IRS-based Energy Efficiency and Admission Control Maximization for IoT Users with Short Packet Lengths

Jalal Jalali *Member, IEEE*, Ata Khalili *Member, IEEE*, Atefeh Rezaei *Student Member, IEEE*, Rafael Berkvens *Member, IEEE*, Maarten Weyn *Member, IEEE*, and Jeroen Famaey *Senior Member, IEEE*.

***Abstract*—In this paper, we study a multiuser multiple-input single-output (MISO) machine type communication (MTC)-enabled intelligent reflecting surface (IRS) system, where a multi-antenna access point (AP) transmits information symbols to a set of Internet of Things (IoT) users with short packet transmission. In particular, the total energy efficiency (EE) and the number of IoT users that could be served fairly are maximized by jointly optimizing active and passive beamformers. An efficient algorithm based on alternating optimization (AO) is proposed to solve the main optimization problem iteratively. To this end, we adopt the difference of convex functions (DC) and successive convex approximation (SCA) to make a concave-convex function. Then, we employ the fractional programming based on the quadratic form to obtain a sub-optimal solution for the active beamformers at the AP and the number of admitted users. In the passive beamforming case, a penalty-based approach is utilized together with the SCA technique to handle the unit-modulus constraints at the IRS. Simulation results unveil an interesting tradeoff between EE and user admissibility performance. Besides, the results show the effectiveness of the IRS deployment in improving EE and successfully admitted users.**

***Index Terms*—Admission control, alternating optimization (AO), energy efficiency (EE), intelligent reflecting surface (IRS), machine type communication (MTC), short packet transmission.**

I. Introduction

Intelligent reflective surfaces (IRS)-aided wireless communication has attracted significant research attention due to its simplistic deployment and favorable wireless propagation environment [1]. The IRS typically operates in the full-duplex mode and has a promising performance in the aspect of spectral efficiency (SE) and energy efficiency (EE) for beyond-fifth-generation (B5G) wireless communication systems. In particular, the maximization of the SE and the weighted sum-rate were studied in [2] and [3], respectively, where the active and passive beamformers were optimized at the base station (BS) and passive beamformers at the IRSs. To provide a more energy-efficient IRS system, the performance of simultaneous wireless information and power transfer (SWIPT) technology in terms of an EE maximization problem was investigated in [4] by jointly optimizing the phase shifts at the IRS and active beamformers at the BS, and power splitting ratio in each user.

Machine-type communication (MTC) is one of the services for future wireless communication that are mainly classified into massive MTC (mMTC) and ultra-reliable MTC (uMTC). MTC encompasses a variety of emerging concepts such as the Internet of Things (IoT), Internet of Vehicles (IoV), Internet of Everything (IoE), and so on. The mMTC service helps make future networks more scalable with efficient connectivity for a massive number of devices that send shorter packets [5]. However, the conventional Shannon capacity formula cannot be affirmed for these services under the short packet regime [6]. In this regard, many works study resource allocation for MTC networks with delay-tolerant devices such as ultra-reliable low-latency (URLLC)-type terminals. For instance, a global optimal resource allocation for a URLLC system was obtained while optimizing the bandwidth, power allocation, and antenna arrangement parameters to minimize the weighted sum of downlink (DL) and uplink (UL) average power consumption in [7]. Moreover, the authors in [8] maximized the sum throughput of a multiple-input single-output (MISO) orthogonal frequency division multiple access (OFDMA) system by designing the active beamforming vectors at the BS. The precoder design of a BS was carried out in [9] to maximize EE in a multi-user MIMO network with finite blocklength codes. The optimal design of the energy-efficient multiple-input multiple-output (MIMO) aided UL URLLC grant-free access systems was elucidated in [10]. [11] considered a hybrid puncturing and superposition policy that jointly maximizes the minimum average throughput of enhanced mobile broadband (eMBB) users' traffic and the number of admitted URLLC users. An IRS platform could be introduced into delay-insensitive systems to further overcome the computational latency. In this sense, an IRS-aided mobile edge computing system was studied in [12], where the latency was minimized by jointly optimizing the edge computing resources, computation offloading, as well as beamforming matrices at the BS and the IRS, respectively. In [13], the authors studied IRS-aided MTC in a factory automation scenario and derived the average data rate and decoding error probability under short packet transmission.

In order to enable MTC service, it is vital to increase the number of successfully admitted MTC/IoT users with higher reliability. IRS could help to maximize the admitted number of such users. To the best of our knowledge, the gain of deploying the IRS platform in an MTC-enabled system with short packet length to maximize the EE simultaneously with the number of admitted IoT users has not been studied in the literature yet [8]–[14]. Besides, it would be interesting to understand if the QoS is met in the face of short packet transmissions in an IRS-aided IoT system. In this paper, we aim to address the

Jalal Jalali, Rafael Berkvens, Maarten Weyn, and Jeroen Famaey are with IDLab research group, University of Antwerp - imec, 2000 Antwerp, Belgium (e-mail: {jalal.jalali,rafael.berkvens,maarten.weyn,jeroen.famaey}@imec.be).

Ata Khalili is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg, Erlangen 91054, Germany. (e-mail: ata.khalili@ieee.org).

This work was supported by the CHIST-ERA grant SAMBAS (CHIST-ERA-20-SICT-003), with funding from FWO, ANR, NKFIH, and UKRI.

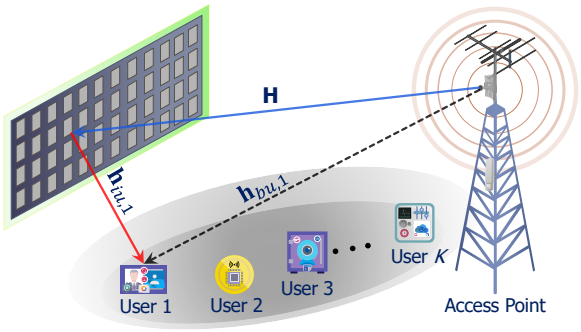

Fig. 1. IRS-Assisted IoT users with finite blocklength.

above issues. To this end, we consider a resource allocation algorithm design for a DL MISO MTC-enabled IRS system. In our study model, a multi-antenna access point (AP) serves multiple single-antenna IoT users with a short packet length using a smart reconfigurable reflector. Moreover, we focus on maximizing the total EE together with admission control in the proposed system, providing valuable insights into the system design. Consequently, the main contributions of this paper can be summarized as follows:

• We maximize the system's total EE together with admission control by jointly optimizing active and passive beamformers at the AP and IRS, respectively, subject to the minimum required data rate for each admitted IoT user with a short packet and unit-modulus constraints at the IRS.

• This problem is formulated as a multi-objective optimization problem (MOOP) which is a non-convex mixed integer non-linear programming (MINLP) problem, and it is non-deterministic polynomial-time (NP) hard. To tackle this issue, we first convert it into a single objective optimization problem via a weighting coefficient. Then, we exploit an alternating optimization (AO) resource allocation algorithm to solve the formulated optimization problem iteratively, which improves the objective function in each step.

• The simulation results reveal that deploying an IRS can increase the system's EE and admission control of the IoT users with a short packet length. Results also reveal an interesting tradeoff region between EE and user admissibility.

II. System Model

We consider a set of IoT users with a finite blocklength-enabled IRS system, shown in Fig. 1, with an N-element IRS, M-antenna AP, and K single-antenna users. The set of IRS elements, AP antennas, and the users are denoted by $\mathcal{N} = \{1, ..., N\}$, $\mathcal{M} = \{1, ..., M\}$, and $\mathcal{K} = \{1, ..., K\}$, respectively. We further assume that B_k information bits are assigned to user k, where the AP encodes these information bits into a block code with the length of m_d (symbols), which is expressed as $z_k^{[l]}$, $l \in \mathcal{L} = \{1, 2, ..., m_d\}$. Subsequently, the transmit signal at the AP can be expressed as $\mathbf{s}^{[l]} = \sum_{k\in\mathcal{K}} u_k \mathbf{w}_k z_k{}^{[l]}$, where $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$ represents the beamforming vector for user k. We can drop superscript l, considering a quasi-static flat-fading channel model, where the wireless channels remain unchanged within each transmission block. Besides, $u_k = 1$ denotes that the k-th IoT user is served in the system, while in the case of $u_k = 0$ the k-th user is dropped. We denote the baseband equivalent channel responses[1] from the AP-to-IRS, IRS-to-user k, and AP-to-user k as $\mathbf{H} \in \mathbb{C}^{N\times M}$, $\mathbf{h}_{iu,k} \in \mathbb{C}^{N\times 1}$, and $\mathbf{h}_{bu,k} \in \mathbb{C}^{M\times 1}$, respectively. Also, we define $\mathbf{\Theta} = \text{diag}(\beta_1 e^{j\alpha_1}, \beta_2 e^{j\alpha_2}, ..., \beta_N e^{j\alpha_N})$ as the reflection-coefficients matrix at the IRS, where $\beta_n \in [0, 1]$ and $\alpha_n \in (0, 2\pi]$, $\forall n \in \mathcal{N}$ are the reflection amplitude and phase shift of the n-th reflection coefficient at the IRS[2], respectively. By defining $\mathbf{h}_k^H \triangleq \mathbf{h}_{iu,k}^H \mathbf{\Theta}\mathbf{H} + \mathbf{h}_{bu,k}^H$, $\forall k \in \mathcal{K}$, as the equivalent channel link, the received signal at user k can be written as:

$$y_k = \mathbf{h}_k^H \mathbf{s} + n_k \triangleq \sum_{k\in\mathcal{K}} u_k \mathbf{h}_k^H \mathbf{w}_k z_k + n_k, \forall k \in \mathcal{K} \quad (1)$$

where the noise is modeled as an additive white Gaussian noise (AWGN) random variable with zero mean and variance σ_k^2, denoted by a circularly symmetric Gaussian distribution referred to as $n_k \sim \mathcal{CN}(0, \sigma_k^2)$. Then, the SINR at user k can be expressed as:

$$\gamma_k = \frac{u_k \left|\mathbf{h}_k^H \mathbf{w}_k\right|^2}{\sum\limits_{i\neq k, i\in\mathcal{K}} u_i \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma_k^2}, \forall k \in \mathcal{K}. \quad (2)$$

Finite and short blocklength is required to guarantee low-latency and high-reliability wireless communication for MTC-type IoT terminals. The precise approximation for the achievable data rate of each user is given by [6]:

$$R_k(u_k, \mathbf{w}_k, \mathbf{\Theta}) = F(u_k, \mathbf{w}_k, \mathbf{\Theta}) - G(u_k, \mathbf{w}_k, \mathbf{\Theta}), \ \forall k \in \mathcal{K}, \quad (3)$$

where

$$F_k(u_k, \mathbf{w}_k, \mathbf{\Theta}) = \log_2(1 + \gamma_k), \qquad \forall k \in \mathcal{K}, \quad (4)$$

$$G_k(u_k, \mathbf{w}_k, \mathbf{\Theta}) = Q^{-1}(\epsilon_k)\sqrt{\frac{1}{m_d} V_k}, \ \forall k \in \mathcal{K}. \quad (5)$$

In addition, ϵ_k is the decoding error, m_d indicates the blocklength, and V_k denotes the channel dispersion which is given by $V_k = a^2\left(1 - (1 + \gamma_k)^{-2}\right)$, where $a = \log_2(e)$. To ensure users' QoS regarding the received number of bits, the reliability, and the latency, a minimum threshold data rate denoted by R_{th}^k which should be provided for each user, is defined as follows:

$$R_k(u_k, \mathbf{w}_k, \mathbf{\Theta}) \geq R_{\text{th}}^k, \ \forall k \in \mathcal{K}. \quad (6)$$

We now define the EE as the ratio of the total system data rate to the corresponding network power consumption in [bits/Joule]:

$$\mathscr{E}_{eff}(u_k, \mathbf{w}_k, \mathbf{\Theta}) = \frac{\sum\limits_{k\in\mathcal{K}} R_k(u_k, \mathbf{w}_k, \mathbf{\Theta})}{\sum\limits_{k\in\mathcal{K}} u_k \left\|\mathbf{w}_k\right\|^2 + P_s + N P_d + P_c^{\text{AP}}}, \quad (7)$$

where P_s indicates the static power consumption as required to maintain the basic circuit operations of the IRS, P_d is the dynamic power dissipation per reflecting component, and P_c^{AP} is the circuit power at the AP.

In what follows, we first formulate the problem to maximize the EE together with admission control while considering the minimum data rate requirement for IoT users with a short packet length. Then, we propose a solution to solve the optimization problem.

[1]We consider the case where the CSI and delay requirements are known at the AP, which offers us insights into the theoretical upper-performance bounds of a system with imperfect CSI as well (see [2]–[4], [8], [14]).

[2]We consider continuous phase shifts, as discrete shifts cause misalignment of IRS-reflected and non-IRS-reflected signals, which degrades performance [1].

III. Problem Formulation

In this section, we aim to maximize the total EE of the considered system together with the number of admitted IoT users by jointly optimizing the active beamformers at the AP and the phase shifts at the IRS. Accordingly, this problem as a MOOP can be mathematically formulated as follows:

$$\mathrm{P}_1: \max_{\mathbf{u},\mathbf{w}_k,\boldsymbol{\Theta}} \quad \mathscr{E}_{eff}(u_k,\mathbf{w}_k,\boldsymbol{\Theta}) \tag{8}$$

$$\max_{\mathbf{u},\mathbf{w}_k,\boldsymbol{\Theta}} \quad \sum_{k\in\mathcal{K}} u_k$$

$$s.t.: R_k(u_k,\mathbf{w}_k,\boldsymbol{\Theta}) \geq u_k R_{\text{th}}^k, \ \forall k\in\mathcal{K}, \tag{8a}$$

$$|\boldsymbol{\Theta}_{nn}| = 1, \qquad \forall n\in\mathcal{N}, \tag{8b}$$

$$\sum_{k\in\mathcal{K}} u_k \|\mathbf{w}_k\|^2 \leq p_{\max}, \tag{8c}$$

$$u_k \in \{0,1\}, \qquad \forall k\in\mathcal{K}. \tag{8d}$$

where (8a) ensures the reliability of each admitted MTC-type finite-blocklength user. (8b) guarantees N unit-modulus elements in the diagonal phase shift matrix. Moreover, (8c) describes the transmission power budget limitation in which $p_{\max}$ is the maximum allowable transmission power. (8d) indicates that u_k is a binary variable, where $\mathbf{u}=[u_1,...,u_K]$ is the optimization decision vector. The optimization problem P_1[3] is a non-convex MINLP due to the non-convexity of the objective function and the constraints, as well as incorporating binary variables in the objective and constraints. In general, finding an optimal solution for such a problem is impossible. However, in the next section, we adopt an approach to find an efficient sub-optimal solution.

IV. Proposed Solution

P_1 is a non-convex optimization problem due to the highly coupled optimization variables. In general, there is no well-organized method to solve P_1. However, we propose an alternating optimization (AO) with low computational complexity to find a sub-optimal solution. In particular, we first convert the MOOP problem into the single objective optimization problem (SOOP) via weighting coefficients that reflect the required preferences of the main MOOP. Then, we decompose the original problem into two sub-problems. The Big-M approach, semidefinite programming (SDP), and fractional programming based on quadratic transform are applied to optimize the active beamformers and the admitted users in the first sub-problem. While in the second one, the phase shifts are optimized by exploiting the penalty approach and SCA technique.

A. Optimizing $\mathbf{w}_k$ with Fixed $\boldsymbol{\Theta}$

At this stage, we assume that the passive reflecting elements at the IRS, i.e., $\boldsymbol{\Theta}$ are fixed to design the active beamformers, $\mathbf{w}_k$ at the AP and determine the admission control for IoT users. By adopting SDP, we have $\mathbf{W}_k = \mathbf{w}_k\mathbf{w}_k^H$ and $\mathbf{H}_k = \mathbf{h}_k\mathbf{h}_k^H, \forall k\in\mathcal{K}$. Besides, we define the product of two variables u_k and $\mathbf{W}_k$ as a new auxiliary variable $\tilde{\mathbf{W}}_k$ based on the big-M method [14], we impose the following additional constraints:

$$0 \preceq \tilde{\mathbf{W}}_k \preceq p_{\max}\mathbf{I}_M u_k, \quad \forall k\in\mathcal{K}, \tag{9a}$$

$$\mathbf{W}_k - (1-u_k)p_{\max}\mathbf{I}_M \preceq \tilde{\mathbf{W}}_k \preceq \mathbf{W}_k, \forall k\in\mathcal{K}. \tag{9b}$$

[3]Please note P_1 ensures user fairness for a subset of the users, i.e., the total number of admitted users.

Next, we relax the binary variable u_k to a continuous one by employing the following constraints:

$$\sum_{k\in\mathcal{K}} u_k - \sum_{k\in\mathcal{K}} (u_k)^2 \leq 0, \ 0\leq u_k\leq 1, \forall k\in\mathcal{K}. \tag{10}$$

Also note that, γ_k in $R_k(\tilde{\mathbf{W}}_k) = F_k(\tilde{\mathbf{W}}_k) - G_k(\tilde{\mathbf{W}}_k)$ can be expressed as:

$$\gamma_k = \frac{\text{Tr}(\mathbf{H}_k\tilde{\mathbf{W}}_k)}{\sum\limits_{i\in\mathcal{K},i\neq k}\text{Tr}(\mathbf{H}_k\tilde{\mathbf{W}}_i)+\sigma_k^2}, \forall k\in\mathcal{K}. \tag{11}$$

It is notable that constraint (8a) in P_1 is not concave. To deal with the non-concavity, we propose a set of auxiliary variables $\xi_k, \ \forall k\in\mathcal{K}$. This provides a lower bound of the SINR. Accordingly, we can write the SINR in (11) as below:

$$0\leq\xi_k\leq\gamma_k = \frac{f_k(\tilde{\mathbf{W}}_k)}{g_k(\tilde{\mathbf{W}}_k)}, \forall k\in\mathcal{K}, \tag{12}$$

where the nominator and denominator of (12) can be expressed as:

$$f_k(\tilde{\mathbf{W}}_k) = \text{Tr}(\mathbf{H}_k\tilde{\mathbf{W}}_k), \tag{13}$$

$$g_k(\tilde{\mathbf{W}}_k) = \sum_{i\in\mathcal{K},i\neq k}\text{Tr}(\mathbf{H}_k\tilde{\mathbf{W}}_i)+\sigma_k^2, \tag{14}$$

respectively. By exploiting the lower bound in (12), SDP, big-M, and by introducing the weighting coefficient $0<\alpha<1$ that indicates the importance of the different objectives, the main optimization problem in the first stage can be restated as:

$$\mathrm{P}_2: \max_{u_k,\tilde{\mathbf{W}}_k,\mathbf{W}_k,\xi_k} \alpha\mathscr{E}_{eff}(\tilde{\mathbf{W}}_k,u_k) + (1-\alpha)\sum_{k\in\mathcal{K}} u_k \tag{15}$$

$$s.t.: 0\leq\xi_k\leq\frac{f_k(\tilde{\mathbf{W}}_k)}{g_k(\tilde{\mathbf{W}}_k)}, \quad \forall k\in\mathcal{K}, \tag{15a}$$

$$R_k(\xi_k)\geq R_{\text{th}}^k, \qquad \forall k\in\mathcal{K}, \tag{15b}$$

$$\text{Rank}(\mathbf{W}_k)\leq 1, \qquad \forall k\in\mathcal{K}, \tag{15c}$$

$$\sum_{k\in\mathcal{K}}\text{Tr}(\tilde{\mathbf{W}}_k)\leq p_{\max}, \tag{15d}$$

$$(9a),(9b),(10),$$

where $R_k(\xi_k) = F_k(\xi_k) - G_k(\xi_k), \forall k\in\mathcal{K}$ in constraint (15b) are given by:

$$F_k(\xi_k) = \log(1+\xi_k), \forall k\in\mathcal{K}, \tag{16}$$

$$G_k(\xi_k) = Q^{-1}(\epsilon_k)\sqrt{\frac{a^2}{m_d}\big(1-(1+\xi_k)^{-2}\big)}. \tag{17}$$

P_2 is still a non-convex optimization problem. To overcome this, we first modify the optimization problem and represent it as the canonical form required for the DC forms. Consequently, we apply first-order Taylor expansion to get a convex approximation of the non-convex terms. In particular, constraint (15a) can be represented as

$$\xi_k g_k(\tilde{\mathbf{W}}_k)\leq f_k(\tilde{\mathbf{W}}_k)\Rightarrow\xi_k A_k(\tilde{\mathbf{W}}_k)\leq f_k(\tilde{\mathbf{W}}_k)-\xi_k\sigma_k^2, \forall k\in\mathcal{K}, \tag{18}$$

where $A_k(\tilde{\mathbf{W}}_k) = \sum_{i\in\mathcal{K},i\neq k}\text{Tr}(\mathbf{H}_k\tilde{\mathbf{W}}_i)$. Nevertheless, (18) is a non-convex constraint since it is the product of two optimization variables, i.e., $\tilde{\mathbf{W}}_i$ and ξ_k, $\forall i,k\in\mathcal{K}$. However, it can be decoupled by adopting the following form [14]:

$$\xi_k A_k(\tilde{\mathbf{W}}_k) = P_k(\xi_k,\tilde{\mathbf{W}}_k) - Q_k(\xi_k,\tilde{\mathbf{W}}_k), \tag{19}$$

where

$$P_k(\xi_k,\tilde{\mathbf{W}}_k) = \frac{1}{2}\left(\xi_k + A_k(\tilde{\mathbf{W}}_k)\right)^2, \forall k\in\mathcal{K}, \tag{20}$$

$$Q_k(\xi_k, \tilde{\mathbf{W}}_k) = \frac{1}{2}(\xi_k)^2 + \frac{1}{2}\left(A_k(\tilde{\mathbf{W}}_k)\right)^2, \forall k \in \mathcal{K}. \quad (21)$$

By denoting $\Omega_k = \{\xi_k, \mathbf{W}_k, \tilde{\mathbf{W}}_k, u_k\}$ as a set of optimization variables, we have $U_k(\Omega_k) = P_k(\Omega_k) - Q_k(\Omega_k)$. Thus, P_2 can be recast as follows:

$$\mathrm{P}_3 : \max_{\Omega_k} \ \alpha \frac{\sum_{k\in\mathcal{K}} R_k(\xi_k)}{\sum_{k\in\mathcal{K}} \mathrm{Tr}(\tilde{\mathbf{W}}_k) + P_\mathrm{s} + NP_d + P_\mathrm{c}^{\mathrm{AP}}} + (1-\alpha)\sum_{k\in\mathcal{K}} u_k - \lambda\left(\sum_{k\in\mathcal{K}}(u_k - u_k^2)\right) \quad (22)$$

$$s.t.: \ U_k(\Omega_k) \le f_k(\tilde{\mathbf{W}}_k, u_k) - \xi_k\sigma_k^2, \forall k \in \mathcal{K}, \quad (22a)$$

$$R_k(\xi_k) \ge u_k R_{\mathrm{th}}^k, \quad \forall k \in \mathcal{K}, \quad (22b)$$

$$\xi_k \ge 0, \forall k \in \mathcal{K}, \quad (22c)$$

$$(9a), (9b), (10), (15c), (15d),$$

where λ is a large constant that acts as a penalty factor. It should be noted that the objective function and constraints (22a) and (22b) belong to the class of DC problems. Thus, the SCA technique can be directly applied to approximate the non-convex problem in each iteration. Indeed, the objective function and constraints (22a) and (22b) are approximated by a more tractable one at a given local point. To this end, we use first-order Taylor expansion to obtain a globally lower-bound of functions $G_k(\xi_k)$ and $Q_k(\Omega_k), \forall k \in \mathcal{K}$. By denoting $\nabla_\square$ as representing the gradient with respect to $\square$, the lower-bounds of these functions at iteration t are respectively given by:

$$G_k(\xi_k) \le \tilde{G}_k(\xi_k) \triangleq G_k(\xi_k^t) + \partial_{\xi_k}^T G_k(\xi_k^t)(\xi_k - \xi_k^t), \forall k \in \mathcal{K}, \quad (23)$$

$$Q_k(\Omega_k) \ge \tilde{Q}_k(\Omega_k) \triangleq Q_k(\Omega_k^t) + \partial_{\Omega_k}^T Q_k\left(\Omega_k^t\right)(\Omega_k - \Omega_k^t) + \mathrm{Tr}\left(\nabla_{\tilde{\mathbf{W}}_k}^H Q_k\left(\Omega_k^t\right)\left(\tilde{\mathbf{W}}_k - \tilde{\mathbf{W}}_k^t\right)\right), \forall k \in \mathcal{K}. \quad (24)$$

Therefore, $\tilde{R}_k(\xi_k) = F_k(\xi_k) - \tilde{G}_k(\xi_k)$ and $\tilde{U}_k(\Omega_k) = P_k(\Omega_k) - \tilde{Q}_k(\Omega_k)$. Then, P_4 with any given local point at iteration t can be approximated as:

$$\mathrm{P}_4 : \max_{\Omega_k} \ \alpha \frac{\sum_{k\in\mathcal{K}} \tilde{R}_k(\xi_k)}{E} + (1-\alpha)\sum_{k\in\mathcal{K}} u_k - \lambda\left(\sum_{k\in\mathcal{K}}(u_k - ((u_k^t)^2 - 2u_k^t(u_k - u_k^t))\right) \quad (25)$$

$$s.t.: \ \tilde{U}_k(\Omega_k) \le f_k(\mathbf{W}_k, u_k) - \xi_k\sigma_k^2, \forall k \in \mathcal{K}, \quad (25a)$$

$$\tilde{R}_k(\xi_k) \ge u_k R_{\mathrm{th}}^k, \quad \forall k \in \mathcal{K}, \quad (25b)$$

$$(9a), (9b), (10), (15c), (15d), (22c),$$

where, $E = \sum_{k\in\mathcal{K}} \mathrm{Tr}(\tilde{\mathbf{W}}_k) + P_\mathrm{s} + NP_d + P_\mathrm{c}^{\mathrm{AP}}$. The objective function in P_4 is now in a format of concave-convex in which we use semidefinite relaxation (SDR) to remove the rank-one constraint (15c). In order to solve P_4, we use the fractional programming method based on the quadratic transformation, which introduces an auxiliary parameter to transform a fractional form function into an equivalent subtractive form. To do so, we utilize the result of Corollary 1 in [15] as follows:

Corollary 1. Consider f as a non-decreasing function, then the sum-of-ratio problem

$$\max_{\mathbf{x}} \ \frac{f_{Obj}(\mathbf{x})}{g_{Obj}(\mathbf{x})} \quad (26a)$$

$$s.t.: \ \mathbf{x} \in \mathcal{X}, \quad (26b)$$

is equivalent to the following problem

$$\max_{\mathbf{x},\ m_{Obj}} \ 2m_{Obj}\sqrt{f_{Obj}(\mathbf{x})} - m_{Obj}^2 g_{Obj}(\mathbf{x}) \quad (27a)$$

$$s.t.: \ \mathbf{x} \in \mathcal{X}, m_{Obj} \in \mathbb{R}, \quad (27b)$$

where m_{Obj} is an auxiliary variable. The proof of the equivalence between (26) and (27) is provided in [15]. When $f_{Obj}(\mathbf{x})$ is a concave function with respect to $\mathbf{x}$ in a convex set $\mathcal{X}$, the subtractive function $2m_{Obj}\sqrt{f_{Obj}(\mathbf{x})} - m_{Obj}^2 g_{Obj}(\mathbf{x})$ would be a concave function with respect to $\mathbf{x}$. Consequently, the resulting problem in (27) is a convex optimization problem for a given m_{Obj}. Finally, we note that the optimal auxiliary variable is given by $m_{Obj} = \sqrt{f_{Obj}(\mathbf{x})}/g_{Obj}(\mathbf{x})$. Thus, we can develop an iterative algorithm with a polynomial-time computational complexity to update $\mathbf{x}$ and m_{Obj} alternatingly. However, the algorithm is only guaranteed to converge to a sub-optimal solution of the main problem in (27) if the transformed problem in (28) can globally be solved [15]. In the following, we demonstrate how to execute the quadratic transformation to achieve a sub-optimal solution of P_4. The problem P_4 can be transformed into the following equivalent optimization problem by adopting the quadratic transformation in (26) and (27):

$$\mathrm{P}_5 : \max_{\Omega_k,\ m_{Obj}} \ \alpha\left(2m_{Obj}\sqrt{\sum_{k\in\mathcal{K}}\tilde{R}_k(\xi_k)} - m_{Obj}^2 E\right) + (1-\alpha)\sum_{k\in\mathcal{K}} u_k - \lambda\left(\sum_{k\in\mathcal{K}} u_k - \left((u_k^t)^2 - 2u_k^t(u_k - u_k^t)\right)\right) \quad (28)$$

$$s.t.: \quad (9a), (9b), (10), (15d), (22c), (25a), (25b),$$

where m_{Obj} denotes the new auxiliary variable corresponding to the objective function of the optimization problem in P_5 and can be updated globally as $m_{Obj} = \sqrt{\sum_{k\in\mathcal{K}}\tilde{R}_k(\xi_k)}/E$. The resulting subtractive function in (28) is concave with respect to the optimization variables for given auxiliary variables. Generally, P_5 yields a solution with a rank higher than one due to constraint (15c). Therefore, to solve (28) for a given m_{Obj}, we use the SDR to remove constraint (15c). Thus, we rewrite the constraint in a mathematically tractable form via the DC method represented as:

$$||\mathbf{W}||_* - ||\mathbf{W}||_2 \le 0. \quad (29)$$

Note that $||\mathbf{W}||_* = \sum_i \sigma_i \ge ||\mathbf{W}||_2 = \max_i\{\sigma_i\}$ holds for any given $\mathbf{W} \in \mathbb{H}^{M\times M}$, where σ_i is the i-th singular value of $\mathbf{W}$. The equality holds if and only if $\mathbf{W}$ achieves rank one i.e., $\mathrm{Rank}(\mathbf{W}) = 1$ [14]. Now, we take the first-order Taylor approximation of $||\mathbf{W}||_2$ as: $\overbrace{\quad\quad\quad\quad}^{=\phi(\mathbf{W})}$

$$||\mathbf{W}||_2 \ge \overbrace{||\mathbf{W}^{(t)}||_2 + \mathrm{Tr}\left(\lambda_{\max}(\mathbf{W}^{(t)})\lambda_{\max}^H(\mathbf{W}^{(t)})(\mathbf{W} - \mathbf{W}^t)\right)}. \quad (30)$$

By resorting to (30), a convex approximation can be obtained for (29) which is given by $\tilde{\phi}^t(\mathbf{W}) \triangleq ||\mathbf{W}||_* - \phi(\mathbf{W}) \le 0$. As a result, by augmenting $\tilde{\phi}^t(\mathbf{W})$ to the objective function of P_6 with $\psi \gg 1$ as a penalty factor to penalize any non-rank-one matrix, the optimization problem in the $(t+1)$-iteration can be written as follows:

$$\mathrm{P}_6 : \max_{\Omega_k} \ \alpha\left(2m_{Obj}\sqrt{\sum_{k\in\mathcal{K}}\tilde{R}_k(\xi_k)} - m_{Obj}^2 E\right) + (1-\alpha)\sum_{k\in\mathcal{K}} u_k$$

$$-\lambda\Big(\sum_{k\in\mathcal{K}} u_k - \Big((u_k^t)^2 - 2u_k^t(u_k - u_k^t)\Big)\Big) - \psi(\tilde{\phi}^t(\mathbf{W}))$$

$$s.t.: \ a(9a),(9b),(10),(15d),(22c),(25a),(25b). \tag{31}$$

Consequently, P_6 is a convex optimization problem and can be efficiently solved.

B. Second-stage: Optimizing $\boldsymbol{\Theta}$

With given $\tilde{\mathbf{W}}_k$, the optimization problem would be converted to the data rate maximization. The main difficulty of optimizing the phase shifts at the IRS is constraint (8b). To be more specific, constraint (8b) is a unit-module constraint, which makes solving the problem intractable. Accordingly, we first define $\boldsymbol{v} = (e^{j\alpha_1}, ..., e^{j\alpha_N})^H \in \mathbb{C}^{N\times 1}$ and $\tilde{\boldsymbol{v}} = [\boldsymbol{v}^T \ \tau]^T \in \mathbb{C}^{(N+1)\times 1}$, respectively, where $\tau \in \mathbb{C}$ is a dummy variable with $|\tau| = 1$. To facilitate the solution design, we also define $\mathbf{V} = \tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^H \in \mathbb{C}^{(N+1)\times(N+1)}$. Thus, we obtain $\big|(\mathbf{h}_{iu,k}^H\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_{bu,k}^H)\tilde{\mathbf{W}}_k\big|^2 \triangleq \mathrm{Tr}(\mathbf{V}\mathbf{X}_k\tilde{\mathbf{W}}_k\mathbf{X}_k^H) = \mathrm{Tr}(\tilde{\mathbf{W}}_k\mathbf{Y}_k)$, where $\mathbf{X}_k = [(\mathrm{diag}(\mathbf{h}_{iu,k}^H)\mathbf{H})^T \ \ \mathbf{h}_{bu,k}^*]^T, \mathbf{Y}_k = \mathbf{X}_k^H\mathbf{V}\mathbf{X}_k$. Similarly, we handle the non-convex constraint (8a) as well as a new objective function, where the contribution of u_k and total power is ignored in the objective function as was solved in the previous sub-problem. To do so, we adopt the same approach as for the beamforming via introducing a set of auxiliary variables (Υ_k), and then we employ the SCA approach, which is omitted here due to lack of space. This means $\tilde{R}_k(\Upsilon_k) = F_k(\Upsilon_k) - \tilde{G}_k(\Upsilon_k)$. Now, we restate the optimization problem as follows

$$\mathrm{P}_7: \max_{\mathbf{V},\Upsilon_k} \ \sum_{k\in\mathcal{K}} \tilde{R}_k(\Upsilon_k) \tag{32}$$

$$s.t.: \Upsilon_k \geq 0, \mathbf{V} \succeq \mathbf{0}, \Omega_k = \{\Upsilon_k, \mathbf{V}\}, \quad \forall k \in \mathcal{K}, \tag{32a}$$

$$\tilde{R}_k(\Upsilon_k) \geq u_k R_{\mathrm{th}}^k, \qquad \forall k \in \mathcal{K}, \tag{32b}$$

$$\tilde{U}_k(\Omega_k) \leq f_k(\mathbf{V}) - \xi_k\sigma_k^2, \forall k \in \mathcal{K}, \tag{32c}$$

$$\mathrm{Rank}(\mathbf{V}) \leq 1. \tag{32d}$$

Similar to P_6, P_7 usually does not give a rank-one solution because of constraint (32d). By rewriting (32d) as $||\mathbf{V}||_* - ||\mathbf{V}||_2 \leq 0$, and owing to (30), a convex approximation, $\tilde{\phi}^t(\mathbf{V}) \leq 0$, of rank-one constraint can be made. Thus, supplementing $\tilde{\phi}^t(\mathbf{V})$ to the objective function of P_8 with $\zeta \gg 1$ as a penalty factor to penalize any non-rank-one matrix, the optimization problem in the $(t+1)$-iteration can be written as follows:

$$\mathrm{P}_8: \max_{\mathbf{V},\Upsilon_k} \sum_{k\in\mathcal{K}} \tilde{R}_k(\Upsilon_k) - \zeta(\tilde{\phi}^t(\mathbf{V})), \ s.t.: (32a) - (32c). \tag{33}$$

The optimization problem P_8 now can be efficiently solved just as P_6. The proposed AO-based algorithm is summarized in **Algorithm 1**, which converges to a locally optimal solution.

Proposition 1: The objective function of P_1 is monotonically non-decreasing through the iterative algorithm.

Proof 1: The proof of Proposition 1 closely follows the step in [16], and is thus omitted here due to page limitation.

V. Computational Complexity

Here, we investigate the computational complexity of our proposed algorithm. The order of complexity for an SDP problem with m SDP constraints, which includes an $n \times n$ positive semi-definite (PSD) matrix, can be found to be $\mathcal{O}\left(\sqrt{n}\log(1/\zeta)(mn^3 + m^2n^2 + m^3)\right)$ where $\zeta{>}0$ is the solution accuracy [14]. As a result, with $m = 6k{+}1$ and $n = M$

Algorithm 1 Proposed Algorithm

Input: Set $m_{Obj}^{(0)}$, $I_{\max}$, and $D_{\max}$.

1: **repeat**
2: Calculate $\tilde{G}_k(\xi_k)$, $\tilde{Q}_k(\xi_k, \tilde{\mathbf{W}}_k)$, and $\tilde{\phi}^2(\mathbf{W})$ via a successive convex approximation (SCA) structure.
3: Solve P_6 for a given $\boldsymbol{\Theta}$, and $m_{Obj}^{(d-1)}$.
4: **if** $|\sqrt{\sum_{k\in\mathcal{K}} \tilde{R}_k^{(d)}(\xi_k)} - m_{\mathrm{Obj}}^{(d-1)} - E^{(d)}| \leq \varepsilon$
5: **return** $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(d)}$, $m_{Obj}^* = m_{Obj}^{(d-1)}$.
6: **else** Update $m_{Obj}^{(d)} = \sqrt{\sum_{k\in\mathcal{K}} \tilde{R}_k(\xi_k)/E}$, **end if**.
7: $d = d + 1$.
8: **until** $d = D_{\max}$
9: Calculate $\tilde{G}_k(\boldsymbol{\Upsilon}_k)$ and $\tilde{\phi}^2(\mathbf{V})$ via an SCA structure.
10: Solve P_8 for the obtained $\tilde{\mathbf{W}}$, u_k, from the previous steps.
11: $i = i + 1$
12: **until** $i = I_{\max}$.
13: **return** $\boldsymbol{\Upsilon}, \mathbf{V}$.

the complexity to solve P_6 follows $O_1 = \mathcal{O}\log(1/\zeta_1)(6K + 1)\big((M)^{3.5} + (6K+1)^{2.5}M^2 + (6K+1)^2\big)$, with the solution accuracy ζ_1. In a similar manner, the computational complexity for the second sub-problem is $O_2 = \mathcal{O}\log(1/\zeta_2)(4K + 1)\big((N)^{3.5} + (4K+1)^{2.5}N^2 + (4K+1)^2\big)$, with ζ_2 solution accuracy. As a result, the overall complexity of the proposed solution, P_8, is $\mathcal{O}(I_{\mathrm{iter}}(O_1 + O_2))$, where I_{iter} is the number of iterations that is required for convergence of the AO approach.

VI. Numerical Results

This section demonstrates the proposed algorithm's effectiveness for maximizing EE and IoT user admission in IRS-enabled systems with short packet lengths. A rectangular area with a dimension of $(100, 100)$ meters is considered, where the AP is placed at $(0, 0)$ m, while the IRS is located at $(50, 0)$ m, and all the users are assumed to be randomly located inside the rectangular area. The path loss model is given by $35.3 + 37.6\log_{10}(d_k)$[dB], where d_k indicates the distance between AP-user k in kilometer. Furthermore, the AO convergence tolerance is set as 10^{-2}, and a thermal noise density of -174 [dBm/Hz] is assumed. Besides, the value of the decoding error probability for user k is given by $\epsilon_k = 10^{-7}$. In addition, it is assumed that $K = 20$, $M = 5$, $m_d = 250$, and $R_{\mathrm{th}}^k = 1.6$ [bits/Sec/Hz] for all simulation setups [8], [9].

Fig. 2 shows the average EE versus different values of the maximum transmit power with $\alpha = 1$. For comparison, we consider three baseline schemes as follows: For baseline scheme 1, we optimize the network's data rate [8]. For baseline scheme 2, we assume that the passive beamforming at the IRS is random, and for baseline scheme 3, we suppose that there is no IRS in the system to assist the network. No matter the scheme considered, it can be observed that the EE first increases and then saturates with increasing $p_{\max}$. However, for baseline scheme 1, the EE starts to decline as the objective of the optimization problem is to maximize the data rate. This conveys the fact that once the system reaches the maximum EE by solely maximizing the data rate, further growth in the transmit power increases the total network power consumption, degrading the EE of the system. This figure also shows the

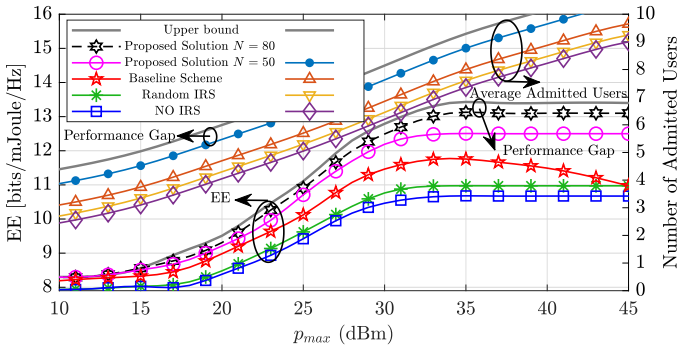

Fig. 2. EE and the average number of admitted users vs. $p_{\max}$

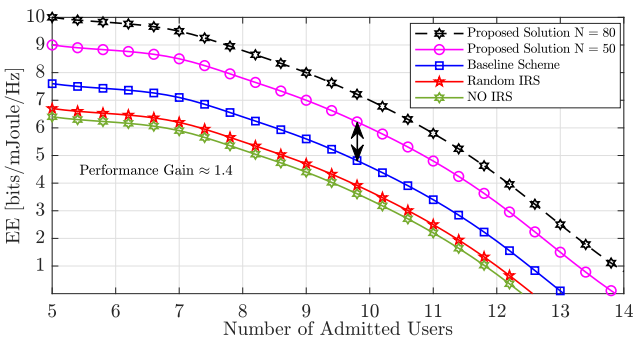

Fig. 3. EE vs. the average number of admitted users.

effectiveness of the phase shift optimization and proves that the EE scales up with an increasing number of reflecting elements. Fig. 2 also plots the average number of admitted users versus $p_{\max}$ with $\alpha = 0$. This figure investigates that the total number of admitted IoT users also increases with increasing $p_{\max}$. This is because the network would be able to support more users and satisfy the quality of finite blocklength IoT users by increasing the power budget. This figure reveals that our proposed scheme performs better than other baseline schemes due to deploying IRS and jointly optimizing the active and passive beamforming matrices at the AP and IRS. Finally, we compared our benchmark algorithm with an (unachievable) performance upper bound, Shannon's capacity formula, when V_k in (5) is set to 0. Fig. 3 plots the tradeoff region between EE and the number of IoT admitted users for different values of $0 < \alpha < 1$ with step size 0.05. There is a non-trivial tradeoff between EE and the number of admitted users — the EE is a monotonically decreasing function of the number of admitted users. In other words, maximizing the EE results in a reduction in the number of admitted users. An interesting observation is that the EE optimization outperforms the number of admitted users for high values of α, i.e., a limited number of users are accepted by the network even though users will enjoy high data rate services in this case. In contrast, the number of admitted users boosts higher for low values of α. Hence, dropping α obliges the optimization problem to focus more on maximizing the number of admitted users while satisfying users' minimum QoS requirements. Thus, although the fairness is improved, the EE performance declines in low α values.

VII. Conclusion and Future Work

This paper investigated the resource allocation for a DL multiuser MISO IRS system by adopting short packet transmission. In particular, the resource allocation design via active/passive beamforming was formulated to maximize the EE together with the number of IoT admitted users while considering QoS requirements for each MTC-typed user. The underlying problem was non-convex. To handle this difficulty, we first employed the AO method to divide the main problem into two sub-problems, i.e., active/passive beamforming sub-problems. Then we adopted the SCA approach and a penalty-based method to solve the beamforming matrices sub-problems. Simulation results investigated our proposed scheme, considering that the IRS could help the IRS system with short packet length users meet the QoS and improve the EE significantly compared to other conventional methods. In our feature work, we will consider designing the IRS based on the physics-based models, which is more practical while considering the imperfect CSI.

References

[1] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, pp. 3313–3351, May 2021.

[2] X. Yu, D. Xu, and R. Schober, "Optimal beamforming for MISO communications via intelligent reflecting surfaces," in *Proc. IEEE SPAWC*, pp. 1–5, 2020.

[3] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for intelligent reflecting surface enhanced wireless networks," in *Proc. IEEE GLOBECOM*, pp. 1–6, 2019.

[4] S. Zargari, A. Khalili, and R. Zhang, "Energy efficiency maximization via joint active and passive beamforming design for multiuser MISO IRS-aided SWIPT," *IEEE Wirel. Commun. Lett.*, vol. 10, pp. 557–561, Mar. 2021.

[5] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, pp. 59–65, Sep. 2016.

[6] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2307–2359, May 2010.

[7] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wirel. Comm.*, vol. 18, pp. 402–415, Jan. 2019.

[8] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFDMA-URLLC systems," *IEEE Trans. Commun.*, vol. 68, pp. 7184–7200, Nov. 2020.

[9] K. .Singh, M.-L. Ku, and M. F. Flanagan, "Energy-efficient precoder design for downlink multi-user MISO networks with finite blocklength codes," *IEEE Trans. Green Commun. Netw.*, vol. 5, pp. 160–173, Mar. 2021.

[10] L. Zhao, S. Yang, X. Chi, W. Chen, and S. Ma, "Achieving energy-efficient uplink URLLC with MIMO-aided grant-free access," *IEEE Trans. Wirel. Commun.*, vol. 21, pp. 1407–1420, Feb. 2022.

[11] M. Darabi, V. Jamali, L. Lampe, and R. Schober, "Hybrid puncturing and superposition scheme for joint scheduling of URLLC and eMBB traffic," *IEEE Commun. Lett.*, vol. 26, pp. 1081–1085, May 2022.

[12] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, pp. 2666–2682, Nov. 2020.

[13] H. Ren, K. Wang, and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *IEEE Trans. Commun.*, vol. 70, pp. 707–723, Jan. 2022.

[14] A. Rezaei, A. Khalili, J. Jalali, H. Shafiei, and Q. Wu, "Energy-efficient resource allocation and antenna selection for IRS-assisted multicell downlink networks," *IEEE Wirel. Commun. Lett.*, vol. 11, pp. 1229–1233, Jun. 2022.

[15] K. Shen and W. Yu, "Fractional programming for communication systems — Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, pp. 2616–2630, May 2018.

[16] S. Zargari, A. Khalili, Q. Wu, M. Robat Mili, and D. W. K. Ng, "Max-Min fair energy-efficient beamforming design for intelligent reflecting surface-aided SWIPT systems with non-linear energy harvesting model," *IEEE Trans. Veh. Technol*, vol. 70, pp. 5848–5864, Jun. 2021.