Robust Beamforming Design for an IRS-Aided NOMA Communication System with CSI Uncertainty

Yasaman Omid, S. M. Mahdi Shahabi, Cunhua Pan, Yansha Deng, Arumugam Nallanathan, *Fellow, IEEE*

Abstract

Intelligent reflecting surface (IRS) is a promising technology that provides high throughput in future communication systems and is compatible with various communication techniques, such as non-orthogonal multiple-access (NOMA). This paper studies the downlink transmission of IRS-assisted NOMA communication, considering the practical case of imperfect channel state information (CSI). Aiming to maximize the system sum rate, a robust IRS-aided NOMA design is proposed to jointly find the optimal beamforming vector for the access point and the passive reflection matrix for the IRS. This robust design is realised using the penalty dual decomposition (PDD) scheme, and it is shown that the results have a close performance to their upper bound obtained from the corresponding perfect CSI scenario. The presented method is compatible with both continuous and discrete phase shift elements of the IRS. Our findings show that the proposed algorithms, for both continuous and discrete IRS, have low computational complexity compared to other schemes in the literature. Furthermore, we conduct a performance comparison between the IRS-aided NOMA and the IRS-aided orthogonal multiple access (OMA). This comparison shows that robust beamforming techniques are crucial for the system to reap the advantages of IRS-aided NOMA communication in the presence of CSI uncertainty.

Yasaman Omid and Arumugam Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. (e-mail: y.omid@qmul.ac.uk; a.nallanathan@qmul.ac.uk).

Cunhua Pan was with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. and he is now with the National Mobile Communications Research Laboratory, Southeast University, China. (e-mail:c.pan@qmul.ac.uk, cpan@seu.edu.cn)

S. M. Mahdi Shahabi and Yansha Deng are with the Department of Engineering, King's College London, U.K. (e-mail: mahdi.shahabi@kcl.ac.uk; yansha.deng@kcl.ac.uk).

Index Terms

Intelligent Reflecting Surface, Reconfigurable Intelligent Surface, Robust Design, NOMA, OMA.

I. Introduction

ONE of the most recognized multiple-access techniques for future communication systems is the non-orthogonal multiple access (NOMA), which enables multiple users to share the same resource block. These resource blocks could be orthogonal bandwidths, different time slots or orthogonal spatial directions, when the system contains multiple-antenna access points (AP). Contrary to NOMA, in the orthogonal multiple access (OMA) techniques, each resource block is dedicated to only one user. Thus, by using NOMA, the system's spectral efficiency is potentially improved by a factor of K compared to OMA systems, where K is the number of users served on each resource block. Despite all its benefits, the application of NOMA is limited to the case where the directions of users' channel vectors are similar [1]. Therefore, to broaden the application of NOMA and improve its performance, intelligent reflecting surfaces (IRS) can be utilized to manipulate the direction of the users' channel vectors [2].

IRS is a promising technology for the next generation of wireless networks, improving both spectral and energy efficiency through passive phase shifters that reconfigure the wireless propagation environment [3]–[5]. IRS is compatible with numerous communication techniques, such as millimeter-wave, Terahertz, physical layer security, simultaneous wireless information and power transfer (SWIPT), unmanned aerial vehicle networks (UAV), MIMO, and NOMA [6]–[9]. Most IRS-assisted system designs assume that cascaded AP-IRS-user channels are available at the AP, as the estimation of IRS-related channels separately is difficult without active elements. Recent works have addressed the issue of channel estimation error in IRS-aided systems, proposing robust solutions [10]–[12]. Thus, in this paper, we focus on designing a robust IRS-aided NOMA communication system.

A. Prior Works Regarding IRS-Aided NOMA

The application of the IRS for NOMA communication has attracted extensive research attention. In [2], the authors proposed a simple IRS-assisted NOMA downlink transmission. First, they employed spatial-division multiple-access (SDMA) at the AP to generate orthogonal beams

by using the spatial directions of nearby user channels. Then, the IRS-assisted NOMA was used to ensure that additional cell-edge users can also be served on these beams by aligning the effective channels of cell-edge users with the predetermined beams. In [13], the performance comparison between NOMA and two types of OMA, namely frequency-division multiple-access (FDMA) and time-division multiple-access (TDMA) was studied in an IRS-aided downlink communication network. The transmit power minimization problem was solved for discrete phase shifters at the IRS. It was shown that TDMA is always more power-efficient than FDMA, but the power consumption of TDMA compared to NOMA depends upon the target rate and the location of the users. The authors of [14] investigated the spectral efficiency improvement of NOMA networks by considering an IRS-aided single-input single-output (SISO) network. Specifically, they attempted to enhance the performance of the user with the best channel gain, while all other users depended on the IRS. The authors of [15] considered the downlink transmit power minimization problem for the IRS-aided NOMA system. They addressed the resulting intractable non-convex bi-quadratic problem by solving the non-convex quadratic problems alternately, and to solve the non-convex quadratic problems, they employed a difference-of-convex (DC) programming algorithm. In [16], for the first time, the sum rate maximization problem was addressed for a MISO IRS-aided NOMA system in the downlink transmission. By using the alternating optimization technique, they designed the passive phase shifts of the IRS and the active AP beamforming vector, alternately, for two cases of ideal and non-ideal IRS phase shifters. In addition, the authors of [17] introduced a fairness-driven solution for optimizing the rate performance in an IRS-assisted NOMA system, where a max-min problem was defined and solved for target decoding signal-to-interference-plus-noise-ratio (SINR) of all users by employing a combined-channel- strength (CCS)-based user ordering method, resulting in readily decoupling the problem. In [18], the downlink transmission of an IRS-aided NOMA system was investigated in terms of the energy efficiency/spectral efficiency trade-off. A multi-objective optimization problem was formulated for jointly optimizing the power allocation, active precoding and passive beamforming. Then, the weighted sum method was used to transform this problem into a single objective problem which was then solved via an iterative alternating optimization algorithm, where in each iteration a convex optimization was solved by CVX. In [19], [20], the authors aimed to maximize the sum rate in the uplink transmission of an IRS-aided NOMA system, and they proposed a semi-definite relaxation (SDR)-based algorithm that reached near-optimal solutions. The authors of [19] showed that the IRS-aided NOMA outperforms the

IRS-aided OMA in terms of the sum rate. The authors of [21], studied both uplink and downlink transmission of IRS-aided NOMA and OMA systems. In [22], the authors investigated the situations in vehicular communications where IRS-aided NOMA outperforms its counterparts realized by OMA, and consequently presented an IRS deployment policy in hybrid access-assisted vehicular networks. In [23], IRS-aided covert communications in the downlink and uplink transmission were studied where a legitimate transmitter applied NOMA in communicating with a covert user terminal and a public user through leveraging the uncertainty appeared in the wireless environments, whereas the legitimate transmitter-covert user link remained hidden. Authors of [24], studied the performance of IRS-aided NOMA systems with a particular focus on 1-bit coding while taking the residual interference into account within successive interference cancellation (SIC) scheme to avoid error propagation and quantization error and consequently decoding errors. In [25], an IRS-aided backscatter NOMA system with only two user terminals was studied and the closed-form outage probability expressions were derived and validated in terms of the power allocation, the number of IRS reflectors, the IRS coefficients and the rate thresholds. In [26], the power optimization of a secure RIS-aided NOMA communication system was investigated to mitigate the information leakage to the eavesdroppers while the legitimate users' reception quality vulnerability was minimized. However, this study considered the channel state information (CSI) uncertainty only for the eavesdropper and assumed that the perfect CSI of the legitimate users were available at the AP through some channel estimation methods. The authors in [27], presented a detailed overview of the IRS-aided NOMA communication systems in the literature.

B. Motivations and Contributions

All of the aforementioned contributions have considered the availability of perfect CSI, while due to the passive nature of the IRS, channel acquisition in IRS-aided systems is quite challenging, especially for IRS-related channels. Although, some papers have presented robust beamforming designs for IRS-aided MIMO systems [10]–[12], [28], [29], to the best of our knowledge, no one has ever considered the effect of CSI uncertainty entirely in IRS-assisted NOMA and IRS-aided OMA systems. To address this issue, a robust design for an IRS-assisted NOMA communication system is devised in this paper. To better clarify, the contributions of this paper are summarized as follows:

- In this paper, for the first time, the channel estimation error is considered in the IRS-aided NOMA communication, and a robust design for IRS-assisted NOMA is presented. To do so, an approximation for the channel estimation error is derived and the rates of the users are formulated based on the CSI uncertainty. Then, using the block successive upper bound minimization/maximization (BSUM) [30], we calculate a tractable lower bound for the sum rate. The penalty dual decomposition (PDD) technique [31] is utilised to design a joint beamforming technique for the IRS and the AP that is robust to the CSI uncertainty. We maximize the lower bound of the sum rate through a low-complexity iterative algorithm, where in each step closed-form solutions are calculated for the optimization variables. The performance of the robust IRS-aided NOMA is compared with that of the non-robust and the perfect CSI scenarios.
- We establish a comparison benchmark by providing a solution for the sum rate maximization problem in an IRS-aided OMA system. To this end, we present a robust IRS-aided OMA design using FDMA and TDMA. By comparing NOMA and OMA in three cases of perfect CSI, imperfect CSI with robust design and imperfect CSI with non-robust design, we provide a new understanding of IRS-aided OMA and IRS-aided NOMA systems and their performance in different situations.
- We consider both discrete and continuous IRS, and we present optimal solutions for passive beamforming in both cases, without increasing the computational complexity. These passive beamforming methods are used for both IRS-aided NOMA and IRS-aided OMA communication systems.
- Based on our results we can claim that: i) the proposed method is robust to channel estimation error, ii) the computational complexity of the proposed algorithm is very low compared to other methods in the literature, iii) IRS-assisted NOMA outperforms IRS-assisted OMA in terms of spectral efficiency under perfect CSI scenario, and iv) in the imperfect CSI scenario, the IRS-aided NOMA undergoes a critical performance loss and in severe channel uncertainties, it performs inferior to IRS-aided OMA unless the robust beamforming design is used. Thus, according to our findings, it can be claimed that consideration of CSI uncertainty is crucial in IRS-aided NOMA systems, and the necessity of a robust beamforming design is then highlighted.

The rest of this paper is organized as follows. The system model is presented in section

II. Section III is dedicated to the problem formulation of IRS-aided NOMA and section IV provides the PDD-based algorithm as a solution. In Section V the IRS-assisted OMA system is presented and the sum rate maximization problem is solved for FDMA and TDMA modes. Section VI provides a complexity analysis for the IRS-aided NOMA and OMA designs. Section VII includes the simulation results, and finally, section VIII concludes this paper. Throughout the paper, the variables, constants, vectors and matrices are represented by small italic letters, capital italic letters, small bold letters and capital bold letters, respectively. If $\mathbf{A}$ represents a matrix, the element in its ith row and jth column is represented by a_{ij}, and its ith column/row is referred to by $\mathbf{a}_i$. The notation $|.|$ denotes the absolute of a variable, the notation $||.||$ stands for the norm of a vector or the Frobenius norm of a matrix, depending on the argument, and $||.||_\infty$ stands for the infinity norm of a matrix. Also, $\mathrm{Re}\{.\}$, $(.)^*$ and $(.)^H$ stand for the real part of a complex variable, the conjugate of a complex variable and the conjugate transpose of a complex vector/matrix, respectively.

II. System Model

Consider the downlink transmission of an AP with N antennas to $2Z$ single-antenna users. This transmission is aided by an IRS equipped with M phase shifters. A smart controller in the AP is responsible for managing the IRS by sharing information and coordinating the transmission. The power of the signals being reflected by the IRS multiple times is much smaller than that of the signal reflected once, and thus it can be ignored [32]. We investigate an indoor application undergoing a rich scattering propagation environment; hence, we consider only non-line-of-sight (NLOS) channels in our model [33]–[37]. As shown in Fig. 1, the users are uniformly distributed in a disc $\mathcal{D}_o$ with the radius r_o, where the AP is located at the centre. In addition, we assume that this disc is broken down into two zones, i.e. a smaller disc $\mathcal{D}_i$ with radius r_i and the same origin is located inside the disc $\mathcal{D}_o$. Specifically, the radius of $\mathcal{D}_i$ is smaller than that of the $\mathcal{D}_o$, that is $r_o > r_i$. Furthermore, L primary users are assumed to be randomly distributed in $\mathcal{D}_i$, while K secondary users are uniformly distributed in the rest of $\mathcal{D}_o$ outside $\mathcal{D}_i$, i.e. $\{\mathcal{D}_o - \mathcal{D}_i\}$. Hence, user $l \in \{1, ..., L\}$ and user $k \in \{1, ..., K\}$ are randomly scheduled and matched together in a cluster. As the users in distinct clusters are allocated with orthogonal resource blocks owing to employing frequency division multiplexing (OFDM), and consequently they might hardly interfere each other, we aim to investigate the effects of intra-cluster interference rather than

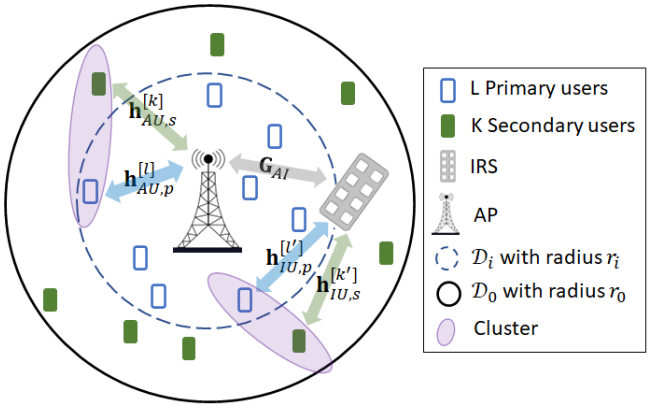


Figure 1: The system model and cluster design [38].

inter-cluster interference.The clusters here refer to a classification of effective channels into superior and inferior channels in a NOMA system. For simplicity, we assume $L = K$, thus we have Z clusters in the system, where $Z = L = K$.

According to Fig. 1, the channel vectors from the AP to the l-th primary user and the k-th secondary user, and from the IRS to these users are denoted by $\mathbf{h}_{AU,p}^{[l]} \in \mathbb{C}^{1\times N}$, $\mathbf{h}_{AU,s}^{[k]} \in \mathbb{C}^{1\times N}$, $\mathbf{h}_{IU,p}^{[l]} \in \mathbb{C}^{1\times M}$ and $\mathbf{h}_{IU,s}^{[k]} \in \mathbb{C}^{1\times M}$, respectively, while each element of these vectors undergoes the Rayleigh flat fading, i.e. $\mathbf{h}_{AU,p}^{[l]} \sim \mathcal{CN}(0, \beta_{AU,p}^{[l]}\mathbf{I})$, $\mathbf{h}_{AU,s}^{[k]} \sim \mathcal{CN}(0, \beta_{AU,s}^{[k]}\mathbf{I})$, $\mathbf{h}_{IU,p}^{[l]} \sim \mathcal{CN}(0, \beta_{IU,p}^{[l]}\mathbf{I})$ and $\mathbf{h}_{IU,s}^{[k]} \sim \mathcal{CN}(0, \beta_{IU,s}^{[k]}\mathbf{I})$, where $\beta_{AU,p}^{[l]}$, $\beta_{AU,s}^{[k]}$, $\beta_{IU,p}^{[l]}$ and $\beta_{IU,s}^{[k]}$ represent the large-scale fading coefficients of their respective channels. The channel matrix between the AP and the IRS is represented by $\mathbf{G}_{AI} \in \mathbb{C}^{M\times N}$ which also follows the complex normal distribution as $\mathrm{vec}(\mathbf{G}_{AI}) \sim \mathcal{CN}(0, \beta_{AI}\mathbf{I})$, with β_{AI} being its large-scale fading coefficient. It is assumed that the cascaded AP-IRS-user channels are available at the AP imperfectly, and the source of this imperfection is the mobility of the users. Hence, we can assume $\mathbf{h}_{AU,p}^{[l]} = \hat{\mathbf{h}}_{AU,p}^{[l]} + \tilde{\mathbf{h}}_{AU,p}^{[l]}$, $\mathbf{h}_{AU,s}^{[k]} = \hat{\mathbf{h}}_{AU,s}^{[k]} + \tilde{\mathbf{h}}_{AU,s}^{[k]}$, $\mathbf{h}_{IU,p}^{[l]} = \hat{\mathbf{h}}_{IU,p}^{[l]} + \tilde{\mathbf{h}}_{IU,p}^{[l]}$ and $\mathbf{h}_{IU,s}^{[k]} = \hat{\mathbf{h}}_{IU,s}^{[k]} + \tilde{\mathbf{h}}_{IU,s}^{[k]}$, where $\hat{\mathbf{h}}_{AU,p}^{[l]}$, $\hat{\mathbf{h}}_{AU,s}^{[k]}$, $\hat{\mathbf{h}}_{IU,p}^{[l]}$ and $\hat{\mathbf{h}}_{IU,s}^{[k]}$ represent the estimated channel vectors, and $\tilde{\mathbf{h}}_{AU,p}^{[l]}$, $\tilde{\mathbf{h}}_{AU,s}^{[k]}$, $\tilde{\mathbf{h}}_{IU,p}^{[l]}$ and $\tilde{\mathbf{h}}_{IU,s}^{[k]}$ stand for the channel estimation error vectors with the following distributions

$$\tilde{\mathbf{h}}_{AU,p}^{[l]} \ , \ \tilde{\mathbf{h}}_{AU,s}^{[k]} \sim \mathcal{CN}\left(0, \sigma_{AU}^2\mathbf{I}\right), \tag{1}$$

and

$$\tilde{\mathbf{h}}_{IU,p}^{[l]} \ , \ \tilde{\mathbf{h}}_{IU,s}^{[k]} \sim \mathcal{CN}\left(0, \sigma_{IU}^2\mathbf{I}\right). \tag{2}$$

We assume only the distribution of channel estimation error caused by the users' mobility with

respect to the AP and IRS, the estimated cascaded AP-IRS-user channels, and the estimated AP-user channels are needed for the beamforming design [10]. Therefore, by defining an error distribution as the estimation error threshold, we aim to compensate for a tolerated estimation error together with optimizing joint AP-IRS elements. Note that the superscripts in the notation of channels demonstrate the number of the user, and the subscripts inform us about the two nodes of the channel and the type of the user, i.e. primary (p) or secondary (s). For example, the subscript AU means that the channel in question is between the AP and a user. For the rest of this paper, all superscripts in $[.]$ represent the number of a user, the letter p in the subscripts refers to a primary user, and the letter s in the subscripts refers to a secondary user.

Prior to optimizing IRS-AP elements, we should consider a user pairing policy for applying NOMA to the users with the same dedicated resource blocks. The users are initially categorized into primary and secondary users, located at the center and edge of a cell defined by the disk $\mathcal{D}_o$. The pairing process involves selecting a primary user from those within $\mathcal{D}_i$ and a secondary user from those in $\mathcal{D}_o - \mathcal{D}_i$ for each primary-secondary user pair. This paper investigates the use of both continuous and discrete phase shifters for IRS:

- **Continuous IRS**: In an ideal case, the IRS phase shifters are continuous, i.e. the set of phase shifts can be expressed as $\mathbf{\Psi}_I = \{\theta_i | \theta_i = e^{j\psi_i}, \psi_i \in [0, 2\pi]\}$.
- **Discrete IRS**: In a non-ideal scenario, the phase shifts are selected from a discrete set of phases, i.e. $\mathbf{\Psi}_N = \{\theta_i | \theta_i = e^{j\psi_i}, \psi_i \in \mathcal{B}\}$, where $\mathcal{B} = \{\frac{2\pi m}{\Delta}, m = 1, ..., \Delta\}$ and Δ denotes the number of possible phases that can be selected by each phase shifter of the IRS. In other words, Δ implies the resolution of the phase shifters.

III. IRS-Aided NOMA

We assume the information-bearing vector $\mathbf{s} \in \mathbb{C}^{(L+K)\times 1}$ is expressed as

$$\mathbf{s} = \left[s_p^{[1]}, \ldots, s_p^{[L]}, s_s^{[1]}, \ldots, s_s^{[K]}\right]^T, \tag{3}$$

where $s_p^{[l]}$ and $s_s^{[k]}$ represent the signal intended for the l-th primary user and the k-th secondary user, respectively. In this case, $\mathbf{x}_p^{[l]} \triangleq \mathbf{w}_{l,p} s_p^{[l]}$ and $\mathbf{x}_s^{[k]} \triangleq \mathbf{w}_{k,s} s_s^{[k]}$ denote the transmit vectors intended for the l-th primary user and the k-th secondary user, respectively, where $\mathbf{w}_{l,p}$ and $\mathbf{w}_{k,s}$ are their corresponding active beamforming vectors at the AP, and $\mathbf{w}_{l,p}\mathbf{w}_{l,p}^H$ and $\mathbf{w}_{k,s}\mathbf{w}_{k,s}^H$ are

their corresponding transmit powers. The signal received by the l-th primary user and the k-th secondary user are then written as

$$y_p^{[l]} = \left(\mathbf{h}_{AU,p}^{[l]} + \mathbf{h}_{IU,p}^{[l]}\mathbf{\Psi}\mathbf{G}_{AI}\right)\left(\mathbf{x}_p^{[l]} + \mathbf{x}_s^{[\mathcal{U}(s,l,p)]}\right) + n_p^{[l]} = \mathbf{h}_p^{[l]}\mathbf{x}_p^{[l]} + \mathbf{h}_p^{[l]}\mathbf{x}_s^{[\mathcal{U}(s,l,p)]} + n_p^{[l]}, \quad (4)$$

$$y_s^{[k]} = \left(\mathbf{h}_{AU,s}^{[k]} + \mathbf{h}_{IU,s}^{[k]}\mathbf{\Psi}\mathbf{G}_{AI}\right)\left(\mathbf{x}_s^{[k]} + \mathbf{x}_p^{[\mathcal{U}(p,k,s)]}\right) + n_s^{[k]} = \mathbf{h}_s^{[k]}\mathbf{x}_s^{[k]} + \mathbf{h}_s^{[k]}\mathbf{x}_p^{[\mathcal{U}(p,k,s)]} + n_s^{[k]}, \quad (5)$$

where $\mathcal{U}(s,l,p)$ stands for the secondary user that occupies the same spectrum as that of the l-th primary user, $\mathcal{U}(p,k,s)$ stands for the primary user that occupies the same spectrum as that of the k-th secondary user, $\mathbf{\Psi}$ represents the diagonal phase shift matrix of the IRS, and $\mathbf{h}_p^{[l]}$ and $\mathbf{h}_s^{[k]}$ are the effective channels between the AP and the users which are defined by $\mathbf{h}_p^{[l]} = \mathbf{h}_{AU,p}^{[l]} + \mathbf{h}_{IU,p}^{[l]}\mathbf{\Psi}\mathbf{G}_{AI}$ and $\mathbf{h}_s^{[k]} = \mathbf{h}_{AU,s}^{[k]} + \mathbf{h}_{IU,s}^{[k]}\mathbf{\Psi}\mathbf{G}_{AI}$, respectively. Finally, $n_p^{[l]} \sim \mathcal{CN}(0,\sigma_n^2)$ and $n_s^{[k]} \sim \mathcal{CN}(0,\sigma_n^2)$ represent the additive white Gaussian noise at their respective users. Now, by reformulating the received signal in (4) and (5) in terms of the estimated channels, we have

$$y_p^{[l]} = \hat{\mathbf{h}}_p^{[l]}\mathbf{x}_p^{[l]} + \hat{\mathbf{h}}_p^{[l]}\mathbf{x}_s^{[\mathcal{U}(s,l,p)]} + \tilde{\mathbf{h}}_p^{[l]}\mathbf{x}_p^{[l]} + \tilde{\mathbf{h}}_p^{[l]}\mathbf{x}_s^{[\mathcal{U}(s,l,p)]} + n_p^{[l]}, \quad (6)$$

$$y_s^{[k]} = \hat{\mathbf{h}}_s^{[k]}\mathbf{x}_s^{[k]} + \hat{\mathbf{h}}_s^{[k]}\mathbf{x}_p^{[\mathcal{U}(p,k,s)]} + \tilde{\mathbf{h}}_s^{[k]}\mathbf{x}_s^{[k]} + \tilde{\mathbf{h}}_s^{[k]}\mathbf{x}_p^{[\mathcal{U}(p,k,s)]} + n_s^{[k]}, \quad (7)$$

where

$$\hat{\mathbf{h}}_{\{p,s\}}^{[i]} \triangleq \hat{\mathbf{h}}_{AU,\{p,s\}}^{[i]} + \mathbf{v}\hat{\mathbf{H}}_{c,\{p,s\}}^{[i]}, \quad (8)$$

$$\tilde{\mathbf{h}}_{\{p,s\}}^{[i]} \triangleq \tilde{\mathbf{h}}_{AU,\{p,s\}}^{[i]} + \mathbf{v}\tilde{\mathbf{H}}_{c,\{p,s\}}^{[i]}, \quad (9)$$

$$\mathbf{v} \triangleq \left(\mathrm{diag}\left(\mathbf{\Psi}\right)\right)^T = \left[\theta_1, \ldots, \theta_M\right]. \quad (10)$$

In (8) and (9), $\hat{\mathbf{H}}_{c,\{p,s\}}^{[i]} = \mathrm{diag}\left(\hat{\mathbf{h}}_{IU,\{p,s\}}^{[i]}\right)\mathbf{G}_{AI}$ is the estimated cascaded AP-IRS-user channel and $\tilde{\mathbf{H}}_{c,\{p,s\}}^{[i]} = \mathrm{diag}\left(\tilde{\mathbf{h}}_{IU,\{p,s\}}^{[i]}\right)\mathbf{G}_{AI}$ represents the cascaded channel estimation error. The distribution of the effective channel estimation error vector $\tilde{\mathbf{h}}_{\{p,s\}}^{[i]}$ for large values of M is approximated by

$$\tilde{\mathbf{h}}_{\{p,s\}}^{[i]} \sim \mathcal{CN}\left(0, \sigma_h^2\mathbf{I}\right), \quad (11)$$

where $\sigma_h^2 = \sigma_{AU}^2 + M\sigma_{IU}^2\beta_{AI}$. The approximation in (11) is derived in the appendix of [10] where it is shown that this approximation becomes more accurate when the number of IRS elements is large. Also, it can be easily shown that this approximation remains valid whether the IRS phase shifts are continuous or discrete. This approximation is valid for uncorrelated channels, however, up to a certain amount of correlation is tolerable. The effect of channel correlation on the system performance can be a topic to be analyzed in future works. We consider the decoding order $\{k,l\}$

for decoding user $k \in \{1, ..., K\}$ and user $l \in \{1, ..., L\}$ while using the successive interference cancellation (SIC) method. In other words, the l-th primary user first decodes the data flow of the $\mathcal{U}(s, l, p)$-th secondary user, then by cancelling its resulting interference, it decodes its own data flow. The k-th secondary user, then, decodes its own data flow by treating the signal from the $\mathcal{U}(p, k, s)$-th primary user as the interference term. With this assumption, the following theorem can be presented.

Theorem 1. *The achievable sum rate of the users is obtained by*

$$R^{[sum]} = \sum_{l=1}^{L} R_p^{[l]} + \sum_{k=1}^{K} R_s^{[k]}, \tag{12}$$

in which

$$R_p^{[l]} = \frac{1}{Z} \log_2 \Big(1 + \frac{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{l,p}|^2}{\frac{\sigma_h^2}{Z} \left(||\mathbf{w}_{l,p}||^2 + ||\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2\right) + \frac{\sigma_n^2}{Z}}\Big), \tag{13}$$

is the minimum achievable rate of the l-th primary user and

$$R_s^{[k]} = \frac{1}{Z} \log_2 \Big(1 + \frac{|\hat{\mathbf{h}}_s^{[k]} \mathbf{w}_{k,s}|^2}{|\hat{\mathbf{h}}_s^{[k]} \mathbf{w}_{\mathcal{U}(p,k,s),p}|^2 + \frac{\sigma_h^2}{Z} \left(||\mathbf{w}_{\mathcal{U}(p,k,s),p}||^2 + ||\mathbf{w}_{k,s}||^2\right) + \frac{\sigma_n^2}{Z}}\Big), \tag{14}$$

is the minimum achievable rate of the k-th secondary user. In addition, the achievable rate of the secondary user $\mathcal{U}(s, l, p)$ when it is decoded by the l-th primary user is obtained as follows

$$R_{s,p}^{[\mathcal{U}(s,l,p)],l} = \frac{1}{Z} \log_2 \Big(1 + \frac{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{l,p}|^2 + \frac{\sigma_h^2}{Z} \left(||\mathbf{w}_{l,p}||^2 + ||\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2\right) + \frac{\sigma_n^2}{Z}}\Big), \tag{15}$$

The proof of Theorem 1 is given in Appendix A. To ensure that the l-th primary user can cancel the interference from the $\mathcal{U}(s, l, p)$-th secondary user, the rate of the $\mathcal{U}(s, l, p)$-th secondary user when it is decoded by the l-th primary user should be no lower than the rate of the $\mathcal{U}(s, l, p)$-th secondary user when it decodes its own data flow, i.e. $R_{s,p}^{[\mathcal{U}(s,l,p)],l} \geq R_s^{[\mathcal{U}(s,l,p)]},\ l \in \{1, ..., L\}$ [17]. Based on (14) and (15), this can also be achieved if $\frac{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{l,p}|^2 + b_l} \geq \frac{|\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]} \mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]} \mathbf{w}_{l,p}|^2 + b_l},\ l \in \{1, ..., L\}$, where $b_l = \frac{\sigma_h^2}{Z} \left(||\mathbf{w}_{l,p}||^2 + ||\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2\right) + \frac{\sigma_n^2}{Z}$. Now, we aim to jointly optimize the active beamforming at the AP and the phase shifts at the IRS to maximize the sum achievable data rate. Specifically, the optimization problem is formulated as follows

$$\begin{aligned} \max_{\mathbf{W}, \mathbf{v}} \quad & \textstyle\sum_{l=1}^{L} R_p^{[l]} + \sum_{k=1}^{K} R_s^{[k]} \\ \text{s. t.} \quad & |\theta_i| = 1,\ i \in \{1, ..., M\}, \\ & \text{trace}\left(\mathbf{W}\mathbf{W}^H\right) \leq P, \\ & \frac{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_p^{[l]} \mathbf{w}_{l,p}|^2 + b_l} \geq \frac{|\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]} \mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]} \mathbf{w}_{l,p}|^2 + b_l},\ l \in \{1, ..., L\}, \end{aligned} \tag{16}$$

where $\mathbf{W} = [\mathbf{w}_{1,p}, \ldots, \mathbf{w}_{L,p}, \mathbf{w}_{1,s}, \ldots, \mathbf{w}_{K,s}]$, and P is the AP power budget. The optimization problem (16) incorporates $R_p^{[l]}$ and $R_s^{[k]}$ which are nonlinear components with respect to (w.r.t) $\mathbf{v}$ and $\mathbf{W}$, and make the problem non-convex. Additionally, the final constraint in (16) implies a strong and mutual coupling between $\mathbf{v}$ and $\mathbf{W}$ which can be hardly dealt with in its initial form. On the other hand, it is obvious that most presented methods for optimizing IRS elements take continuous phases in consideration for the reflection coefficients, whereas practical scenarios impose discrete phase shifts making the optimization problem intractable. Hence, in this paper, we aim to address this non-convex problem towards finding near-optimal solutions with a low complexity form in both cases of continuous and discrete phase shifts.

IV. The PDD Technique

In this section, inspired by the PDD method [31], [39], we propose a novel algorithm to solve the problem in (16). To this end, we first introduce a set of auxiliary variables as $\mathcal{S} = \{\mathbf{T}, \mathbf{v}, \mathbf{W}, \bar{\mathbf{W}}\}$, where $\mathbf{T} = \mathbf{W}^H \hat{\mathbf{H}}^H$, $\bar{\mathbf{W}} = \mathbf{W}$, and we define

$$\hat{\mathbf{H}} \triangleq \left[\hat{\mathbf{h}}_p^{[1]\ T}, \ldots, \hat{\mathbf{h}}_p^{[L]\ T}, \hat{\mathbf{h}}_s^{[1]\ T}, \ldots, \hat{\mathbf{h}}_s^{[K]\ T}\right]^T. \tag{17}$$

With these definitions, the optimization problem becomes

$$\begin{aligned}
\max_{\mathcal{S}} \quad & \textstyle\sum_{l=1}^{L} R_p^{[l]} + \sum_{k=1}^{K} R_s^{[k]} \\
\text{s.t.} \quad & |\theta_i| = 1,\ i \in \{1, ..., M\}, \\
& \text{trace}\left(\bar{\mathbf{W}}\bar{\mathbf{W}}^H\right) \leq P, \\
& |t_{l,l}|^2 \leq \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2,\ l \in \{1, ..., L\}, \\
& \left|t_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)}\right|^2 \leq \min\left\{\left|t_{l,\mathcal{U}(s,l,p)+L}\right|^2, \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2\right\},\ l \in \{1, ..., L\}, \\
& \mathbf{T} = \mathbf{W}^H \hat{\mathbf{H}}^H,\ \bar{\mathbf{W}} = \mathbf{W}.
\end{aligned} \tag{18}$$

Note that if the third and fourth constraints of (18) hold, then the third constraint of (16) will definitely hold, but not vice versa. Please refer to Appendix B for further explanation. Also, unlike the third constraint of (16), the two new constraints in (18) no longer have any limitations on the level of estimation error and noise, but this feature is created by limiting the feasibility region. In other words, by limiting the feasibility region, the third constraint of (16) has been simplified. To obtain the augmented Lagrangian problem, all equality conditions except for $|\theta_i| = 1, i \in \{1, ..., M\}$, should be brought into the objective function [10], [39]. The reason is that by keeping this condition in the constrains, a closed-form solution can be calculated

with projection. The details are explained later in subsection IV-A. This leads to the following optimization problem

$$\begin{aligned}
\max_{\mathcal{S}} \quad & \textstyle\sum_{l=1}^{L} R_p^{[l]} + \sum_{k=1}^{K} R_s^{[k]} - Q_\gamma(\mathcal{S}) \\
\text{s.t.} \quad & |\theta_i| = 1,\ i \in \{1, ..., M\}, \\
& \text{trace}\left(\bar{\mathbf{W}}\bar{\mathbf{W}}^H\right) \leq P, \\
& |t_{l,l}|^2 \leq \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2,\ l \in \{1, ..., L\}, \\
& \left|t_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)}\right|^2 \leq \min\left\{\left|t_{l,\mathcal{U}(s,l,p)+L}\right|^2, \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2\right\},\ l \in \{1, ..., L\},
\end{aligned} \tag{19}$$

where $Q_\gamma(\mathcal{S})$ is calculated by

$$Q_\gamma(\mathcal{S}) = \frac{1}{2\gamma}\Big(\sum_{i=1}^{L+K} ||\mathbf{w}_i - \bar{\mathbf{w}}_i + \gamma\boldsymbol{\lambda}_{w_i}||^2 + \sum_{i=1}^{L+K}\sum_{j=1}^{L+K} |t_{i,j} - \mathbf{w}_i^H\hat{\mathbf{h}}^{[j]H} + \gamma\lambda_{h_{ij}}|^2\Big). \tag{20}$$

In (20), $\mathbf{w}_i$ and $\bar{\mathbf{w}}_i$ stand for the ith column of $\mathbf{W}$ and $\bar{\mathbf{W}}$, respectively. Also, $\hat{\mathbf{h}}^{[j]}$ represents the j-th row of the matrix $\hat{\mathbf{H}}$. The penalty parameter for the Lagrangian function is denoted by γ, and $\boldsymbol{\lambda}_{w_i}$ and $\lambda_{h_{ij}}$ are the dual variables of the equality conditions in problem (18). The vector $\boldsymbol{\lambda}_{w_i}$ is the ith column of the matrix $\boldsymbol{\Lambda}_w$ and $\lambda_{h_{ij}}$ is the element in the ith row and jth column of $\boldsymbol{\Lambda}_h$. The constraint $\mathbf{T} = \mathbf{W}^H\hat{\mathbf{H}}^H$ is only necessary to hold for the diagonal elements of $\mathbf{T}$ and the elements of $\mathbf{T}$ which represent the primary and the secondary users that are paired together. Thus (20) can be rewritten by

$$Q_\gamma(\mathcal{S}) = \frac{1}{2\gamma}\Big(\sum_{i=1}^{L+K} ||\mathbf{w}_i - \bar{\mathbf{w}}_i + \gamma\boldsymbol{\lambda}_{w_i}||^2 + \sum_{i=1}^{L+K}\sum_{\substack{j=1,\\ j\in\mathcal{J}_i}}^{L+K} |t_{i,j} - \mathbf{w}_i^H\hat{\mathbf{h}}^{[j]H} + \gamma\lambda_{h_{ij}}|^2\Big), \tag{21}$$

where $\mathcal{J}_i = \{j | j = i\ ,\ j = L + \mathcal{U}(s,i,p)\ or\ j = \mathcal{U}(p, i-L, s)\}$. With this in mind, all other elements of $\mathbf{T}$ can be set to zero. Now, in order to apply BSUM to (19), we introduce the following theorem.

Theorem 2. *The objective function of the optimization problem in (18) can be rewritten as*

$$\max_{q_i,r_i,c_i,d_i} \sum_{l=1}^{L}\Big(\log(c_l) - c_l f_l(q_l,\mathcal{S})\Big) + \sum_{k=1}^{K}\Big(\log(d_k) - d_k g_k(r_k,\mathcal{S})\Big), \tag{22}$$

where

$$f_l(q_l,\mathcal{S}) = |1 - q_l^*\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{l,p}|^2 + \frac{\sigma_h^2}{Z}(||q_l^*\mathbf{w}_{l,p}||^2 + ||q_l^*\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2) + \frac{\sigma_n^2}{Z}|q_l|^2, \tag{23}$$

$$g_k(r_k,\mathcal{S}) = |1 - r_k^*\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{k,s}|^2 + |r_k^*\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{\mathcal{U}(p,k,s),p}|^2 + \frac{\sigma_h^2}{Z}(||r_k^*\mathbf{w}_{\mathcal{U}(p,k,s),p}||^2 + ||r_k^*\mathbf{w}_{k,s}||^2) + \frac{\sigma_n^2}{Z}|r_k|^2. \tag{24}$$

The proof of this theorem can be simply achieved by using the first-order optimality condition

on (22). Note that the function in (22) is convex w.r.t. the variables q_l, r_k, c_l and d_k. To obtain the optimal values for these variables, we set the derivative of (22) w.r.t. q_l^*, c_l^*, r_k^*, and d_k^*, to zero, which results in the following solutions

$$q_l^{opt}(\mathcal{S}) = \frac{\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{l,p}}{|\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{l,p}|^2 + \frac{\sigma_h^2}{Z}\left(||\mathbf{w}_{l,p}||^2 + ||\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2\right) + \frac{\sigma_n^2}{Z}}, \tag{25}$$

$$c_l^{opt}(\mathcal{S}) = \frac{1}{f_l(q_l,\mathcal{S})} = 1 + \frac{|\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{l,p}|^2}{\frac{\sigma_h^2}{Z}\left(||\mathbf{w}_{l,p}||^2 + ||\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2\right) + \frac{\sigma_n^2}{Z}}, \tag{26}$$

$$r_k^{opt}(\mathcal{S}) = \frac{\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{k,s}}{|\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{k,s}|^2 + |\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{\mathcal{U}(p,k,s),p}|^2 + \frac{\sigma_h^2}{Z}\left(||\mathbf{w}_{\mathcal{U}(p,k,s),p}||^2 + ||\mathbf{w}_{k,s}||^2\right) + \frac{\sigma_n^2}{Z}}, \tag{27}$$

$$d_k^{opt}(\mathcal{S}) = \frac{1}{g_k(r_k,\mathcal{S})} = 1 + \frac{|\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{k,s}|^2}{|\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{\mathcal{U}(p,k,s),p}|^2 + \frac{\sigma_h^2}{Z}\left(||\mathbf{w}_{\mathcal{U}(p,k,s),p}||^2 + ||\mathbf{w}_{k,s}||^2\right) + \frac{\sigma_n^2}{Z}}. \tag{28}$$

This theorem gives a lower bound for the objective function of (18) as

$$Z\Big(\sum_{l=1}^{L} R_p^{[l]} + \sum_{k=1}^{K} R_s^{[k]}\Big) \geq \sum_{l=1}^{L}\Big(\log(c_l) - c_l f_l(q_l,\mathcal{S})\Big) + \sum_{k=1}^{K}\Big(\log(d_k) - d_k g_k(r_k,\mathcal{S})\Big). \tag{29}$$

This locally tight lower bound is tractable and by using it as the objective function, the following optimization problem is driven

$$\begin{aligned}
\min_{\mathcal{S}} \quad & \sum_{l=1}^{L} c_l f_l(q_l,\mathcal{S}) + \sum_{k=1}^{K} d_k g_k(r_k,\mathcal{S}) + Q_\gamma(\mathcal{S}) \\
\text{s.t.} \quad & |\theta_i| = 1,\ i \in \{1,...,M\}, \\
& \text{trace}\left(\bar{\mathbf{W}}\bar{\mathbf{W}}^H\right) \leq P, \\
& |t_{l,l}|^2 \leq \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2,\ l \in \{1,...,L\}, \\
& \left|t_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)}\right|^2 \leq \min\left\{\left|t_{l,\mathcal{U}(s,l,p)+L}\right|^2, \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2\right\},\ l \in \{1,...,L\}.
\end{aligned} \tag{30}$$

Note that (30) is obtained by using the right hand side of (29) instead of the objective function in (19), and removing the constant terms, $\log(c_l)$ and $\log(d_k)$ from the objective function.

A. The BSUM Algorithm

By applying the BSUM method, the optimization problem in (30) can be solved via the following steps, each resulting in a closed-form solution for a sub-set of the variables.

1) Step 1: Solving (30) for $\mathbf{W}$: In order to determine the minimum point of the objective function presented in equation (30) w.r.t. $\mathbf{W}$, which is expressed as $p(\mathbf{W}) = \sum_{l=1}^{L} c_l f_l(q_l,\mathcal{S}) + \sum_{k=1}^{K} d_k g_k(r_k,\mathcal{S}) + Q_\gamma(\mathcal{S})$, the first-order optimality condition is employed. This involves setting

the derivative of the objective function w.r.t. $\mathbf{w}_{l,p}^H$ and $\mathbf{w}_{k,s}^H$ to zero, as $\frac{\partial p(\mathbf{W})}{\partial \mathbf{w}_{l,p}^H} = \mathbf{0}$ and $\frac{\partial p(\mathbf{W})}{\partial \mathbf{w}_{k,s}^H} = \mathbf{0}$. These equations result in

$$\begin{aligned}\mathbf{w}_{l,p} =&\Big(2\gamma|q_l|^2 c_l \frac{\sigma_h^2}{Z}\mathbf{I} + 2\gamma|q_l|^2 c_l \hat{\mathbf{h}}_p^{[l]H}\hat{\mathbf{h}}_p^{[l]} + 2\gamma|r_{\mathcal{U}(s,l,p)}|^2 d_{\mathcal{U}(s,l,p)}\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]H}\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]} + \mathbf{I}\\ &+ 2\gamma|r_{\mathcal{U}(s,l,p)}|^2 d_{\mathcal{U}(s,l,p)}\frac{\sigma_h^2}{Z}\mathbf{I} + \hat{\mathbf{h}}_p^{[l]H}\hat{\mathbf{h}}_p^{[l]} + \hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]H}\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]}\Big)^{-1}\Big(\bar{\mathbf{w}}_{l,p} - \gamma\lambda_{w_{l,p}} + 2\gamma c_l q_l \hat{\mathbf{h}}_p^{[l]H}\\ &+ \hat{\mathbf{h}}_p^{[l]H}(t_{l,l}^* + \gamma\lambda_{h_{l,l}}^*) + \hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]H}(t_{l,L+\mathcal{U}(s,l,p)}^* + \gamma\lambda_{h_{l,L+\mathcal{U}(s,l,p)}}^*)\Big),\ l \in \{1,\dots,L\}, \end{aligned} \tag{31}$$

$$\begin{aligned}\mathbf{w}_{k,s} =&\Big(2\gamma|q_{\mathcal{U}(p,k,s)}|^2 c_{\mathcal{U}(p,k,s)}\frac{\sigma_h^2}{Z}\mathbf{I} + 2\gamma|r_k|^2 d_k \hat{\mathbf{h}}_s^{[k]H}\hat{\mathbf{h}}_s^{[k]} + 2\gamma|r_k|^2 d_k\frac{\sigma_h^2}{Z}\mathbf{I} + \mathbf{I} + \hat{\mathbf{h}}_p^{[\mathcal{U}(p,k,s)]H}\hat{\mathbf{h}}_p^{[\mathcal{U}(p,k,s)]}\\ &+ \hat{\mathbf{h}}_s^{[k]H}\hat{\mathbf{h}}_s^{[k]}\Big)^{-1}\Big(\bar{\mathbf{w}}_{k,s} - \gamma\lambda_{w_{k,s}} + 2\gamma d_k r_k \hat{\mathbf{h}}_s^{[k]H} + \hat{\mathbf{h}}_s^{[k]H}(t_{L+k,L+k}^* + \gamma\lambda_{h_{L+k,L+k}}^*)\\ &+ \hat{\mathbf{h}}_p^{[\mathcal{U}(p,k,s)]H}(t_{L+k,\mathcal{U}(p,k,s)}^* + \gamma\lambda_{h_{L+k,\mathcal{U}(p,k,s)}}^*)\Big),\ k \in \{1,\dots,K\}, \end{aligned} \tag{32}$$

for the primary users and the secondary users, respectively. The results in (31) and (32) are global minimum points of $p(\mathbf{W})$, since this function is differentiable and convex w.r.t $\mathbf{W}$.

2) Step 2: Solving (30) for $\bar{\mathbf{W}}$: Now, we need to consider the auxiliary variable $\bar{\mathbf{W}}$. The corresponding optimization problem w.r.t $\bar{\mathbf{W}}$ is a projection of a point onto a ball centered at the origin[1]. The closed-form solution exists as

$$\bar{\mathbf{W}} = \mathcal{P}_P(\mathbf{W} + \gamma\mathbf{\Lambda}_w), \tag{33}$$

where $\mathcal{P}_P(\mathbf{X})$ denotes the projection of $\mathbf{X}$ onto the convex set P.

3) Step 3: Solving (30) for $\mathbf{T}$: In this step, the optimization problem is solved w.r.t $\mathbf{T}$. Similar to step 2, we can obtain the variable $\mathbf{T}$ through projection. To do so, we project the points $t_{l,l}$ and $t_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)}$, $\forall l \in \{1,...,L\}$, into the balls centered at the origin with radii $|t_{L+\mathcal{U}(s,l,p),l}|^2$ and $\min\left\{\left|t_{l,\mathcal{U}(s,l,p)+L}\right|^2, \left|t_{L+\mathcal{U}(s,l,p),l}\right|^2\right\}$, $\forall l \in \{1,...,L\}$, respectively. In other words, $2L$ projections take place to solve the problem w.r.t $\mathbf{T}$, as

$$t_{l,l} = \mathcal{P}_{P_1}(\mathbf{w}_{l,p}^H\mathbf{h}_p^{[l]H} + \gamma\lambda_{h_{l,l}}), \tag{34}$$

$$t_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)} = \mathcal{P}_{P_2}(\mathbf{w}_{\mathcal{U}(s,l,p),s}^H\mathbf{h}_s^{[\mathcal{U}(s,l,p)]H} + \gamma\lambda_{h_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)}}), \tag{35}$$

where $\mathcal{P}_{P_1}(x)$ and $\mathcal{P}_{P_2}(x)$ denote the projection of x onto the convex sets P_1 and P_2, respectively.

4) Step 4: Solving (30) for $\mathbf{v}$: The final step is to solve the problem for the variable $\mathbf{v}$. To do so, we first replace $\hat{\mathbf{h}}^{[i]}$ by its definition in (8) and we write the objective function in (30) w.r.t

[1]The projection of a point $\mathbf{X}$ into a set $\mathcal{X}$ is defined by $\min_{\mathbf{P}\in\mathcal{X}} \|\mathbf{X}-\mathbf{P}\|$. In the case $\mathcal{X}$ is a sphere centered at the origin with radius of R ($\mathcal{X} = \{\mathbf{X}|\|\mathbf{X}\| \le R\}$), the projection of $\mathbf{X}$ is obtained by $R\frac{\mathbf{X}}{\|\mathbf{X}\|+\max(0,R-\|\mathbf{X}\|)}$.

the variable $\mathbf{v}$. Assuming we have continuous phase shifters in the IRS, the problem could be rewritten as

$$\min_{|\theta_i|=1,\ i\in\{1,\ldots,M\}} \quad \mathbf{v}\mathbf{A}\mathbf{v}^H - 2\,\mathrm{Re}\{\mathbf{n}\mathbf{v}^H\}, \tag{36}$$

where,

$$\begin{aligned}\mathbf{A} = &\sum_{k=1}^{K} d_k|r_k|^2\hat{\mathbf{H}}_c^{[L+k]}\mathbf{w}^{[L+k]}\mathbf{w}^{[L+k]H}\hat{\mathbf{H}}_c^{[L+k]H} + d_k|r_k|^2\hat{\mathbf{H}}_c^{[L+k]}\mathbf{w}^{[\mathcal{U}(p,k,s)]}\mathbf{w}^{[\mathcal{U}(p,k,s)]H}\hat{\mathbf{H}}_c^{[L+k]H}\\ &+\sum_{l=1}^{L} c_l|q_l|^2\hat{\mathbf{H}}_c^{[l]}\mathbf{w}^{[l]}\mathbf{w}^{[l]H}\hat{\mathbf{H}}_c^{[l]H} + \sum_{i=1}^{L+K}\sum_{\substack{j=1,\\ j\in\mathcal{J}_i}}^{L+K}\hat{\mathbf{H}}_c^{[j]}\mathbf{w}_i\mathbf{w}_i^H\hat{\mathbf{H}}_c^{[j]H},\end{aligned} \tag{37}$$

$$\begin{aligned}\mathbf{n} = &\sum_{i=1}^{L+K}\sum_{\substack{j=1,\\ j\in\mathcal{J}_i}}^{L+K} t_{i,j}^*\mathbf{w}_i^H\hat{\mathbf{H}}_c^{[j]H} - \hat{\mathbf{h}}_{AU}^{[j]}\mathbf{w}_i\mathbf{w}_i^H\hat{\mathbf{H}}_c^{[j]H} + \gamma\lambda_{h_{i,j}}^*\mathbf{w}_i^H\hat{\mathbf{H}}_c^{[j]H}\\ &+\sum_{k=1}^{K} d_k r_k\mathbf{w}^{[L+k]H}\hat{\mathbf{H}}_c^{[L+k]H} - d_k|r_k|^2\hat{\mathbf{h}}_{AU}^{[L+k]}\mathbf{w}^{[L+k]}\mathbf{w}^{[L+k]H}\hat{\mathbf{H}}_c^{[L+k]H}\\ &+\sum_{k=1}^{K} d_k|r_k|^2\hat{\mathbf{h}}_{AU}^{[L+k]}\mathbf{w}^{[\mathcal{U}(p,k,s)]}\mathbf{w}^{[\mathcal{U}(p,k,s)]H}\hat{\mathbf{H}}_c^{[L+k]H}\\ &+\sum_{l=1}^{L} c_l q_l\mathbf{w}^{[l]H}\hat{\mathbf{H}}_c^{[l]H} - c_l|q_l|^2\hat{\mathbf{h}}_{AU}^{[l]}\mathbf{w}^{[l]}\mathbf{w}^{[l]H}\hat{\mathbf{H}}_c^{[l]H}.\end{aligned} \tag{38}$$

In (37) and (38), $\hat{\mathbf{H}}_c^{[j]}$ is the cascaded channel of the j-th user when $j \in \{1, ..., L+K\}$, and it is defined by $\hat{\mathbf{H}}_c^{[j]} = \mathrm{diag}(\hat{\mathbf{h}}_{IU}^{[j]})\mathbf{G}_{AI}$, where $\hat{\mathbf{h}}_{IU}^{[j]}$ is the j-th row of $[\hat{\mathbf{h}}_{IU,p}^{[1]T}, ..., \hat{\mathbf{h}}_{IU,p}^{[L]T}, \hat{\mathbf{h}}_{IU,s}^{[1]T}, ..., \hat{\mathbf{h}}_{IU,s}^{[K]T}]^T$. Now for one specified $\theta_k, k = 1, ..., M$, the optimization problem becomes

$$\max_{|\theta_k|=1} \quad 2\,\mathrm{Re}\{e_k\theta_k^*\}, \tag{39}$$

in which $e_k = n_k - \sum_{i\neq k}^{M}\theta_i a_{ik}$. The solution to this optimization problem is

$$\theta_k = \frac{e_k}{|e_k|}. \tag{40}$$

For a scenario where a discrete IRS is used with the discrete set of phases $\mathbf{\Psi}_N$, the following optimization problem should be solved

$$\max_{\theta_k\in\mathbf{\Psi}_N, k=1,\ldots,M} \quad 2\,\mathrm{Re}\{e_k\theta_k^*\}. \tag{41}$$

The optimization problem (41) can be further simplified as

$$\max_{\theta_k\in\mathbf{\Psi}_N, k=1,\ldots,M} \quad 2\,|e_k|\cos(\psi_k - \angle e_k). \tag{42}$$

where $\angle e_k$ denotes the argument of e_k. The optimal solution to the problem in (42) can be found

as follows

$$\theta_k = \left\{ e^{j\psi_m} \middle| m = \arg\max_{m'} \left\{ \cos\left(\frac{2\pi m'}{\Delta} - \angle e_k \right),\ m' \in \{1, \ldots, \Delta\} \right\} \right\}. \tag{43}$$

For instance, the solution to this optimization problem, given 1-bit phase shifters ($\Delta = 2$), for a specified θ_k, $k = 1, ..., M$ is given by

$$\theta_k = \begin{cases} e^{j\psi_1}, & \mathrm{Re}\{e_k\} \leq 0 \\ e^{j\psi_2}, & \text{otherwise}, \end{cases} \tag{44}$$

while for 2-bit phase shifters ($\Delta = 4$), the solution to Problem (42) is

$$\theta_k = \begin{cases} e^{j\psi_1}, & |\,\mathrm{Re}\{e_k\}| \leq |\mathrm{Im}\{e_k\}| \text{ and } \mathrm{Im}\{e_k\} \geq 0, \\ e^{j\psi_2}, & |\,\mathrm{Re}\{e_k\}| \geq |\mathrm{Im}\{e_k\}| \text{ and } \mathrm{Re}\{e_k\} \leq 0, \\ e^{j\psi_3}, & |\,\mathrm{Re}\{e_k\}| \leq |\mathrm{Im}\{e_k\}| \text{ and } \mathrm{Im}\{e_k\} \leq 0, \\ e^{j\psi_4}, & |\,\mathrm{Re}\{e_k\}| \geq |\mathrm{Im}\{e_k\}| \text{ and } \mathrm{Re}\{e_k\} \geq 0. \end{cases} \tag{45}$$

Given the solutions for a specific $k \in \{1, ..., M\}$ in (43), an iterative algorithm is proposed to alternately optimize one phase shift while keeping the others fixed until final convergence. This algorithm is guaranteed to converge in both cases of continuous and discrete IRS phase shifts.

The BSUM algorithm to solve the optimization problem in (30) is summarised in Algorithm 1. The parameters are initialized at the beginning of the overall algorithm which will be explained in the following sub-section. The termination criterion of Algorithm 1, is achieving a certain accuracy for the variables, i.e. Algorithm 1 is terminated when $b(\bar{\mathbf{W}}, \mathbf{W}, \mathbf{v}, \mathbf{T}) = ||\bar{\mathbf{W}}(i) - \bar{\mathbf{W}}(i-1)|| + ||\mathbf{W}(i) - \mathbf{W}(i-1)|| + ||\mathbf{v}(i) - \mathbf{v}(i-1)|| + ||\mathbf{T}(i) - \mathbf{T}(i-1)|| < \eta_0 a(\bar{\mathbf{W}}, \mathbf{W}, \mathbf{v}, \mathbf{T})$, where i is the number of iteration, $\eta_0 = 0.1$, and $a(\bar{\mathbf{W}}, \mathbf{W}, \mathbf{v}, \mathbf{T}) = ||\bar{\mathbf{W}}|| + ||\mathbf{W}|| + ||\mathbf{v}|| + ||\mathbf{T}||$.

B. The Overall PDD Algorithm and Convergence Analysis

The PDD algorithm for solving Problem (30) is described in Algorithm 2. As described in Algorithm 2, the PDD method is a double loop algorithm, where in the inner loop the optimization problem is solved via the BSUM method and in the outer loop the dual Lagrangian variables and the penalty parameter are updated. A detailed convergence analysis for the PDD algorithm using the BSUM method is provided in [31], thus to avoid redundancy we only refer to [31] regarding the convergence of this algorithm. Specifically, it is shown that under appropriate conditions, the sequence generated by the PDD method tends to a KKT point of the main problem. In Algorithm

Algorithm 1 The BSUM method to optimize (30)

Repeat

1. Compute q_l^{opt}, r_k^{opt}, c_l^{opt} and d_k^{opt} by (25)-(28)
2. Compute $\mathbf{W}$ by (31), (32)
3. Compute $\bar{\mathbf{W}}$ by (33)
4. Compute $\mathbf{T}$ via (34) and (35)
5. Calculate $\mathbf{A}$ and $\mathbf{n}$ by (37) and (38)
6. **Repeat**:

   For $k \in \{1, ..., M\}$: Compute θ_k by solving either (39) or (42)

   Until $\mathbf{v}$ converges
7. Recalculate $\hat{\mathbf{H}}$ with the updated $\mathbf{v}$

Until some termination criterion is met.

2, the update method for the dual variable $\mathbf{\Lambda}_w$ in the kth iteration is

$$\mathbf{\Lambda}_w(k) = \mathbf{\Lambda}_w(k-1) + \frac{1}{\gamma}(\mathbf{W}(k) - \bar{\mathbf{W}}(k)). \tag{46}$$

The other dual Lagrangian variable, $\mathbf{\Lambda}_h$, is updated in the same manner. Also, the penalty parameter in the kth iteration is updated by

$$\gamma(k) = \zeta\gamma(k-1), \tag{47}$$

where $\zeta \in (0, 1)$ is a decreasing parameter.

Algorithm 2 The overall PDD algorithm

Initialize $\gamma, \mathbf{\Lambda}_h, \mathbf{\Lambda}_w, \eta, \zeta$ $\mathbf{v}, \mathbf{W}$ such that all constraints are met

Compute $\hat{\mathbf{H}}$ based on (17), $\mathbf{T} = \mathbf{W}^H\hat{\mathbf{H}}^H, \bar{\mathbf{W}} = \mathbf{W}$.

Repeat

1. Optimize (30) via the BSUM method in Algorithm 1
2. If $\left\| \left[(\mathbf{W} - \bar{\mathbf{W}})^T, (\mathbf{T} - \mathbf{W}^H\hat{\mathbf{H}}^H)^T \right] \right\|_\infty \leq \eta$: Update $\mathbf{\Lambda}_h, \mathbf{\Lambda}_w$ as in (46)

   Else: Update γ by (47)

Until convergence criteria is met.

V. IRS-AIDED OMA

In this section we consider IRS-aided FDMA and TDMA communication systems to provide a comparison benchmark for IRS-aided NOMA. Here, we consider a system containing U single-antenna users. The IRS-user, AP-user, and AP-IRS channels are denoted by $\mathbf{h}_{IU}^{[i]}$, $\mathbf{h}_{AU}^{[i]}$, and $\mathbf{G}_{AI}$, respectively, where $i \in \{1, ..., U\}$, and the effective channel between the ith user and the AP is represented by $\mathbf{h}^{[i]} = \mathbf{h}_{AU}^{[i]} + \mathbf{v}\,\mathrm{diag}(\mathbf{h}_{IU}^{[i]})\mathbf{G}_{AI}$. The estimated effective channel is $\hat{\mathbf{h}}^{[i]} = \hat{\mathbf{h}}_{AU}^{[i]} + \mathbf{v}\hat{\mathbf{H}}_c^{[i]}$, where $\hat{\mathbf{H}}_c^{[i]} = \mathrm{diag}(\hat{\mathbf{h}}_{IU}^{[i]})\mathbf{G}_{AI}$ represents the cascaded channel, and $\hat{\mathbf{h}}_{AU}^{[i]}$ and $\hat{\mathbf{h}}_{IU}^{[i]}$ are imperfect estimations for $\mathbf{h}_{AU}^{[i]}$ and $\mathbf{h}_{IU}^{[i]}$, respectively. The channels and estimation errors are generated similar to the IRS-NOMA scenario. Also, similar to the IRS-NOMA scenario, it is assumed that the estimation error of the acquired effective channel follows the distribution $\mathcal{CN}(\mathbf{0}, \sigma_h^2\mathbf{I})$ where σ_h^2 is calculated by (11). In the following, the FDMA and TDMA cases are considered and a robust PDD-based solution is given that maximizes their sum rates. The overall PDD algorithm for the FDMA and TDMA modes are similar to that of the NOMA mode presented in Algorithm 2. The steps of the BSUM method are different, and they are presented in the following sub-section with the revised formulas and parameters.

A. FDMA Mode

In the FDMA mode, the AP communicates with U users over U adjacent frequency resource blocks. In this case, the achievable sum rate is calculated by

$$R_{\text{FDMA}}^{[sum]} = \frac{1}{U}\sum_{k=1}^{U}\log_2\Big(1 + \frac{|\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}|^2}{\frac{1}{U}\sigma_h^2||\mathbf{w}^{[k]}||^2 + \frac{1}{U}\sigma_n^2}\Big), \tag{48}$$

where $\mathbf{w}^{[k]}$ is the active beamforming vector for the kth user. The proof can be readily derived with a similar approach to that of Theorem 1 in [10]. Assuming $\mathbf{W} = [\mathbf{w}^{[1]}, ..., \mathbf{w}^{[U]}]$, the resultant optimization problem of maximizing the sum rate is written as

$$\begin{aligned} \max_{\mathbf{W},\mathbf{v}} \quad & R_{\text{FDMA}}^{[sum]} \\ \text{s.t.} \quad & |\theta_i| = 1,\ i \in \{1, ..., M\}, \\ & \mathrm{trace}\left(\mathbf{W}\mathbf{W}^H\right) \le P. \end{aligned} \tag{49}$$

To solve this optimization problem, we employ the same PDD-based approach as we conducted for IRS-NOMA. We introduce a set of auxiliary variables $\mathcal{X} = \{\mathbf{W}, \bar{\mathbf{W}}, \mathbf{v}\}$, such that $\bar{\mathbf{W}} = \mathbf{W}$.

We now obtain the following augmented Lagrangian problem

$$\begin{aligned} \max_{\mathcal{X}} \quad & \sum_{k=1}^{U} \log_2\left(1+\frac{|\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}|^2}{\frac{1}{U}\sigma_h^2||\mathbf{w}^{[k]}||^2+\frac{1}{U}\sigma_n^2}\right) - Q_\gamma^{\text{FDMA}}(\mathcal{X}) \\ \text{s.t.} \quad & |\theta_i|=1,\ i\in\{1,...,M\}, \\ & \text{trace}\left(\bar{\mathbf{W}}\bar{\mathbf{W}}^H\right)\leq P, \end{aligned} \tag{50}$$

where $Q_\gamma^{\text{FDMA}}(\mathcal{X}) = \frac{1}{2\gamma}\sum_{i=1}^{U}||\mathbf{w}^{[i]}-\bar{\mathbf{w}}^{[i]}+\gamma\boldsymbol{\lambda}_{w_i}||^2$. To solve the problem using BSUM, we employ the same procedure as in Theorem 2 and [39], to find a tractable locally tight lower bound for the achievable sum rate. This lower bound is given by

$$\log_2\left(1+\frac{|\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}|^2}{\frac{1}{U}\sigma_h^2||\mathbf{w}^{[k]}||^2+\frac{1}{U}\sigma_n^2}\right) = \max_{c_k,q_k}\log_2(c_k) - c_k f(q_k), \tag{51}$$

where $f(q_k) = |1-q_k^*\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}|^2+\frac{\sigma_h^2}{U}||q_k^*\mathbf{w}^{[k]}||^2+\frac{\sigma_n^2}{U}|q_k|^2$. The optimal values of c_k and q_k are

$$q_k^{opt}(\mathcal{X}) = \frac{\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}}{|\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}|^2+\frac{1}{U}\sigma_h^2||\mathbf{w}^{[k]}||^2+\frac{1}{U}\sigma_n^2}, \tag{52}$$

$$c_k^{opt}(\mathcal{X}) = 1+\frac{|\hat{\mathbf{h}}^{[k]}\mathbf{w}^{[k]}|^2}{\frac{1}{U}\sigma_h^2||\mathbf{w}^{[k]}||^2+\frac{1}{U}\sigma_n^2}. \tag{53}$$

The optimization problem can now be rewritten as

$$\begin{aligned} \min_{\mathcal{X}} \quad & \sum_{k=1}^{U} c_k f(q_k) + Q_\gamma^{\text{FDMA}}(\mathcal{X}) \\ \text{s.t.} \quad & |\theta_i|=1,\ i\in\{1,...,M\}, \\ & \text{trace}\left(\bar{\mathbf{W}}\bar{\mathbf{W}}^H\right)\leq P. \end{aligned} \tag{54}$$

Similar to the PDD-based algorithm we used for IRS-NOMA, by using the first-order optimality condition, this problem can be solved in three steps w.r.t each of the variables individually. Then through an iterative algorithm similar to Algorithm 2, the problem is solved.

1) Step 1: Solving (54) for $\mathbf{W}$: The solution to this optimization problem w.r.t $\mathbf{w}^{[k]}$, $k\in\{1,...,U\}$, is given by

$$\mathbf{w}^{[k]} = \left(2\gamma c_k|q_k|^2\hat{\mathbf{h}}^{[k]H}\hat{\mathbf{h}}^{[k]}+\frac{1}{U}2\gamma c_k|q_k|^2\sigma_h^2\mathbf{I}+\mathbf{I}\right)^{-1}\times\left(\bar{\mathbf{w}}^{[k]}+2\gamma c_k q_k\hat{\mathbf{h}}^{[k]H}-\gamma\lambda_{w_k}\right). \tag{55}$$

2) Step 2: Solving (54) for $\bar{\mathbf{W}}$: The optimization problem w.r.t $\bar{\mathbf{W}}$ is equivalent to the projection of a point onto a ball centred at the origin, and it is solved similar to (33).

3) Step 3: Solving (54) for $\mathbf{v}$: The solution to the problem w.r.t $\mathbf{v}$ is exactly the same as what it was in IRS-NOMA, with $\mathbf{A}$ and $\mathbf{n}$ replaced by $\bar{\mathbf{A}}$ and $\bar{\mathbf{n}}$, respectively, defined by

$$\bar{\mathbf{A}} = \sum_{k=1}^{U} c_k|q_k|^2\hat{\mathbf{H}}_c^{[k]}\mathbf{w}^{[k]}\mathbf{w}^{[k]H}\hat{\mathbf{H}}_c^{[k]H}, \tag{56}$$

$$\bar{\mathbf{n}} = \sum_{k=1}^{U} c_k q_k \mathbf{w}^{[k]H} \hat{\mathbf{H}}_c^{[k]H} - c_k |q_k|^2 \hat{\mathbf{h}}_{AU}^{[k]} \mathbf{w}^{[k]} \mathbf{w}^{[k]H} \hat{\mathbf{H}}_c^{[k]H}. \tag{57}$$

To be more specific, in the case of continuous IRS, the solution is given by (40) and in the case of discrete IRS, the solution is driven by (43), depending on the resolution of the IRS, with the new definitions of $\bar{\mathbf{A}}$ and $\bar{\mathbf{n}}$ in (56) and (57), respectively.

B. TDMA Mode

In the TDMA mode, the communication between the AP and the users occurs in adjacent time-domain resource blocks. In this case, the achievable sum rate is obtained by

$$R_{\text{TDMA}}^{[sum]} = \frac{1}{U} \sum_{k=1}^{U} \log_2 \Big(1 + \frac{|\hat{\mathbf{h}}^{[k]} \mathbf{w}_k|^2}{\frac{1}{U}\sigma_h^2 ||\mathbf{w}_k||^2 + \frac{1}{U}\sigma_n^2} \Big). \tag{58}$$

Hence, the optimization problem becomes

$$\begin{aligned} \max_{\mathbf{W},\mathbf{v}} \quad & R_{\text{TDMA}}^{[sum]} \\ \text{s.t.} \quad & |\theta_i| = 1, \ i \in \{1, ..., M\}, \\ & \text{trace}\left(\mathbf{W}\mathbf{W}^H\right) \leq P. \end{aligned} \tag{59}$$

Solving this optimization problem is similar to solving (49). We use the PDD-based scheme to solve the optimization problem in an iterative algorithm, where in each iteration closed-form solutions are given for each of the variables. To this end, the optimization problem is rewritten based on the auxiliary variables in $\mathcal{X}$ as an augmented Lagrangian problem, and then the sum rate is replaced by its tight lower bound, which results in the tractable optimization problem in (54). The solution to the optimization problem in (54) in the TDMA mode, for the variables $\mathbf{W}$ and $\bar{\mathbf{W}}$ are given by (55) and (33), respectively. However, the solution of (54) for $\mathbf{v}$ is not the same as FDMA or NOMA modes. Note that, unlike NOMA and FDMA, in TDMA the IRS phase shifts for each user can be optimized separately over different time slots. The IRS passive beamforming can be time-selective, however, it cannot be frequency-selective, thus in TDMA, we have $\hat{\mathbf{h}}^{[k]} = \hat{\mathbf{h}}_{AU}^{[k]} + \mathbf{v}^{[k]} \hat{\mathbf{H}}_c^{[k]}$, in which $\mathbf{v}^{[k]}$ denotes the value of $\mathbf{v}$ in the k-th time slot, where user k is being served. With this fundamental difference in mind, the following optimization problem should be solved for each $k \in \{1, ..., U\}$

$$\begin{aligned} \min_{\mathbf{v}^{[k]}} \quad & \mathbf{v}^{[k]} \check{\mathbf{A}}^{[k]} \mathbf{v}^{[k]H} - 2\,\text{Re}\{\check{\mathbf{n}} \mathbf{v}^{[k]H}\} \\ \text{s.t.} \quad & |\theta_i^{[k]}| = 1, \ i \in \{1, ..., M\}, \end{aligned} \tag{60}$$

where,

$$\check{\mathbf{A}}^{[k]} = c_k |q_k|^2 \hat{\mathbf{H}}_c^{[k]} \mathbf{w}^{[k]} \mathbf{w}^{[k]H} \hat{\mathbf{H}}_c^{[k]H}, \tag{61}$$

$$\check{\mathbf{n}}^{[k]} = c_k q_k \mathbf{w}^{[k]H} \hat{\mathbf{H}}_c^{[k]H} - c_k |q_k|^2 \hat{\mathbf{h}}_{AU}^{[k]} \mathbf{w}^{[k]} \mathbf{w}^{[k]H} \hat{\mathbf{H}}_c^{[k]H}. \tag{62}$$

To solve the equivalent problem (60), we can follow the same way as in (36). In this case, for each time slot k and for one specific phase shift $\theta_j^{[k]}$, The solution would be obtained as

$$\theta_j^{[k]} = \frac{\check{e}_j^{[k]}}{\left|\check{e}_k^{[k]}\right|}, \tag{63}$$

where $\check{e}_j^{[k]} = \check{n}_j^{[k]} - \sum_{i \neq j}^{M} \theta_i \check{a}_{ij}^{[k]}$. In the case of a discrete IRS, the solution would be calculated with the same approach as in (43).

VI. Complexity Analysis

In this section, the complexity of the proposed method is calculated and it is compared with other techniques in the literature. The proposed scheme in Algorithm 2 is iterative, and each iteration contains Algorithm 1. As the number of outer iterations increases, the accuracy required for the termination of Algorithm 1 is achieved in fewer inner loop iterations. In the final iterations of the outer loop, this accuracy is very close, and the inner loop ends in only one iteration. Thus, on average, the number of inner loop iterations to achieve the required accuracy is very low and does not affect the order of the computational complexity. In each outer iteration, the computational complexity involves calculating the variables in steps 1 through 7 of Algorithm 1. This results in a complexity of $\mathcal{O}(U^2N + U(N^3 + (3.5 + N)M^2 + 13MN))$ for a total of $U = L + K$ users in the system. Additionally, either the dual Lagrangian variables or the penalty parameter need to be updated during this process. Note that in step 2 of Algorithm 2, the order of complexity is the highest when the dual variables are updated, i.e. $\mathcal{O}(2U^2N)$. Also, for updating $\mathbf{v}$, the maximum computational complexity is to calculate the values of $\mathbf{A}$ and $\mathbf{n}$, which is done only once in each iteration of BSUM (refer to step 5 and 6 of Algorithm 1). Considering a scenario where $M > N, U$, and assuming the BSUM technique terminates in less than 10 steps, the order of complexity for each iteration of Algorithm 2, for the IRS-aided NOMA, is

$$\mathcal{O}\left(3U^2N + U(N^3 + (3.5 + N)M^2 + 13MN)\right). \tag{64}$$

For the IRS-aided OMA, the order of complexity for calculating the dual variables is negligible, thus only the complexity of Algorithm 1 needs to be considered here. In this case, the order of complexity of each iteration of Algorithm 2 for IRS-OMA becomes

$$\mathcal{O}\left(U(N^3+(1+N)M^2+4MN)\right). \tag{65}$$

It is apparent that for large values of M, the computational complexity of IRS-NOMA and IRS-OMA are in the same order.

The computational complexity of the proposed PDD-based algorithm for IRS-aided NOMA is lower than that of the methods in the literature focusing on sum rate maximization in the downlink. In this section, we compare the complexity of our proposed algorithm to those presented in [16], [18]. In [16], an alternating optimization technique was used to find the optimal active AP beamforming vector and the passive IRS phase shifts for the sum rate maximization problem. The authors used the interior-point solver in CVX to solve the resulting convex optimization problems in each iteration of their algorithm which scales up the order of complexity. For instance, the complexity of each iteration of the algorithm proposed in [16] is

$$\mathcal{O}(\max(N,3U(U-1)^4)\sqrt{N}\log(\frac{1}{\mu_c})+(3U^2+M)^{3.5}), \tag{66}$$

where μ_c is the accuracy defined in [16]. In [18], the authors presented an iterative alternating optimization algorithm, where each iteration included three steps of solving convex optimization problems using the CVX toolbox. The Computational complexity for each iteration of this algorithm is

$$\mathcal{O}((N^{4.5}+(M+1)^{4.5}+U^{3.5})\log\frac{1}{\mu}), \tag{67}$$

where μ represents the accuracy of CVX. Now, consider a simple case where $U=2$, $\mu_c=0.1$ and $\mu=0.1$. In this scenario, the complexity ratio of the PDD-based algorithm to the methods in [16] and [18] is shown in Fig. 2. As shown in Fig. 2 the computational complexity of our method is far less than the algorithms proposed in [16] and [18] for different system dimensions.

VII. Numerical Results

In this section, the performance of our proposed algorithm is evaluated in terms of the average sum rate of the system. The NOMA system model under study contains $L=3$ primary users and $K=3$ secondary users, which is equivalent to $U=6$ single-antenna users in the OMA scenario.

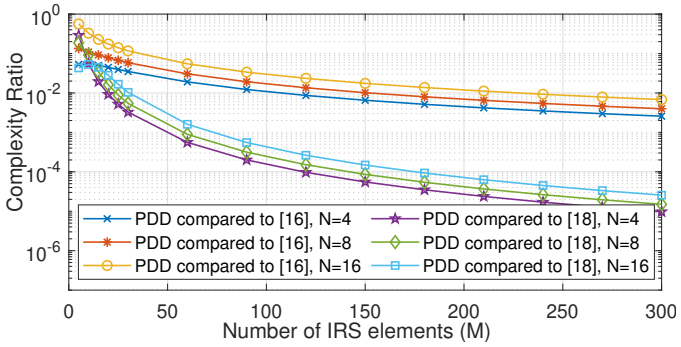
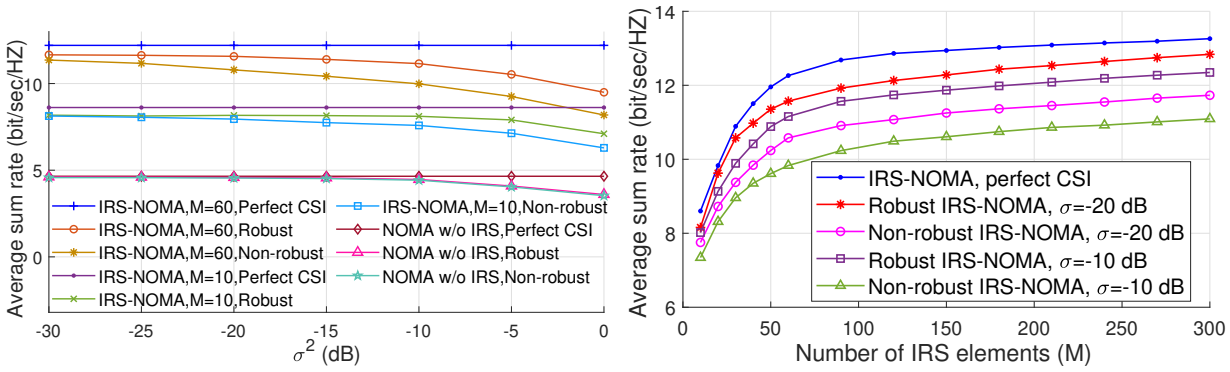


Figure 2: The complexity ratio of the PDD-based algorithm to the algorithm in [16], i.e. $\frac{(64)}{(66)}$, and the algorithm in [18], i.e. $\frac{(64)}{(67)}$.

An AP with $N = 4$ antennas communicates with the users and an IRS with M elements assists this communication. Without loss of generality, we assume that the AP is located at the point $[0, 0]$ at the center of $\mathcal{D}_0$, and the IRS is located at the point $[-0.25, 0.25]$. The radius of $\mathcal{D}_0$ and $\mathcal{D}_i$ are $r_0 = 1$ km and $r_i = 0.3$ km, respectively. The large-scale fading coefficients, β_{AU}, β_{IU} and β_{AI} are modeled by 3GPP standards in [40], where each large scale fading coefficient is calculated by $10\log_{10}(\beta) = -10 - 20\log_{10}(d) + z_{shad}$. In this model, d represents the distance between the two nodes in km and the parameter z_{shad} is a random variable with the distribution of $z_{shad} \sim \mathcal{CN}(0, \sigma^2_{shad})$ where $\sigma^2_{shad} = -22$ dB represents the shadowing. Unless stated otherwise, it is assumed that the noise variance at each user is $\sigma^2_n = -10$ dB, the transmission power of the AP is $P = 1$, and the estimation error variances are $\sigma^2 = \sigma^2_{AU} = \sigma^2_{IU} = -10$ dB. In the IRS-NOMA scenario, we utilized the policy presented in [41], where the users are paired based on their direct channels to AP. To be specific, for a NOMA system with $U = 2Z$ users, the optimal pairing policy is $u_{i,j} = \begin{cases} 1, & i + j = U + 1; \\ 0, & \text{Otherwise}, \end{cases}$, where $u_{i,j},\ i, j \in \{1, \ldots, U\}$ stands for a coefficient showing that the i-th user is paired with the j-th user if its value equals 1. Therefore, the overall pairing policy follows an appropriate condition for using NOMA by having a weak channel and a strong channel respectively for the edge user and the central user. The simulation parameters for Algorithm 2 are $\zeta = 0.95$, $\eta = 0.3$, $\gamma = 4.518$, and the dual Lagrangian variables are initialized by setting all of their values to 0.1. To initialize the variables $\mathbf{W}$ and $\mathbf{v}$ at the beginning of the algorithm (iteration zero), we set $\mathbf{W}^0 = 0.9\sqrt{\frac{P}{2UN}}(\mathbf{1} + j\mathbf{1})$ and $\mathbf{v}^0 = \mathbf{1}$, where $\mathbf{1}$ is an all-one matrix/vector with the appropriate size. In our simulations, "perfect CSI" refers

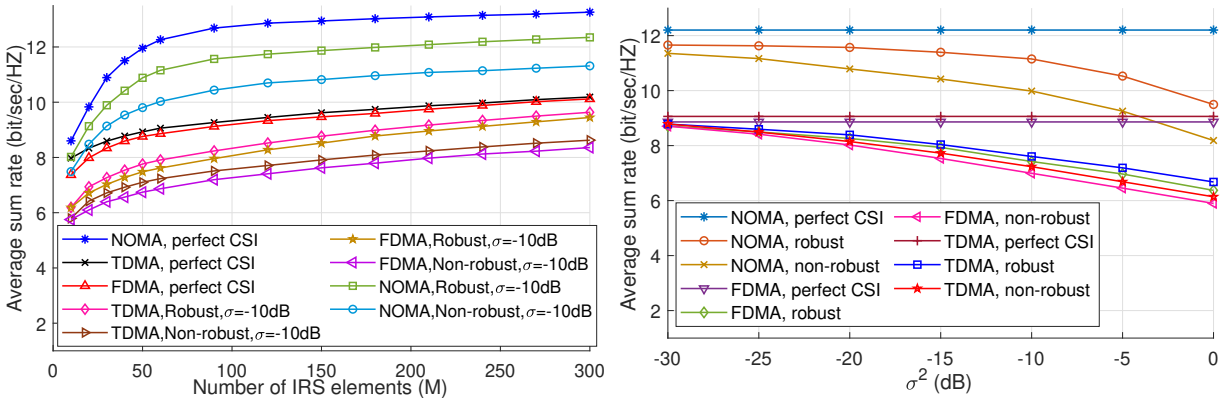


(a) Comparison between NOMA and IRS-NOMA. (b) IRS-NOMA in different levels of CSI uncertainty.

Figure 3: Evaluation of the IRS-NOMA system in terms of robustness to CSI uncertainty.

to the scenario where there is no CSI uncertainty ($\sigma^2 = \sigma_{AU}^2 = \sigma_{IU}^2 = 0$). "Robust" refers to a case where $\sigma^2 \neq 0$ and we use the beamforming design that is robust to the CSI uncertainty, which is presented in Algorithms 1 and 2. "Non-robust" refers to a case where $\sigma^2 \neq 0$, however, the AP is unaware of this CSI uncertainty. This can be realized by setting $\sigma_h^2 = 0$ in all of the formulations that are used in Algorithms 1 and 2. This inherently effects the output rate.

In Fig. 3 we evaluate the robustness of the proposed scheme to CSI uncertainty. Specifically, Fig. 3.(a) compares the IRS-aided NOMA system with the case where NOMA is used without the assistance of an IRS. In Fig. 3.(a), the average sum rate is demonstrated versus the variance of channel estimation error, for three NOMA systems, one without IRS, one with an IRS of size $M = 10$, and one with an IRS of size $M = 60$. It is shown that even with a small-scale IRS with only a few tens of antennas, the system performance is much better than when IRS is not employed. The advantages of IRS-NOMA are due to the capability of the IRS to modify the propagation environment such that the direction of users' channel vectors become more alike. Furthermore, the ability of the system to cancel out the channel estimation error is improved by the number of passive elements in the IRS, since the IRS creates more degrees of freedom for the system. In Fig. 3.(b) the sum rate is depicted versus the number of IRS elements (M) for three cases of robust design, non-robust design and perfect CSI. It is shown that the performance of the proposed robust method is better than the non-robust case. Note that, as M increases, the gap between the robust design and the non-robust structure widens as a consequence of using the approximation in (11) which is only accurate when M is large [10]. In other words, by



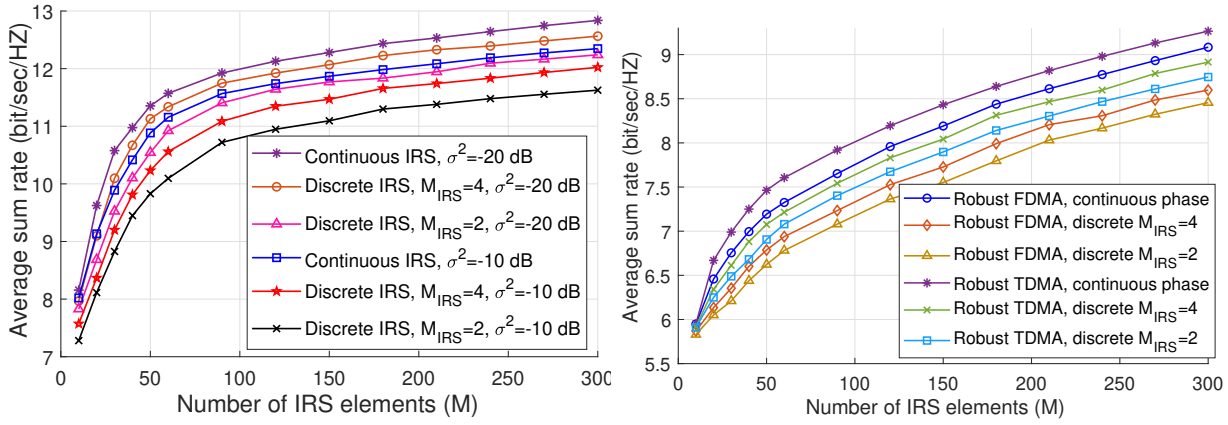
(a) Comparison in terms of Number of IRS elements.

(b) Comparison in terms of CSI uncertainty.

Figure 4: Comparison among IRS-aided FDMA, TDMA and NOMA.

increasing M, the system becomes more robust. Through simulation, we verified the validity of the approximation in (11) for $M > 50$, where the amount of uncertainty in this approximation drops to less than 2% of σ_h^2.

Fig. 4 presents a comparison between IRS-assisted NOMA and OMA systems, i.e. TDMA and FDMA, in three cases of perfect CSI, imperfect CSI with robust design, and imperfect CSI with non-robust design. In Fig. 4.(a), the average sum rate is shown versus the number of IRS elements, for $\sigma^2 = -10$ dB. Here, it is demonstrated how IRS-aided NOMA outperforms OMA in terms of average sum rate. It is noteworthy that, in case non-overlapping resource blocks are dedicated to users, the throughput of NOMA should be theoretically twice that of OMA thanks to employing twice as many resource blocks as OMA, and IRS is a beneficial tool for realizing such a situation. However, in practice, there would be some residual co-channel interference due to non-ideal similarity in the directions of users' channels. It is also shown that TDMA has a superior performance than FDMA. This is driven by the IRS's time selectivity through TDMA, i.e. in TDMA the IRS phase shift matrix is adjusted for each user separately, while in FDMA the phase shift matrix should be in charge of optimizing both users' propagation environments at the same time, making the IRS less flexible to the channel variation [13]. In other words, FDMA is not able to fully leverage the degree of freedom rendered by the IRS. Fig. 4.(b) depicts the system average sum rate versus the variance of the channel estimation error (σ^2). It is assumed here that an IRS with $M = 60$ elements aids either of the NOMA, TDMA or FDMA communications. As shown in Fig. 4.(b), in the perfect CSI scenario, IRS-NOMA always



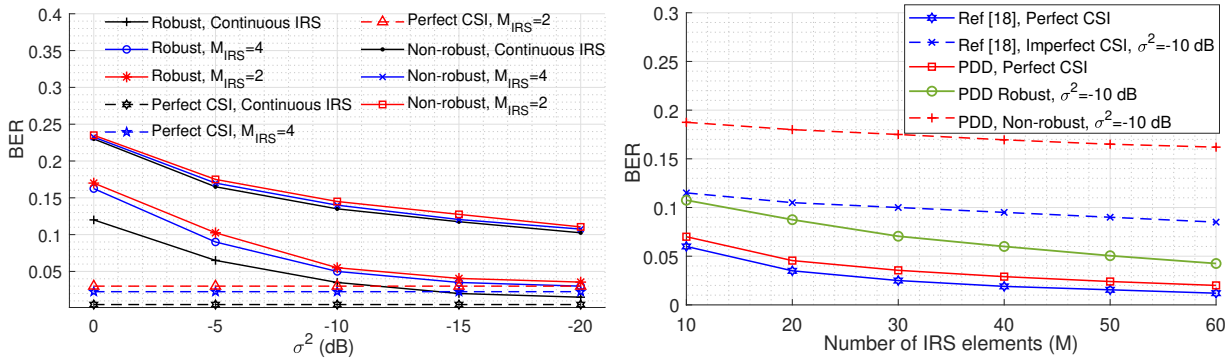
(a) IRS-aided NOMA.

(b) IRS-aided OMA.

Figure 5: Comparison between the two cases of continuous and discrete IRS.

renders superior performance compared to IRS-OMA due to channel reconfiguration abilities of the IRS. However, by increasing the channel estimation error, the performance of the non-robust IRS-NOMA drops quickly until it becomes fairly close to the performance of IRS-TDMA in larger values of σ^2. This limitation stems from the fact that for a high CSI uncertainty level, the system operates in a noise-dominant environment, and the promising features of the IRS can hardly extend to the case with severe CSI uncertainty. As shown in Fig. 4.(b), the robust IRS-NOMA design can compensate for the performance losses due to the CSI uncertainty. Thus, Fig. 4.(b) proves that to benefit from the advantages of the IRS-NOMA, having a beamforming design robust to the channel estimation error is crucial.

Fig. 5 demonstrates the effectiveness of the proposed robust PDD-based methods for a discrete IRS with phase resolutions of $\Delta \in \{2, 4\}$ by comparing their performances to those of a continuous IRS. In Fig 5.(a), the average sum rate of the robust IRS-NOMA scheme is shown versus M for different levels of CSI uncertainty. The performance of the discrete method where $\Delta = 4$ is very close to that of the continuous IRS-NOMA. In case $\Delta = 2$, there is a performance loss in the system due to the smaller resolution of the IRS elements, yet this performance loss is negligible. In summary, the proposed method using a discrete IRS can achieve a high performance in IRS-NOMA with a less complex structure. Fig. 5.(b) displays the average sum rate versus M for robust IRS-aided OMA under CSI uncertainty of $\sigma^2 = -10$ dB. It is revealed that the performance loss of the discrete IRS designs are negligible compared to the continuous IRS



(a) Continuous and discrete IRS-NOMA

(b) Comparison between PDD and the method in [18].

Figure 6: Performance of the system in terms of BER.

designs. Thus, by using a discrete IRS, we can reduce the hardware complexity while avoiding severe performance loss in IRS-OMA systems.

In Fig. 6, the performance of the IRS-aided NOMA system is studied in terms of bit error rate (BER) for an uncoded BPSK modulation. In Fig. 6.(a), we employ both continuous and discrete phase shifters with different resolutions, in an IRS with $M = 150$ elements. Fig. 6.(a) shows the significance of estimation error cancellation in an IRS-aided NOMA system. Note that in the IRS-NOMA system, the transmission power, i.e. $\mathrm{trace}(\mathbf{W}\mathbf{W}^H)$, depends on many system variables such as the value of CSI uncertainty σ_h^2. Thus, the SNR changes with system variables. The SNR for achieving the results in Fig. 6.(a) was in the range of $[0, 4.5]\ dB$. Fig. 6.(b) presents a performance comparison between the presented method in the Algorithm 2 and the method proposed in [18]. It is shown that in a perfect CSI scenario, the method presented in [18] achieves a better performance, but in the case of imperfect CSI, the proposed robust design can compensate for the CSI uncertainty up to an acceptable level. As discussed in Section VI, the order of computational complexity of the proposed method is far less than those presented in the literature. However, there is a trade-off between performance and complexity. To be specific, by choosing the accuracy threshold for Algorithm 1 as $\eta_0 = 0.1$, we can maintain low complexity at the cost of performance, but in case we set a better accuracy for the termination criterion, i.e. for any $\eta_0 < 0.1$, better results can be achieved at the cost of high computational complexity. Thus, in the presented PDD-based scheme, we have a trade-off between performance and complexity and based on our system requirements and limitations, we can choose the right value for η_0 to

achieve the desired result. Based on our simulations, by setting $\eta_0 = 0.1$, we can maintain a low complexity while the algorithm performs well.

VIII. Conclusion

This paper investigated the benefits of using IRS in a NOMA communication system. We jointly optimized the AP beamforming vector and the IRS entries under imperfect CSI, to maximize the sum rate. The PDD algorithm was applied to design a robust joint beamforming at the IRS and the AP. The proposed method was an iterative algorithm with closed-form solutions in each step, resulting in a low computational complexity. Two cases of continuous and discrete IRS were considered, and a low-complexity solution was proposed for the discrete phase shift selection. Numerical results demonstrated the efficiency of the robust design compared to the non-robust and the perfect CSI scenarios. In addition, the study compared NOMA to OMA in two scenarios involving IRS-assisted FDMA and TDMA systems, and concluded that the robust IRS-aided NOMA technique outperforms the robust IRS-OMA in terms of spectral efficiency.

Appendix A
Proof of Theorem 1

Let $\mathcal{I}\left(\mathbf{x}_{\{p,s\}}^{[i]}; y_{\{p,s\}}^{[i]} | \hat{\mathbf{h}}_{\{p,s\}}^{[i]}\right)$ be the conditional mutual information of primary user i conditioned on estimated channel vector $\hat{\mathbf{h}}_{\{p,s\}}^{[i]}$. Note that $\{p,s\}$ is replaced with p if user i is a primary user, and replaced with s if user i is a secondary user. Expanding $\mathcal{I}\left(\mathbf{x}_{\{p,s\}}^{[i]}; y_{\{p,s\}}^{[i]} | \hat{\mathbf{h}}_{\{p,s\}}^{[i]}\right)$ in terms of the differential entropies results in

$$\mathcal{I}\left(\mathbf{x}_{\{p,s\}}^{[i]}; y_{\{p,s\}}^{[i]} | \hat{\mathbf{h}}_{\{p,s\}}^{[i]}\right) = \mathcal{H}\left(\mathbf{x}_{\{p,s\}}^{[i]} | \hat{\mathbf{h}}_{\{p,s\}}^{[i]}\right) - \mathcal{H}\left(\mathbf{x}_{\{p,s\}}^{[i]} | y_{\{p,s\}}^{[i]}, \hat{\mathbf{h}}_{\{p,s\}}^{[i]}\right). \tag{68}$$

The first term on the right-hand side of (68) simplifies to $\log_2 \det\left(2\pi e \mathbf{V}_{\{p,s\}}^{[i]}\right)$, where $\mathbf{V}_{\{p,s\}}^{[i]} \triangleq \mathbb{E}\left\{\mathbf{x}_{\{p,s\}}^{[i]} \mathbf{x}_{\{p,s\}}^{[i]H}\right\}$ denotes the transmit co-variance matrix related to $\mathbf{x}_{\{p,s\}}^{[i]}$ [42]. The second term of the right-hand side of (68) is upper bounded by the entropy of a Gaussian random variable [43] as follows

$$\mathcal{H}(\mathbf{x}_{\{p,s\}}^{[i]} | y_{\{p,s\}}^{[i]}, \hat{\mathbf{h}}_{\{p,s\}}^{[i]}) \leq \log_2 \det\left(2\pi e (\mathbf{V}_{\{p,s\}}^{[i]} - \frac{\mathbf{V}_{\{p,s\}}^{[i]} \hat{\mathbf{h}}_{\{p,s\}}^{[i]H} \hat{\mathbf{h}}_{\{p,s\}}^{[i]} \mathbf{V}_{\{p,s\}}^{[i]}}{\hat{\mathbf{h}}_{\{p,s\}}^{[i]} \mathbf{V}_{\{p,s\}}^{[i]} \hat{\mathbf{h}}_{\{p,s\}}^{[i]H} + \Gamma_{\{p,s\}}^{[i]}})\right), \tag{69}$$

in which

$$\Gamma_p^{[l]} \triangleq \frac{1}{Z}\left(\sigma_h^2 \operatorname{trace}\left(\mathbf{V}_p^{[l]}\right) + \sigma_h^2 \operatorname{trace}\left(\mathbf{V}_s^{[\mathcal{U}(s,l,p)]}\right) + \sigma_n^2\right), \tag{70}$$

$$\Gamma_s^{[k]} \triangleq \hat{\mathbf{h}}_s^{[k]H}\mathbf{V}_p^{[\mathcal{U}(p,k,s)]}\hat{\mathbf{h}}_s^{[k]} + \frac{1}{Z}\left(\sigma_h^2 \operatorname{trace}\left(\mathbf{V}_p^{[\mathcal{U}(p,k,s)]}\right) + \sigma_h^2 \operatorname{trace}\left(\mathbf{V}_s^{[k]}\right) + \sigma_n^2\right). \tag{71}$$

Now, we employ the Woodbury matrix identity as follows

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B}\right)^{-1}\mathbf{DA}^{-1}. \tag{72}$$

Assuming $\mathbf{A}=\mathbf{I}$, $\mathbf{B}=\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}$, $\mathbf{C}=\Gamma_{\{p,s\}}^{[i]^{-1}}$ and $\mathbf{D}=\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}$, and using (72), it is concluded

$$\mathbf{I} - \frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}}{\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}+\Gamma_{\{p,s\}}^{[i]}} = \left(\mathbf{I}+\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}}{\Gamma_{\{p,s\}}^{[i]}}\right)^{-1}. \tag{73}$$

Thus, the right hand side of (69) can be rewritten as

$$\begin{aligned}
&\log_2\det\left(2\pi e\mathbf{V}_{\{p,s\}}^{[i]}\left(\mathbf{I}-\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}}{\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}+\Gamma_{\{p,s\}}^{[i]}}\right)\right)\\
&=\log_2\det\left(2\pi e\mathbf{V}_{\{p,s\}}^{[i]}\left(\mathbf{I}+\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}}{\Gamma_{\{p,s\}}^{[i]}}\right)^{-1}\right)\\
&=\log_2\det\left(2\pi e\mathbf{V}_{\{p,s\}}^{[i]}\right) - \log_2\det\left(\mathbf{I}+\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}}{\Gamma_{\{p,s\}}^{[i]}}\right).
\end{aligned} \tag{74}$$

Exploiting (74) and employing Sylvester's determinant theorem, i. e., $\det(\mathbf{I}+\mathbf{AB}) = \det(\mathbf{I}+\mathbf{BA})$, (69) is rewritten as

$$\begin{aligned}
\mathcal{H}\left(\mathbf{x}_{\{p,s\}}^{[i]}|y_{\{p,s\}}^{[i]},\hat{\mathbf{h}}\right) &\le \log_2\det\left(2\pi e\mathbf{V}_{\{p,s\}}^{[i]}\left(\mathbf{I}-\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}}{\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}+\Gamma_{\{p,s\}}^{[i]}}\right)\right)\\
&= \log_2\det\left(2\pi e\mathbf{V}_{\{p,s\}}^{[i]}\right) - \log_2\det\left(1+\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}}{\Gamma_{\{p,s\}}^{[i]}}\right).
\end{aligned} \tag{75}$$

Consequently, assuming $\mathbb{E}\{|s_{\{p,s\}}^{[i]}|^2\}=1$ and utilizing (75), (68) leads to

$$\mathcal{I}\left(\mathbf{x}_{\{p,s\}}^{[i]};y_{\{p,s\}}^{[i]}|\hat{\mathbf{h}}^{[i]}\right) \ge \log_2\det\left(1+\frac{\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{V}_{\{p,s\}}^{[i]}\hat{\mathbf{h}}_{\{p,s\}}^{[i]H}}{\Gamma_{\{p,s\}}^{[i]}}\right) = \log_2\left(1+\frac{\left|\hat{\mathbf{h}}_{\{p,s\}}^{[i]}\mathbf{w}_{i,\{p,s\}}\right|^2}{\Gamma_{\{p,s\}}^{[i]}}\right). \tag{76}$$

Thus, the minimum achievable rates of the l-th primary user and the k-th secondary user are

$$R_p^{[l]} = \log_2\left(1+\frac{|\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{l,p}|^2}{\sigma_h^2\left(||\mathbf{w}_{l,p}||^2+||\mathbf{w}_{\mathcal{U}(s,l,p),s}||^2\right)+\sigma_n^2}\right), \tag{77}$$

$$R_s^{[k]} = \log_2\left(1+\frac{|\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{k,s}|^2}{|\hat{\mathbf{h}}_s^{[k]}\mathbf{w}_{\mathcal{U}(p,k,s),p}|^2+\sigma_h^2\left(||\mathbf{w}_{\mathcal{U}(p,k,s),p}||^2+||\mathbf{w}_{k,s}||^2\right)+\sigma_n^2}\right). \tag{78}$$

APPENDIX B
FURTHER EXPLANATIONS ON THE CONSTRAINTS OF PROBLEM (18)

In this appendix, we provide further explanations on deriving the third and fourth constraints of (18) from the third constraint of (16). In (16), we have $\frac{|\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_p^{[l]}\mathbf{w}_{l,p}|^2+b_l} \geq \frac{|\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]}\mathbf{w}_{\mathcal{U}(s,l,p),s}|^2}{|\hat{\mathbf{h}}_s^{[\mathcal{U}(s,l,p)]}\mathbf{w}_{l,p}|^2+b_l}$, $l \in \{1, ..., L\}$. Based on our definitions of $\mathbf{T} = \mathbf{W}^H\hat{\mathbf{H}}^H$, $\mathbf{W} = [\mathbf{w}_{1,p}, \ldots, \mathbf{w}_{L,p}, \mathbf{w}_{1,s}, \ldots, \mathbf{w}_{K,s}]$ and $\hat{\mathbf{H}} \triangleq \left[\hat{\mathbf{h}}_p^{[1]\ T}, \ldots, \hat{\mathbf{h}}_p^{[L]\ T}, \hat{\mathbf{h}}_s^{[1]\ T}, \ldots, \hat{\mathbf{h}}_s^{[K]\ T}\right]^T$, this ratio can be rewritten by $\frac{|t_{L+\mathcal{U}(s,l,p),l}|^2}{|t_{l,l}|^2+b_l} \geq \frac{|t_{L+\mathcal{U}(s,l,p),L+\mathcal{U}(s,l,p)}|^2}{|t_{l,L+\mathcal{U}(s,l,p)}|^2+b_l}$. Now, this constraint can be replaced by the third and fourth constraints of (18), based on the following theorem.

Theorem 3. *Given the four non-negative real variables $a, b, c, d \geq 0$, if $a \geq c$, $a \geq b$ and $d \geq c$, for all non-negative real values of K we have $\frac{a}{b+K} \geq \frac{c}{d+K}$.*

Proof. It is apparent that if $a \geq c$ and $d \geq b$, the inequality $\frac{a}{b+K} \geq \frac{c}{d+K}$ holds. Thus, the proof is complete if $d \geq b$. In the case where $b \geq d$, only one option exists to arrange the variables, that is $0 \leq c \leq d \leq b \leq a$. In this case, suppose that the inequality does not hold. We have

$$\frac{a}{b+K} < \frac{c}{d+K} \rightarrow \ \frac{a(d+K)-c(b+k)}{(b+K)(d+K)} < 0 \rightarrow \ a(d+K)-c(b+K) < 0 \rightarrow$$
$$(ad-cb)+K(a-c) < 0 \xrightarrow{\{1\}} \ ad < cb \rightarrow \ \frac{a}{b} < \frac{c}{d}, \tag{79}$$

where $\{1\}$ results from the fact that $K(a-c) \geq 0$. Now, based on our original assumption of $0 \leq c \leq d \leq b \leq a$, we have $\frac{a}{b} \geq 1$ and $\frac{c}{d} \leq 1$. Hence, the resulting inequality in (79) is inaccurate which proves that the inequality $\frac{a}{b+K} \geq \frac{c}{d+K}$ holds. Thus, the proof is conducted. □

REFERENCES

[1] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the Application of Quasi-Degradation to MISO-NOMA Downlink," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6174–6189, 2016.

[2] Z. Ding and H. V. Poor, "A Simple Design of IRS-NOMA Transmission," *IEEE Communications Letters*, vol. 24, no. 5, pp. 1119–1123, 2020.

[3] C. Pan, H. Ren, K. Wang, J. F. Kolb, M. Elkashlan, M. Chen, M. Di Renzo, Y. Hao, J. Wang, A. L. Swindlehurst *et al.*, "Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 14–20, 2021.

[4] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO Communications Relying on Intelligent Reflecting Surfaces," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5218–5233, 2020.

[5] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent Reflecting Surface Aided MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1719–1734, 2020.

[6] Q. Zhang, W. Saad, and M. Bennis, "Reflections in the Sky: Millimeter Wave Communication with UAV-Carried Intelligent Reflectors," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[7] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and Secure Wireless Communications via Intelligent Reflecting Surfaces," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2637–2652, 2020.

[8] Q. Wu and R. Zhang, "Weighted Sum Power Maximization for Intelligent Reflecting Surface Aided SWIPT," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 586–590, 2019.

[9] J. Zuo, Y. Liu, E. Basar, and O. A. Dobre, "Intelligent Reflecting Surface Enhanced Millimeter-Wave NOMA Systems," *IEEE Communications Letters*, vol. 24, no. 11, pp. 2632–2636, 2020.

[10] Y. Omid, S. M. Shahabi, C. Pan, Y. Deng, and A. Nallanathan, "Low-Complexity Robust Beamforming Design for IRS-Aided MISO Systems with Imperfect Channels," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1697–1701, 2021.

[11] G. Zhou, C. Pan, H. Ren, K. Wang, M. Di Renzo, and A. Nallanathan, "Robust Beamforming Design for Intelligent Reflecting Surface Aided MISO Communication Systems," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1658–1662, 2020.

[12] J. Zhang, Y. Zhang, C. Zhong, and Z. Zhang, "Robust Design for Intelligent Reflecting Surfaces Assisted MISO Systems," *IEEE communications letters*, vol. 24, no. 10, pp. 2353–2357, 2020.

[13] B. Zheng, Q. Wu, and R. Zhang, "Intelligent Reflecting Surface-Assisted Multiple Access with User Pairing: NOMA or OMA?" *IEEE Communications Letters*, vol. 24, no. 4, pp. 753–757, 2020.

[14] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "Reconfigurable Intelligent Surface Aided NOMA Networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2575–2588, 2020.

[15] M. Fu, Y. Zhou, and Y. Shi, "Intelligent Reflecting Surface for Downlink Non-Orthogonal Multiple Access Networks," in *2019 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2019, pp. 1–6.

[16] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting Intelligent Reflecting Surfaces in NOMA Networks: Joint Beamforming Optimization," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6884–6898, 2020.

[17] G. Yang, X. Xu, Y.-C. Liang, and M. Di Renzo, "Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access," *IEEE Transactions on Wireless Communications*, vol. 20, no. 5, pp. 3137–3151, 2021.

[18] F. Xu and H. Zhang, "Energy Efficiency and Spectral Efficiency Tradeoff in IRS-Assisted Downlink MmWave NOMA Systems," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1433–1437, 2022.

[19] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum Rate Maximization for IRS-Assisted Uplink NOMA," *IEEE Communications Letters*, vol. 25, no. 1, pp. 234–238, 2020.

[20] X. Li, Z. Xie, G. Huang, J. Zhang, M. Zeng, and Z. Chu, "Sum Rate Maximization for RIS-Aided NOMA With Direct Links," *IEEE Networking Letters*, vol. 4, no. 2, pp. 55–58, 2022.

[21] Y. Cheng, K. H. Li, Y. Liu, K. C. Teh, and H. V. Poor, "Downlink and Uplink Intelligent Reflecting Surface Aided Networks: NOMA and OMA," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3988–4000, 2021.

[22] A. A. Salem, M. Rihan, L. Huang, and A. Benaya, "Intelligent Reflecting Surface Assisted Hybrid Access Vehicular Communication: NOMA or OMA Contributes the Most?" *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18 854–18 866, 2022.

[23] L. Lv, Q. Wu, Z. Li, Z. Ding, N. Al-Dhahir, and J. Chen, "Covert Communication in Intelligent Reflecting Surface-Assisted NOMA Systems: Design, Analysis, and Optimization," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1735–1750, 2021.

[24] X. Yue and Y. Liu, "Performance Analysis of Intelligent Reflecting Surface Assisted NOMA Networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 4, pp. 2623–2636, 2021.

[25] S. Li, L. Bariah, S. Muhaidat, A. Wang, and J. Liang, "Outage Analysis of NOMA-Enabled Backscatter Communications with Intelligent Reflecting Surfaces," *IEEE Internet of Things Journal*, 2022.

[26] Z. Zhang, L. Lv, Q. Wu, H. Deng, and J. Chen, "Robust and Secure Communications in Intelligent Reflecting Surface Assisted NOMA Networks," *IEEE Communications Letters*, vol. 25, no. 3, pp. 739–743, 2020.

[27] Z. Ding, L. Lv, F. Fang, O. A. Dobre, G. K. Karagiannidis, N. Al-Dhahir, R. Schober, and H. V. Poor, "A State-of-the-Art Survey on Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access Networks," *Proceedings of the IEEE*, 2022.

[28] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Matrix-Calibration-Based Cascaded Channel Estimation for Reconfigurable Intelligent Surface Assisted Multiuser MIMO," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2621–2636, 2020.

[29] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A Framework of Robust Transmission Design for IRS-Aided MISO Communications with Imperfect Cascaded Channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5092–5106, 2020.

[30] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[31] Q. Shi and M. Hong, "Penalty Dual Decomposition Method for Nonsmooth Nonconvex Optimization—Part I: Algorithms and Convergence Analysis," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4108–4122, 2020.

[32] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[33] X. Yu, D. Xu, and R. Schober, "Enabling Secure Wireless Communications via Intelligent Reflecting Surfaces," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[34] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, 2019.

[35] Z. Abdullah, G. Chen, S. Lambotharan, and J. A. Chambers, "Optimization of Intelligent Reflecting Surface Assisted Full-Duplex Relay Networks," *IEEE Wireless Communications Letters*, vol. 10, no. 2, pp. 363–367, 2020.

[36] ——, "A Hybrid Relay and Intelligent Reflecting Surface Network and its Ergodic Performance Analysis," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1653–1657, 2020.

[37] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "MIMO Assisted Networks Relying on Large Intelligent Surfaces: A Stochastic Geometry Model," *arXiv preprint arXiv:1910.00959*, vol. 1, 2019.

[38] W. Liang, Z. Ding, Y. Li, and L. Song, "User Pairing for Downlink Non-Orthogonal Multiple Access Networks Using Matching Algorithm," *IEEE Transactions on communications*, vol. 65, no. 12, pp. 5319–5332, 2017.

[39] Q. Shi, M. Hong, X. Fu, and T.-H. Chang, "Penalty Dual Decomposition Method for Nonsmooth Nonconvex Optimization—Part II: Applications," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4242–4257, 2020.

[40] 3GPP, "Technical Specification Group Radio Access Network; 3GPP TR 25.996, Spatial Channel Model for MIMO Simulations," vol. 1, no. Release 10, 2011.

[41] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal User Pairing for Downlink Non-Orthogonal Multiple Access (NOMA)," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 328–331, 2018.

[42] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications. New York, USA: John Wiley & Sons, Inc., 1991.

[43] M. Medard, "The Effect Upon Channel Capacity in Wireless Communications of Perfect and Imperfect Knowledge of the Channel," *IEEE Transactions on Information theory*, vol. 46, no. 3, pp. 933–946, 2000.