. Modern off-the-shelf conic solvers can efficiently exploit this sparsity, and thus the practical run time for solving (10) is much less than the worst-case time complexity. A further improvement in computational complexity can be obtained by using first-order optimization methods to solve the given problem in future study.

C. A Corresponding OMA-MEC System

In order to showcase the advantages of the NOMA-based system, in this subsection, we present a corresponding baseline OMA-based system. Following the two-user (pure) OMA system in [5], we consider a multi-user OMA-MEC system, where a dedicated time slot

is allocated to every user. Note that we reuse the notations from the NOMA-MEC system in this subsection when appropriate. Specifically, U_m offloads its data to the MEC server *only* during $\hat{D}_m$ such that $\sum_{j=1}^{m} \hat{D}_j \leq D_m$. The data offloaded (in nats) by U_m to the server is given by $\hat{N}_{\text{off},m} = B\hat{D}_m \ln(1 + g_m\hat{P}_m)$, where $\hat{P}_m$ is the power transmitted by U_m during $\hat{D}_m$. The problem of maximizing the minimum number of offloaded nats among all users in the OMA-MEC can be easily formulated but is skipped here due to the space constraint. Similar to the case of NOMA, we can derive a SCA-based method to solve the resulting problem. Explicitly in iteration $n + 1$, an SOCP can be formulated as follows:

$$\underset{\hat{\mathbf{p}}\geq 0, \hat{\mathbf{d}}\geq 0}{\text{maximize}} \; \varphi, \quad (12a)$$

$$\text{subject to } B\hat{f}_m\left(\hat{P}_m, \hat{D}_m; \hat{P}_m^{(n)}, \hat{D}_m^{(n)}\right) \geq \varphi, \quad (12b)$$

$$\sum_{m=1}^{M} \hat{q}_m(\hat{P}_m, \hat{D}_m; \hat{P}_m^{(n)}, \hat{D}_m^{(n)}) \leq E_{\text{th}}, m \in \mathcal{M}, \quad (12c)$$

$$\sum_{j=1}^{m} \hat{D}_j \leq D_m, m \in \mathcal{M}, \quad (12d)$$

$$\hat{P}_m \leq P_m, m \in \mathcal{M}, \quad (12e)$$

where $\hat{\mathbf{p}} \triangleq [\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_M]^{\mathsf{T}}$, $\hat{\mathbf{d}} \triangleq [\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_M]^{\mathsf{T}}$, and $B\hat{f}_m(\hat{P}_m, \hat{D}_m; \hat{P}_m^{(n)}, \hat{D}_m^{(n)})$ is a concave lower bound on $\hat{N}_{\text{off},m}$, $\hat{f}_m(\hat{P}_m, \hat{D}_m; \hat{P}_m^{(n)}, \hat{D}_m^{(n)}) \triangleq \hat{a}_m + \hat{b}_m\hat{D}_m - \hat{c}_m\hat{P}_m - \hat{d}_m(\hat{D}_m+1)^2 /(1 + g_m\hat{P}_m)$, $\hat{a}_m \triangleq 0.25[1 - (\hat{D}_m^{(n)})^2 + (\hat{D}_m^{(n)} - 1)^2 g_m\hat{P}_m^{(n)}/(1 + g_m\hat{P}_m^{(n)})]$, $\hat{b}_m \triangleq [1 + \ln(1 + g_m\hat{P}_m^{(n)}) + 0.5(\hat{D}_m^{(n)}-1)]$, $\hat{c}_m \triangleq 0.25g_m(\hat{D}_m^{(n)} - 1)^2/(1 + g_m\hat{P}_m^{(n)})$ and $\hat{d}_m \triangleq 0.25(1 + g_m\hat{P}_m^{(n)})$. Similarly, we have $\hat{q}_m(\hat{P}_m, \hat{D}_m; \hat{P}_m^{(n)}, \hat{D}_m^{(n)}) \triangleq 0.25(\hat{D}_m + \hat{P}_m)^2 + 0.25(\hat{D}_m^{(n)} - \hat{P}_m^{(n)})^2 - 0.5(\hat{D}_m^{(n)} - \hat{P}_m^{(n)})(\hat{D}_m - \hat{P}_m) \geq \hat{D}_m\hat{P}_m$.

Although a dedicated interference free time slot is allocated to each user in OMA system to offload data to MEC server, yet a combination of power controlled data offloading of users to previous time slots and low interference at delay sensitive users due to SIC make NOMA-MEC perform better for MEC purposes. In the section to follow, we perform various numerical experiments that corroborate this observation.

IV. NUMERICAL RESULTS

In this section we conduct numerical experiments to assess the performance of our algorithm to offload computational tasks to the MEC server. Here, g_m is modeled as $|h_m|^2/\zeta(\mathfrak{d})$ where $h_m \sim \mathcal{CN}(0, 1)$ denotes Rayleigh fading, and $[\zeta(\mathfrak{d})]_{\text{dB}} \triangleq 128.1 + 37.6\log_{10}(\mathfrak{d})$ is the distance-dependent path loss with $\mathfrak{d}$ denoting the distance between the transmitter and receiver in km [26]. We consider $B = 20$ MHz, which results in the noise power of -92 dBm. Note also that the channel gains are normalized w.r.t. the noise power at the MEC server. The distance of the users from the BS are given by $\{200, 250, 300, \ldots, 200 + 50(M - 1)\}$ m. Similarly, the time delay deadlines for the users are given by $\{D_1, D_1 + \Delta, D_1 + 2\Delta, \ldots, D_1 + \Delta(M - 1)\}$, where $D_1 = 0.5$ s and $\Delta = 0.05$ s, unless stated otherwise. Note that, for simplicity, we have considered a constant difference between the users' deadlines, however, the proposed SCA-based algorithm holds good also for the case when the difference between the deadlines are random. Also, throughout this section, we assume that all of the users have the same transmit power budget, P_{t}, i.e., $P_m = P_{\text{t}}, \forall m \in \mathcal{M}$. We execute the proposed algorithm using CVXPY with MOSEK as the internal solver on a Linux PC with 7.5 GiB memory and Intel Core i5-7200 U CPU.
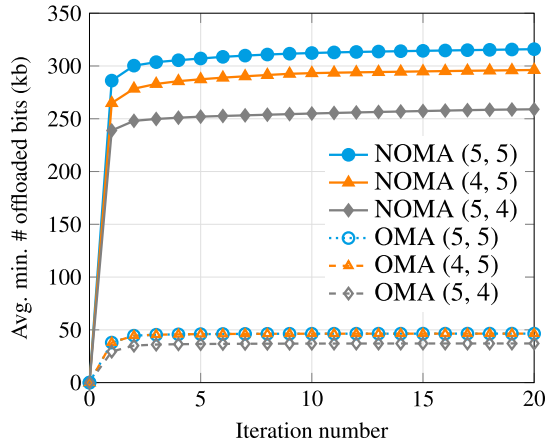
---

[3]Following [23], a quadratic proximal term can be added to the non-strict convex objective in (10), without effecting the optimal point, so as to ensure the existence of such limit of the sequence.

[4]Big-O notation has been used in literature to represent the number of variables, SOC constraints and their dimension [25].

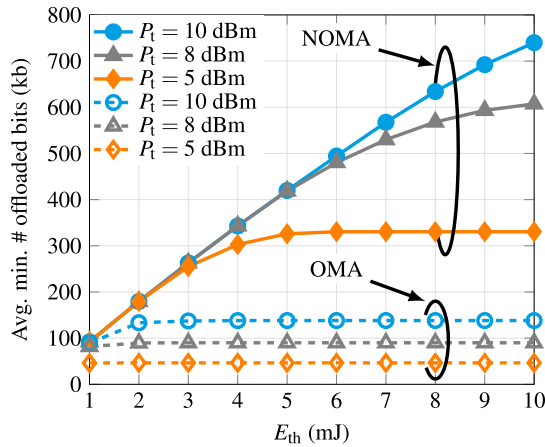Fig. 2.　Convergence of the NOMA-MEC and OMA-MEC systems.

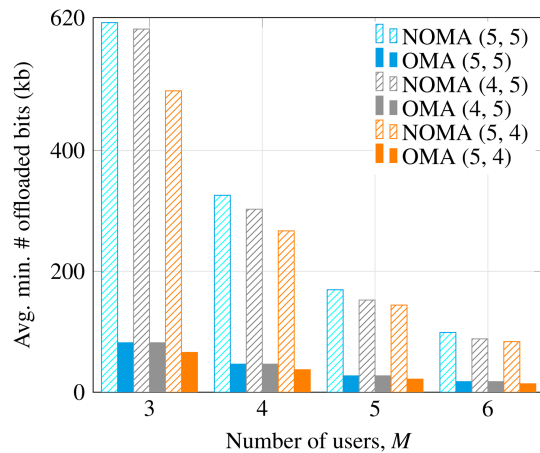Fig. 3.　Average minimum number of offloaded bits versus E_{th}.

Fig. 4.　Average minimum number of offloaded bits versus M.

Note that in this section, we use the logarithm with base 2 and thus the offloaded tasks are expressed in bits. For all of the figures in this section, the average values of offloaded bits are computed over the same set of 100 independent channel realizations. Moreover, in Figs. 2, 3 and 5, we consider $M = 4$.

In Fig. 2, where the numbers in the parentheses denote (E_{th} in mJ, P_{t} in dBm), we show the convergence behavior of the proposed algorithm for the NOMA-MEC system. We also show the convergence of the corresponding OMA-MEC system for comparison. It can be noted from
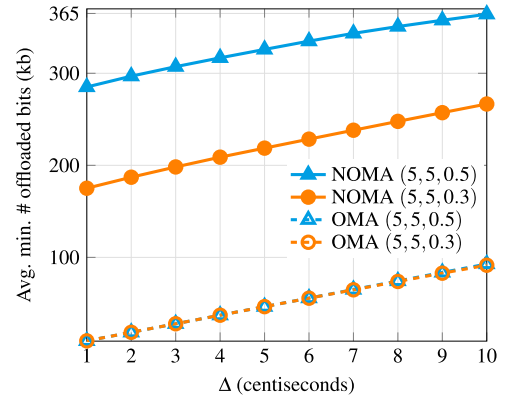

Fig. 5.　Average minimum number of offloaded bits versus Δ.

TABLE I
AVERAGE PROBLEM SOLVING TIME FOR NOMA-MEC SYSTEM
USING ALGORITHM 1

# of users, M	4	8	12	16	20
Time (in seconds)	0.63	1.74	4.85	12.27	137.63

the figure that both NOMA and OMA systems converge within a small number of iterations. Moreover, for the given system parameters, the NOMA-based system results in a much larger number of offloaded bits within the same number of iterations, compared to the OMA-based counterpart. Also, in Table I, we show the average problem solving time for the NOMA-MEC system for $E_{\text{th}} = 5$ mJ, $P_{\text{t}} = 5$ dBm and different number of users with number of iterations fixed at 20.

In Fig. 3 we report the average minimum number of offloaded bits as a function of energy threshold, E_{th} for given P_{t}. From the figure, the superiority of the NOMA-based system is clearly evident over its OMA-based counterpart. More interestingly, for the case of NOMA-based system, the objective first increases for the small values of E_{th} (which we refer to as the *energy-constrained* regime), and then saturates for large values of E_{th} (which we refer to as the *power-constrained* regime). It is noteworthy that for two different NOMA systems with the same E_{th} but different P_{t}, the performance in the energy-constrained regime remains the same. It can also be noted from the figure that compared to NOMA, the effect of easing the energy constraint is less pronounced in the OMA-MEC system.

To investigate the impact of number of users, in Fig. 4, we plot the variation in the average minimum number of offloaded bits against the number of users. Note that in the figure legend, the numbers in the parentheses denote (E_{th} in mJ, P_{t} in dBm). It can be inferred from the figure that as the number of users, M, increases, the performance of both NOMA-based and OMA-based systems deteriorate. It is also evident that with an increase in the number of users, the performance gap between the NOMA-MEC and OMA-MEC systems also decreases. Note that for the NOMA-MEC system, as M increases, the inter-user interference for the higher-indexed users also increases, reducing the achievable rate for such users.

To investigate the impact of users' delay requirements, in Fig. 5 we investigate the variation of number of offloaded bits with Δ. Note that in the legend of the figure, the numbers in the parentheses denote (E_{th} in mJ, P_{t} in dBm, D_1 in s). Remarkably, the performance of both NOMA-MEC and OMA-MEC systems increases linearly with a linear increase in Δ. The poor performance of OMA is due to the fact that U_m has to offload all its data only during the mth time slot. Under the same circumstances NOMA is able to offload more data due to availability of additional time slot. Thus, NOMA-based MEC systems

can be considered as a suitable candidate to achieve gains associated with edge servers under stringent delay constraints of users.

V. CONCLUSION

In this correspondence, we have considered using NOMA to offload computational tasks for edge computing in a multi-user scenario. For the system model under consideration, we have formulated a non-convex optimization problem to maximize the minimum number of offloaded bits among the users, constrained by delay, transmit power and energy requirements. The original non-convex problem is solved using successive convex optimization strategy that leads to a provably convergent algorithm, each iteration of which can be solved efficiently by the SOCP solvers. The numerical results have showed that in NOMA systems with given power budget, the minimum number of offloaded bits first increases with energy budget in the energy-constrained regime, and then saturates in the power-constrained regime. It has also been confirmed that as the number of users increases, the performance gap between NOMA and OMA systems decreases. Our results have shown that *for strict delay constraints*, NOMA substantially outperforms OMA-based approach in terms of maximizing tasks to be offloaded to the MEC server.

APPENDIX A
PROOF OF THEOREM 1

We can write $\ln(1 + \frac{u}{v}) = \ln(u + v) - \ln(v)$. Then, due to the concavity of $\ln(v)$ we have $\ln(v) \leq \ln(v_0) + \frac{1}{v_0}(v - v_0) = -1 + \ln(v_0) + \frac{v}{v_0}$. Next, we recall the following well-known inequality: $\ln(u) \geq 1 - \frac{1}{u}$, for $u > 0$, where the equality occurs if and only if $u = 1$. Therefore, it follows that $\ln(\frac{u+v}{u_0+v_0}) \geq 1 - \frac{u_0+v_0}{u+v} \Rightarrow \ln(u + v) \geq 1 + \ln(u_0 + v_0) - \frac{u_0+v_0}{u+v}$. Using the preceding inequalities for $\ln(v)$ and $\ln(\frac{u+v}{u_0+v_0})$ yields $\ln(1 + \frac{u}{v}) \geq 2 + \ln(u_0 + v_0) - \ln(v_0) - \frac{u_0+v_0}{u+v} - \frac{v}{v_0}$, and thus $x \ln(1 + \frac{u}{v}) \geq x(2 + \ln(1 + \frac{u_0}{v_0})) - \frac{(u_0+v_0)x}{u+v} - \frac{xv}{v_0} = x(2 + \ln(1 + \frac{u_0}{v_0})) - (u_0 + v_0)\frac{(x+1)^2-(x-1)^2}{4(u+v)} - \frac{(x+v)^2-(x-v)^2}{4v_0}$.

Since the term $\frac{(x-1)^2}{4(u+v)}$ is jointly convex in all variables involved, approximating the term $\frac{(x-1)^2}{4(u+v)}$ around x_0, u_0, and v_0 yields $\frac{(x-1)^2}{4(u+v)} \geq \frac{(x_0-1)^2}{4(u_0+v_0)} + \frac{(x_0-1)}{2(u_0+v_0)}(x - x_0) - \frac{(x_0-1)^2}{4(u_0+v_0)^2}(u + v - u_0 - v_0) = \frac{1-x_0}{2(u_0+v_0)} + \frac{x(x_0-1)}{2(u_0+v_0)} - \frac{(x_0-1)^2(u+v)}{4(u_0+v_0)}$. Similarly we have $(x - v)^2 \geq -(x_0 - v_0)^2 + 2(x_0 - v_0)(x - v)$.

Combining the preceding expressions for $x \ln(1 + \frac{u}{v})$, $\frac{(x-1)^2}{4(u+v)}$ and $(x - v)^2$ produces (4), which completes the proof.

REFERENCES

[1] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

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

[3] Y. Wen, X. Zhou, F. Fang, H. Zhang, and D. Yuan, "Joint time and power allocation for cooperative NOMA based MEC system," in *Proc. IEEE 92nd Veh. Technol. Conf.*, 2020, pp. 1–5.

[4] K. Wang, F. Fang, D. B. da Costa, and Z. Ding, "Sub-channel scheduling, task assignment, and power allocation for OMA-based and NOMA-based MEC systems," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2692–2708, Apr. 2021.

[5] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1875–1879, Dec. 2018.

[6] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal resource allocation for delay minimization in NOMA-MEC networks," *IEEE Trans. Commun*, vol. 68, no. 12, pp. 7867–7881, Dec. 2020.

[7] M. Zeng, N.-P. Nguyen, O. A. Dobre, and H. V. Poor, "Delay minimization for NOMA-assisted MEC under power and energy constraints," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1657–1661, Dec. 2019.

[8] M. Sheng, Y. Dai, J. Liu, N. Cheng, X. Shen, and Q. Yang, "Delay-aware computation offloading in NOMA MEC under differentiated uploading delay," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2813–2826, Apr. 2020.

[9] Y. Wu, L. P. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE J. Sel. Topics Signal Process*, vol. 13, no. 3, pp. 392–407, Jun. 2019.

[10] K. Wang, Z. Ding, D. K. C. So, and G. K. Karagiannidis, "Stackelberg game of energy consumption and latency in MEC systems with NOMA," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2191–2206, Apr. 2021.

[11] L. Li et al., "Resource allocation for NOMA-MEC systems in ultra-dense networks: A learning aided mean-field game approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1487–1500, Mar. 2021.

[12] V. D. Tuong, T. P. Truong, T.-V. Nguyen, W. Noh, and S. Cho, "Partial computation offloading in NOMA-assisted mobile-edge computing systems using deep reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13196–13208, Sep. 2021.

[13] L. Qian, Y. Wu, F. Jiang, N. Yu, W. Lu, and B. Lin, "NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5688–5698, Aug. 2021.

[14] X. Huang, S. Zeng, D. Li, P. Zhang, S. Yan, and X. Wang, "Fair computation efficiency scheduling in NOMA-aided mobile edge computing," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1812–1816, Nov. 2020.

[15] X. Yu, F. Xu, J. Cai, X. Dang, and K. Wang, "Computation efficiency optimization for millimeter-wave mobile edge computing networks with NOMA," *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 4578–4593, Aug. 2023.

[16] W. Wu, X. Wang, F. Zhou, K. K. Wong, C. Li, and B. Wang, "Resource allocation for enhancing offloading security in NOMA-enabled MEC networks," *IEEE Syst. J.*, vol. 15, no. 3, pp. 3789–3792, Sep. 2021.

[17] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access NOMA with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, 2015.

[18] J. Kim, J. Kim, and S.-H. Park, "Joint design of power control and SIC decoding order for max-min fairness optimization in uplink NOMA systems," in *Proc. IEEE Int. Conf. Inf. Netw.*, 2021, pp. 339–342.

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

[20] M. Zeng, N.-P. Nguyen, O. A. Dobre, Z. Ding, and H. V. Poor, "Spectral and energy efficient resource allocation for multi-carrier uplink NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9293–9296, Sep. 2019.

[21] Z. Wei, D. W. K. Ng, and J. Yuan, "Joint pilot and payload power control for uplink MIMO-NOMA with MRC-SIC receivers," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 692–695, Apr. 2018.

[22] L.-N. Tran, M. F. Hanif, A. Tolli, and M. Juntti, "Fast converging algorithm for weighted sum rate maximization in multicell MISO downlink," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 872–875, Dec. 2012.

[23] I. Sugathapala, M. F. Hanif, B. Lorenzo, S. Glisic, M. Juntti, and L.-N. Tran, "Topology adaptive sum rate maximization in the downlink of dynamic wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3501–3516, Aug. 2018.

[24] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, no. 1/3, pp. 193–228, 1998.

[25] I. I. W. Tsang and J. T.-Y. Kwok, "Efficient hyperkernel learning using second-order cone programming," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 48–58, Jan. 2006.

[26] 3rd Generation Partnership Project (3GPP), "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) requirements for LTE pico node B.(Release 13)," 3GPP, Sophia Antipolis, France, Tech. Rep. TR 36.931 v13.0.0, 2016.