IEEE ComSoc IEEE Open Journal of the Communications Society

Geometric Mean Rate Maximization in RIS-aided mmWave ISAC Systems Relying on a Non-Diagonal Phase Shift Matrix

JITENDRA SINGH[1] (Graduate Student Member, IEEE), ADITYA K. JAGANNATHAM[1] (Senior Member, IEEE), AND LAJOS HANZO[2] (Life Fellow, IEEE)

[1]Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India
[2]School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

CORRESPONDING AUTHOR: LAJOS HANZO (e-mail: lh@ecs.soton.ac.uk).

The work of Aditya K. Jagannatham was supported in part by the Qualcomm Innovation Fellowship; in part by the Qualcomm 6G UR Gift; in part by the Arun Kumar Chair Professorship; and in part by the DST, Govt. of India and –UKRI EPSRC Project Intelligent Spectrum Innovation ICON. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council (EPSRC) projects under grant EP/Y026721/1, EP/W032635/1, EP/Y037243/1 and EP/X04047X/1 as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028).

ABSTRACT The joint optimization of the hybrid transmit precoders (HTPCs) and reflective elements of a millimeter wave (mmWave) integrated sensing and communication (ISAC) system is considered. The system also incorporates a reconfigurable intelligent surface (RIS) relying on a non-diagonal RIS (NDRIS) phase shift matrix. Specifically, we consider a hybrid architecture at the ISAC base station (BS) that serves multiple downlink communication users (CUs) via the reflected links from the RIS, while concurrently detecting multiple radar targets (RTs). We formulate an optimization problem that aims for maximizing the geometric mean (GM) rate of the CUs, subject to the sensing requirement for each RT. Additional specifications related to the limited transmit power and unit modulus (UM) constraints for both the HTPCs and the reflective elements of the NDRIS phase shift matrix make the problem challenging. To solve this problem, we first transform the intractable GM rate expression to a tractable weighted sum rate objective and next split the transformed problem into sub-problems. Consequently, we propose an iterative alternating optimization approach that leverages the majorization-minimization (MM) framework and block coordinate descent (BCD) method to solve each sub-problem. Furthermore, to tackle the UM constraints in the sub-problem of the HTPC design, we propose a penalty-based Riemannian manifold optimization (PRMO) algorithm, which optimizes the HTPCs on the Riemannian manifold. Similarly, the phases of the reflective elements of the NDRIS are optimized by employing the Riemannian manifold, and the locations of the non-zero entries of the NDRIS phase shift matrix are obtained by the maximal ratio combining (MRC) criterion. Finally, we present our simulation results, which show that deploying an NDRIS achieves additional gains for the CUs over a conventional RIS, further enhancing both the communication efficiency and sensing reliability. Furthermore, we compare the results to the pertinent benchmarks, which validate the effectiveness of our proposed algorithms.

INDEX TERMS Reconfigurable intelligent surface, integrated sensing and communication, millimeter wave, and geometric mean rate.

I. INTRODUCTION

INtegrated sensing and communication (ISAC) is an innovative technology that seeks to merge the traditionally separate domains of sensing and communication into a unified framework. This integration promises numerous advantages, including improved efficiency, enhanced performance, and novel capabilities in various applications such as autonomous vehicles, smart cities, unmanned aerial vehicles, healthcare, industrial automation, and a plethora of others [1], [2]. Thus, ISAC technology is envisaged to play a crucial role in the evolution of next-generation (NG) wireless networks, which jointly optimize both sensing and

communications functionalities in an integrated hardware platform, thereby reducing the hardware cost and enhancing the overall system efficiency and performance [3], [4].

Recently, several compelling advantages have emerged with the deployment of millimeter wave (mmWave) technology in ISAC systems. At the forefront of this development is the use of a wide bandwidth facilitated by the vast mmWave range spanning from 30 GHz to 300 GHz, which is capable of supporting ultra-high data rates [5] for the communication users (CUs). Another advantage of mmWave technology is its ability to facilitate precise localization and sensing capabilities due to its short wavelengths, enabling accurate radar target (RT) detection and tracking in complex environments. However, the conventional fully digital transmit precoder (TPC) designs are not suitable for mmWave systems due to their exclusive hardware costs and increased power consumption, especially when deployed in large-scale array systems [6]. To overcome this impediment, the well-known hybrid transmit precoder (HTPC) is employed in mmWave systems in which the TPC at the ISAC base station (BS) is split into the baseband (BB) TPC and digitally controllable RF TPC. Due to this, the HPTC significantly reduces the number of RF chains required, resulting in moderate hardware complexity and cost compared to conventional systems [7], [8].

Despite the above advantages, mmWave-aided systems face several challenges, such as their high pathloss and susceptibility to blockages by physical obstacles. To handle these issues, reconfigurable intelligent surfaces (RIS) have been investigated, which offer an additional layer of enhancement for mmWave systems by providing a favorable propagation environment and additional reflecting links [9]–[11]. Specifically, an RIS comprises passive reflective elements having adjustable reflection coefficients, which modify the phase shifts of the incident signal to optimize the signal propagation towards the intended CUs. In the conventional diagonal RIS (DRIS), the incident wave impinges on a specific element of the RIS and it is reflected from the same element, resulting in a diagonal phase shift matrix. However, while conventional DRIS enhances the signal of the intended CUs, it does not reach its full potential due to passive beamforming based on the equal gain criterion (EGC) [12]. To achieve a higher passive beamforming gain based on the optimal maximal ratio combining (MRC) criterion, an enhanced version of the RIS, namely a non-diagonal RIS (NDRIS), was proposed in [12]–[14]. In the NDRIS design, the incident wave impinges on an element and gets reflected from another element via a configuration of switches, resulting in a non-diagonal phase shift matrix. In doing so, one can achieve a higher beamforming gain by suitably optimizing the phases and locations of the non-zero RIS elements [14].

Furthermore, rate fairness among the CUs in mmWave ISAC systems is a serious problem due to the sensing requirements of the RTs and higher channel attenuation in the mmWave band. To achieve this, one can employ the max-min rate (MMR) objective, as discussed in [15]–[17] for maximizing the rate of the CU having poor channel conditions. However, this leads to an undesirable degradation in the overall sum rate (SR) of the system. As a remedy, the geometric mean (GM) rate maximization [18]–[20] strikes a compelling trade-off between excellent rate fairness and high sum rate. Therefore, we harness GM rate maximization for achieving rate fairness in mmWave ISAC systems.

Inspired by the above developments, we consider an RIS-assisted mmWave ISAC system relying on the NDRIS architecture that jointly optimizes the HPTCs and the reflection coefficients of the NDRIS phase shift matrix for maximizing the GM rate of CUs, while meeting the sensing requirements of the RTs. The next subsection presents a comprehensive literature survey on ISAC, mmWave technology, RIS, GM rate optimization, and their integration in modern wireless communications.

A. Literature review

The recent literature [5], [6], [21]–[27] investigated potent HTPC design techniques conceived for mmWave ISAC systems. Specifically, Xiao *et al.* [22] proposed a cost-effective RF TPC-based beamforming design, which is inspired by the idea of the multiple-beam sweeping method for serving the CU and RT in a mmWave ISAC system. Qi *et al.* [23] minimized the error between the radar beampattern and the desired transmit beamformer, incorporating SINR constraints for each CU and total power constraints to design HTPCs for mmWave ISAC systems. Gong *et al.* [24] optimized the HTPCs by maximizing the weighted sum rate of the CUs, while considering the Cramér-Rao bound (CRB) for the estimation of the RT angle and power constraint. The authors therein proposed the Riemannian conjugate gradient (RCG) based framework to address the unit modulus (UM) constraint on each element of the RF TPC. Wang *et al.* [25] considered a partially connected hybrid architecture at the ISAC base station (BS) and proposed a triple-stage alternating minimization technique for designing the HPTCs, which accomplishes the tasks of joint sensing and communication. As a further advance, the authors of [26], [27] conceived HPTC design for the mmWave ISAC-aided Internet of Vehicles (IoVs). Furthermore, Chowdary *et al.* [28] proposed a hybrid radar fusion framework for ISAC, where a DFRC base station integrates the downlink and uplink signals to jointly estimate target angle of arrivals (AoAs) using efficient maximum likelihood-based algorithms.

To overcome signal blockages in mmWave-aided systems, the authors of [9]–[11], [15], [29], [30] explored the benefit of RISs in mmWave communications, highlighting the benefits of deploying an RIS for enhancing the hybrid beamforming gain. Specifically, Gong *et al.* [9] considered a RIS-aided mmWave system, wherein they advocated minimizing the mean squared error (MSE) of the transmitted and received signal for jointly optimizing the HPTCs and

reflective elements at the ISAC BS and RIS, respectively. To solve the problem, the authors employed the RCG as well as majorization and minimization (MM) techniques to tackle the non-convex UM constraints. Furthermore, Zargari *et al.* [11] proposed the joint active and passive beamforming deign for the RIS-aided simultaneous wireless information and power transfer (SWIPT) system to maximize energy efficiency (EE) through the MM and Dinkelbach algorithms. By contrast, the authors of [15] addressed a QoS problem aiming for minimizing the total transmit power at the BS, while ensuring that the individual SINR constraints of the users are met. In our previous work [29], we jointly optimized the HTPCs, receiver combiners (RCs), and RIS phase shift matrix for maximizing the sum rate of the secondary system in RIS-aided mmWave cognitive radio systems operating in the underlay mode. In addition, in [30], we maximized the EE of an RIS-aided mmWave system for the optimization of the HTPCs, RCs, and the RIS phase shift matrix. Furthermore, Li *et al.* [31]–[33] have investigated the impact of imperfect CSI and hardware impairments on RIS-aided systems. The authors revealed performance floors at high transmit power due to estimation errors and RIS phase noise. These effects must be carefully considered in practical ISAC system design to ensure robust performance.

It is worth noting that an important feature of the RIS discussed in [9], [10], [15], [29], [30] is the diagonal nature of the phase shift matrix. This type of RIS configuration, commonly referred to as a DRIS, operates without any cooperation among its elements. Furthermore, due to the underlying passive beamforming, the DRIS amalgamates the incident and reflected signals by harnessing the suboptimal EGC criterion. Therefore, a DRIS is unable to fully harness the potential advantages of RIS technology. To address this limitation, Shen *et al.* [34] proposed a more advanced fully/group-connected RIS architecture, wherein the RIS elements cooperate to create a favorable wave propagation environment. In these architectures, the fully-connected and group-connected configurations contain $\frac{M(M+1)}{2}$ and $\frac{M(G+1)}{2}$ non-zero entries, respectively, where M denotes the number of RIS elements and G represents the group size. Optimizing these configurations and feeding the results back to the RIS is essential for achieving optimal performance, but it comes at the cost of increased complexity and overhead. To mitigate this drawback, Li *et al.* [12] introduced a non-diagonal RIS (NDRIS) phase shift matrix. This design allows an wave incident on one RIS element to be reflected from any other RIS element, significantly enhancing the system performance. However, implementing an NDRIS necessitates additional switches within the RIS to enable inter-element routing. As discussed in [12], these switches rely on RF micro-electromechanical systems (MEMS), which are cost-effective and energy-efficient components suitable for practical communication systems [35]. For practical implementation of the NDRIS structure, including its additional power requirements and hardware complexity, motivated readers might like to consult [14], [35], [36]. In particular, [14] considered an NDRIS-aided mmWave system for jointly optimizing the HTPCs and phase shift matrix for the maximization of the minimum user rate, which vindicated our proposition that an NDRIS phase shift matrix can lead to a significantly enhanced performance over the DRIS phase shift matrix in terms of both spectral- and energy- efficiency. Moreover, the authors of [36] presented a transistor-based realization of nonreciprocal, non-gyrotropic phase gradient metasurfaces suitable for NDRIS design.

However, only a few studies [37], [38] have investigated the use of RISs in ISAC-aided mmWave systems for enhancing both the sensing and communication performance. Specifically, Lyu *et al.* [37] jointly optimized the HTPCs and RIS phase shift matrix to minimize the CRB for sensing the angle of the RT, while meeting the minimum rate requirement of each CU in the RIS-aided mmWave ISAC system. The authors reformulated the non-convex problem considered by using auxiliary variables and the Schur complement and subsequently proposed a Remanninan manifold-based algorithm to handle the non-convex UM constraints. As a further advance, Hao *et al.* [38] considered a pair of interesting problems for optimizing the joint active and passive beamforming in an RIS-aided mmWave ISAC system. The first one focused on maximizing the worst-case illumination power of the RTs, while maintaining the quality of services of the CUs. By contrast, the second one aimed for maximizing the rate of CUs, while maintaining the illumination power requirements of the RTs. Although Hao *et al.* [38] in their second problem maximized the achievable sum rate of the CUs, this approach leads to rate unfairness among the CUs, possibly resulting in near zero data rates, especially as the sensing performance improves. To this end, the GM rate [18]–[20] is a beneficial communication metric for RIS-aided mmWave ISAC systems, which achieves a rate fairness similar to that of the MMR optimization framework without significantly compromising the sum rate. It is worth noting that the above studies have initiated the investigation of RISs in mmWave ISAC systems, yet key challenges remain, including ensuring rate fairness among users and managing the non-convex optimization of active and passive beamforming. Therefore, considering the GM rate maximization as the communication performance metric along with the NDRIS in a mmWave ISAC system is eminently suitable for achieving rate fairness of the CUs and meeting the sensing requirements of the RTs. However, maximizing the GM rate in an RIS-aided mmWave ISAC system with the NDRIS phase shift matrix is challenging due to the intractable expression of the GM rate and the non-convex constraints of the HPTCs and the NDRIS phase shift matrix. Furthermore, optimizing the locations of the non-zero elements of the NDRIS phase shift matrix exacerbates the challenges.

Motivated by these facts and to address the challenging problems described above, we consider a RIS-aided

TABLE 1. Overview of Literature on RIS-aided mmWave ISAC Systems

	[9]	[10]	[12]	[14]	[15]	[18]	[19]	[22]	[23]	[24]	[25]	[37]	[38]	Proposed
RIS-aided mmWave systems	✓	✓		✓	✓							✓	✓	✓
Joint active and passive beamforming	✓		✓	✓	✓	✓	✓					✓	✓	✓
NDRIS	✓		✓	✓										✓
mmWave ISAC systems								✓	✓	✓	✓	✓	✓	✓
Multiple communication users (CUs)		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Multiple radar targets (RTs)								✓		✓	✓		✓	✓
Block coordinate descent (BCD) approach					✓					✓		✓		✓
Riemannian manifold optimization (RMO)	✓				✓				✓	✓		✓	✓	✓
Majorization and minimization (MM)	✓				✓									✓
NDRIS-aided mmWave ISAC														✓
Geometric mean rate maximization						✓	✓							✓
Penalty-based RMO														✓
MM-based BCD														✓

mmWave ISAC system relying on an NDRIS phase shift matrix, where an ISAC BS serves multiple CUs with the aid of an RIS and simultaneously ensures meeting the minimum sensing requirements of the RTs. To this end, we formulate an optimization problem for maximizing the GM rate of the CUs, while considering the minimum beampattern gain of the RTs, transmit power, and non-convex UM constraints of the RF TPC and the NDRIS phase shift matrix. To solve this challenging problem, we propose efficient iterative algorithms based on the MM and BCD techniques. Our simulation results comprehensively compare the performance achieved to that of the benchmark schemes, which demonstrates the efficacy of the proposed methods and the suitability of the NDRIS phase shift matrix in RIS-enabled mmWave ISAC systems. We boldly contrast our novel contributions to the related literature in Table 1, and the main contributions of this paper are enumerated next.

B. Contributions of this work

1) We conceive a RIS-aided mmWave ISAC system, where an ISAC BS provides communication services to single-antenna users, while simultaneously sensing multiple radar targets. In this system model, the RIS supports communication by enhancing the signal quality and coverage, thus enabling improved link reliability and increased capacity.
2) We formulate a communication-centric optimization problem that aims for optimizing the joint HBF and RIS phase shift matrix to maximize the geometric mean rate for all users, subject to the constraint of a minimum acceptable radar transmit power from the BS for the targets, along with additional limitations related to the finite transmit power and unit modulus constraints for both the RF precoder and the RIS reflective elements.
3) To tackle this novel problem, we initially reformulate the intractable GM rate objective in an equivalent tractable form by converting the non-linear product of user rates to a weighted sum rate of the users. However, the reformulated problem remains a highly coupled non-convex problem. To solve the above problem, we propose an efficient Block coordinate descent (BCD) method by decoupling it into three sub-problems, each one focusing on the separate design of the BB TPC $\mathbf{F}_{\mathrm{BB}}$, RF TPC $\mathbf{F}_{\mathrm{RF}}$, and the RIS phase shift matrix $\boldsymbol{\Theta}$, respectively.
4) At each stage of optimizing these variables, we transform the non-convex objective function using the theory of majorization-minimization (MM), which seeks a suitable surrogate function that acts as a lower-bound for the rate expression. For BB TPC optimization, we reformulate the pertinent problem as a convex quadratic program. Subsequently, the Riemannian Conjugate Gradient (RCG) method is utilized for optimizing both the RF TPC and the RIS phase shift matrix.
5) Finally, our findings demonstrate that optimizing the GM rate strikes a compelling balance between the reliability of sensing and efficiency of communication, ultimately leading to significantly enhanced overall system performance in comparison to other benchmarks.

C. Notations

The following notations are used throughout the paper: Vectors and matrices are denoted as $\mathbf{x}$ and $\mathbf{X}$, respectively. The matrices operations $\mathbf{X}(i,j)$, $\mathbf{X}^T$, and $\mathbf{X}^H$ represent the (i,j)th element, transpose, and Hermitian transpose of $\mathbf{X}$, respectively. The trace, Frobenius norm, and vectorization of a matrix $\mathbf{X}$ are denoted as $\mathrm{tr}(\cdot)$, $|\mathbf{X}|_F$, and $\mathrm{vec}(\cdot)$, respectively. A diagonal matrix with the vector $\mathbf{x}$ on its main diagonal is written as $\mathcal{D}(\mathbf{x})$. The expectation operator is denoted by $\mathbb{E}\cdot$, and the real part of a number is written as $\Re\cdot$. The $M \times M$ identity matrix is represented by $\mathbf{I}_M$. The symmetric complex Gaussian distribution with mean μ and covariance matrix σ^2 is denoted as $\mathcal{CN}(\mu, \sigma^2)$. Additionally,

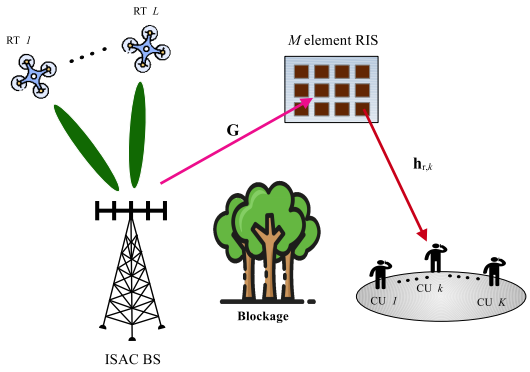

FIGURE 1. Illustration of a RIS-aided mmWave ISAC system

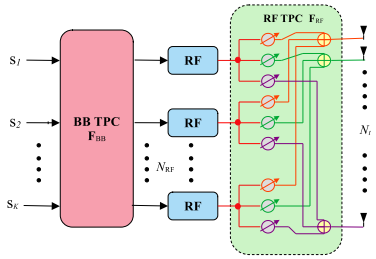

FIGURE 2. Block diagram of HTPC at the ISAC BS.

the operators $\odot$ and $\otimes$ represent the Hadamard product and Kronecker product, respectively.

II. System Model

As shown in Fig. 1, we consider a RIS-assisted millimeter-wave ISAC system, where an ISAC BS equipped having N_{t} transmit antennas is serving K CUs with a single antenna each, while simultaneously detecting L RTs, which are indexed by $k \in \mathcal{K} \triangleq \{1, \ldots, K\}$ and $l \in \mathcal{L} \triangleq \{1, \ldots, L\}$, respectively. Furthermore, a RIS comprising an NDRIS phase shift matrix with M elements indexed as $m \in \mathcal{M} \triangleq \{1, \ldots, M\}$ is deployed closer to the CUs to enhance the SINR of CUs. Assuming that the RTs are located far from the RIS ensures that the RIS does not interfere with the sensing task, while enhancing the performance of the CUs. Note that even though the NDRIS is capable of enhancing not only the communication signal, but also radar target detection, we assume that radar sensing is performed primarily via the direct line-of-sight path between the ISAC BS and RTs. This represents a mmWave scenario, where signals reflected by the RIS suffer significant path loss, making their contribution to sensing negligible compared to the direct link. As shown in Fig. 2, the overall HTPC in the hybrid architecture splits into a BB TPC as well as an RF TPC, and both are connected via a limited number of RFCs N_{RF}. To ensure the feasibility of the HPTC design, we assume the condition $K < N_{\mathrm{RF}} << N_{\mathrm{t}}$.

A. Signal model

Let us consider $\mathbf{s} = [s_1, s_2, \ldots, s_K]^T \in \mathbb{C}^{K\times 1}$ as the BB data stream vector, where the data symbol s_k is intended for the kth CU. Furthermore, the data symbols are assumed to be mutually independent, satisfying $\mathbb{E}\{\mathbf{s}\} = \mathbf{0}$ and $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}_K$. Following the hybrid MIMO architecture, the downlink signal $\mathbf{x} \in \mathbb{C}^{N_{\mathrm{t}}\times 1}$ transmitted by employing the HPTC from the ISAC BS, which is used to serve K CUs as well as to sense L RTs, is expressed as

$$\mathbf{x} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} = \mathbf{F}_{\mathrm{RF}} \sum_{k\in\mathcal{K}} \mathbf{f}_{\mathrm{BB},k} s_k. \tag{1}$$

The matrix $\mathbf{F}_{\mathrm{BB}} = [\mathbf{f}_{\mathrm{BB},1}, \ldots, \mathbf{f}_{\mathrm{BB},K}] \in \mathbb{C}^{N_{\mathrm{RF}}\times K}$ represents the BB TPC, in which $\mathbf{f}_{\mathrm{BB},k} \in \mathbb{C}^{N_{\mathrm{RF}}\times 1}$ denotes the BB TPC vector intended for the kth CU. Furthermore, $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{RF}}}$ represents the RF TPC, with each entry having a constant modulus of unity, i.e., $|\mathbf{F}_{\mathrm{RF}}(i,j)| = 1, \forall i,j$. Moreover, the signal $\mathbf{x}$ transmitted from the ISAC BS is also directed to the RIS, where the phase shift of the incident is adjusted for ensuring constructive interference at the intended CUs. Since this paper considers an NDRIS phase shift matrix having only M non-zero quantities, let us define the non-zero quantities of the phase shift matrix $\mathbf{\Theta} \in \mathbb{C}^{M\times M}$ as $|\theta_{f(m),m}| = 1, \forall m \in \mathcal{M}$. Here the quantity $\theta_{f(m),m}$ denotes the phase shift that occurs for the wave impinging at the mth and reflected from the $f(m)$th element. Furthermore, $f(m)$ is the bijective function that maps the indices for the incident wave to the wave reflected via the configuration of the switches. For more details about the NDRIS, readers are encouraged to refer to [12]–[14].

B. Communication model

In the system under consideration, we define the channels spanning from ISAC BS to RIS as $\mathbf{G} \in \mathbb{C}^{M\times N_{\mathrm{t}}}$, and from the RIS to the kth CU as $\mathbf{h}_{\mathrm{r},k}^H \in \mathbb{C}^{1\times M}$. Since the mmWave channels have a limited effective number of scatters in the propagation environment, we adopt the well-known Saleh-Valenzuela sparse geometric multipath mmWave channel model [14], [29], [30], which effectively captures key propagation features like sparsity and limited scattering. While real-world scenarios involve more complex effects, this model provides a strong foundation for our analysis, with future work aiming to incorporate blockage and dynamic environments. Following this model, the channel matrix $\mathbf{G}$ can be expressed as

$$\mathbf{G} = \sqrt{\frac{N_{\mathrm{t}}M}{L_p}} \sum_{\ell=0}^{L_{\mathrm{p}}} \alpha_\ell \mathbf{a}_{\mathrm{B}}\left(\psi_\ell^{\mathrm{B}}\right) \mathbf{a}_{\mathrm{R}}^{\mathrm{H}}\left(\psi_\ell^{\mathrm{R}}\right), \tag{2}$$

where L_p and $\alpha^{(\ell)}$ denote the number of propagation paths and the complex path gain of the ℓth path. Furthermore, $\mathbf{a}_{\mathrm{B}}\left(\psi_\ell^{\mathrm{B}}\right) \in \mathbb{C}^{N_{\mathrm{t}}\times 1}$ and $\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}\left(\psi_\ell^{\mathrm{R}}\right) \in \mathbb{C}^{N_{\mathrm{r}}\times 1}$ are the array steering vectors corresponding to the azimuth angles of

departures (AoDs), ψ_l^{B} and to the angle of arrivals (AoAs), ψ_l^{R} of the ISAC BS and the RIS, respectively. Considering uniform linear arrays (ULAs) at the ISAC BS and RIS, the steering vectors can be written as

$$\mathbf{a}_z\left(\psi^z\right)=\frac{1}{\sqrt{N_z}}\Bigg[1,\ldots,e^{j\frac{2\pi d}{\lambda}\sin(\psi^z)i},\ldots,e^{j\frac{2\pi d}{\lambda}\sin(\psi^z)(N_z^h-1)}\Bigg]^T, \tag{3}$$

where we have $z\in\{\mathrm{B},\ \mathrm{R}\}$, d is the antenna spacing. $0\leq i<N_z^h$, where N_z^h denotes the number of elements of the ULA in the 1D plane. By contrast, the RIS is often placed at high elevations, where it establishes LoS channels with the CUs, mitigating the impact of multiple reflections. Thus, the channel $\mathbf{h}_{\mathrm{r},k}^H\in\mathbb{C}^{1\times M}$ is given by

$$\mathbf{h}_{\mathrm{r},k}^H=\sqrt{M}\beta_k\mathbf{a}_{\mathrm{R}}^H(\psi_k), \tag{4}$$

where β_k represent the complex path gain. We assume that all the CSI involved in the system model are perfectly known. As a result, the signal received at the kth CU is expressed as

$$y_k=\mathbf{h}_{\mathrm{r},k}^H\boldsymbol{\Theta}\mathbf{G}\mathbf{F}_{\mathrm{RF}}\sum_{j=1}^{K}\mathbf{f}_{\mathrm{BB},j}s_j+n_k,\ \forall k\in\mathcal{K}, \tag{5}$$

where n_k is the i.i.d. complex additive white Gaussian noise (AWGN) at the kth CU with the distribution $n_k\sim\mathcal{CN}(0,\sigma_k^2)$. Upon defining $\mathbf{h}_{\mathrm{br},k}^H(\boldsymbol{\Theta})\triangleq\mathbf{h}_{\mathrm{r},k}^H\boldsymbol{\Theta}\mathbf{G}\in\mathbb{C}^{1\times N_{\mathrm{t}}}$, the SINR of the kth user is given by

$$\gamma_k(\boldsymbol{\Theta},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}})=\frac{\left|\mathbf{h}_{\mathrm{br},k}^H(\boldsymbol{\Theta})\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},k}\right|^2}{\sum\limits_{j\neq k}^{K}\left|\mathbf{h}_{\mathrm{br},k}^H(\boldsymbol{\Theta})\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},j}\right|^2+\sigma_k^2},\forall k\in\mathcal{K}. \tag{6}$$

Thus, the achievable rate of the kth CU in nats/s is expressed as

$$R_k\left(\boldsymbol{\Theta},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}\right)=\ln\left[1+\gamma_k\left(\boldsymbol{\Theta},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}\right)\right],\forall k\in\mathcal{K}. \tag{7}$$

C. Radar model

We assume that the ISAC BS operates in tracking mode to detect the RTs located at azimuth angles $\varphi_l,\forall l\in\mathcal{L}$, while also providing communication services to CUs. Since the radar sensing considers LoS propagation, the signal transmitted in the direction of the lth RT is given by

$$y_l(\varphi_l)=\mathbf{a}_{\mathrm{B}}^H(\varphi_l)\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s},\quad\forall l\in\mathcal{L}. \tag{8}$$

To achieve effective radar sensing, it is crucial to radiate the transmitted signal power from the ISAC BS in the directions of the RTs. This is because, maximizing the radiated power towards $\varphi_l,\forall l\in\mathcal{L}$ results in a strong received radar echo signal. Therefore, we express the power radiated toward the direction of the lth RT, as follows

$$P(\varphi_l)=\mathbb{E}\{|y_l(\varphi_l)|^2\},\ \forall l\in\mathcal{L}, \tag{9a}$$
$$=\mathbf{a}_{\mathrm{B}}^H(\varphi_l)\mathbf{R}_{\mathrm{x}}\mathbf{a}_{\mathrm{B}}(\varphi_l),\ \forall l\in\mathcal{L}, \tag{9b}$$

where $\mathbf{R}_{\mathrm{x}}=\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^H\mathbf{F}_{\mathrm{RF}}^H\in\mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{t}}}$ is the covariance matrix of the transmitted signal $\mathbf{x}$.

D. Problem formulation

The main objective of this paper is to jointly optimize the BB TPC $\mathbf{F}_{\mathrm{BB}}$, RF TPC $\mathbf{F}_{\mathrm{RF}}$, and the NDRIS phase shift matrix $\boldsymbol{\Theta}$ to achieve rate fairness among CUs in the RIS-assisted mmWave ISAC systems. Specifically, we aim for maximizing the GM rate of CUs, while maintaining the minimum power radiated in the direction of the RTs required for sensing, and mmWave, the NDRIS hardware constraints. Mathematically, we formulate the pertinent optimization problem as follows

$$\mathcal{P}_0:\quad\max_{\mathbf{F}_{\mathrm{RF}},\ \mathbf{F}_{\mathrm{BB}},\ \boldsymbol{\Theta}}\left(\prod_{k\in\mathcal{K}}R_k(\boldsymbol{\Theta},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}})\right)^{\frac{1}{K}} \tag{10a}$$
$$\text{s. t. } P(\varphi_l)\geq\Gamma_l,\forall l\in\mathcal{L}, \tag{10b}$$
$$\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2\leq P_{\mathrm{t}}, \tag{10c}$$
$$|\mathbf{F}_{\mathrm{RF}}(i,j)|=1,\forall i,j, \tag{10d}$$
$$\left|\theta_{f(m),m}\right|=1,\forall m\in\mathcal{M}, \tag{10e}$$

where constraint (10b) is to guarantee that the minimum power radiated in the direction of the lth RT is not lower than the specified threshold Γ_l, while (10c) denotes the transmit power constraint of the BS. Furthermore, the last two constraints (10d) and (10e) are the UM constraints imposed due to phase shifters in the RF TPC and the NDRIS phase shift matrix. The non-diagonal components of the phase shift matrix not indexed by $f(m)$ in (10e) are structurally zero and not included in the optimization, reflecting the inherent hardware design of the NDRIS. Note that this work assumes the availability of full CSI and known sensing thresholds $\Gamma_l,\forall l$, at the ISAC BS, which reflects a centralized processing framework similar to that in [39], wherein all relevant information is aggregated at a central node to enable joint optimization. This assumption allows us to explore the theoretical performance limits of the proposed GM rate maximization framework. However, in practical deployments, CSI estimation errors may degrade the SINR performance due to beam misalignment and increased interference. Investigating robust optimization strategies that account for such imperfections is an important avenue for future research. Nevertheless, it is clear that the above optimization problem (10) is computationally challenging to solve due to the non-tractable expression of the GM rate (10a) and highly non-convex UM constraints (10d) and (10e). Additionally, the bijective function $f$involves combinatorial optimization as well as tightly coupled optimization variables in the objective function and the above constraints, rendering the problem highly challenging. To solve Problem (10), we propose an highly efficient majorization-minimization-based block coordinate descent (MM-BCD) framework for jointly optimizing the BB TPC, RF TPC, and the NDRIS phase shift matrix, which is discussed in the next section.

III. Proposed approach

To solve Problem (10), let us first transform the intractable expression of the GM rate (10) into a tractable form. To this end, we define the function $f_{\mathrm{GM}}(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) \triangleq \left(\prod_{k\in\mathcal{K}} R_k(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})\right)^{\frac{1}{K}}$ corresponding to the GM rate of the system. Observe that the function $f_{\mathrm{GM}}(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ is highly non-linear due to the composition of the convex function $\left(\prod_{k\in\mathcal{K}} R_k\right)^{\frac{1}{K}}$ and the non-convex functions $R_k(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}), k \in \mathcal{K}$. To convert $f_{\mathrm{GM}}(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ into an equivalent linear form, let us consider the point $(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)})$ as the optimal feasible solution obtained from the $(\kappa-1)$st iteration. As a result, a linearized form of the function $f_{\mathrm{GM}}(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ around the point $(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)})$ can be derived as [18]

$$\frac{f_{\mathrm{GM}}(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)})}{K} \sum_{k\in\mathcal{K}} \frac{R_k(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})}{R_k(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)})}. \tag{11}$$

Consequently, Problem (10) is transformed into the following steepest descent optimization problem,

$$\max_{\mathbf{F}_{\mathrm{RF}},\, \mathbf{F}_{\mathrm{BB}},\, \mathbf{\Theta}} \frac{1}{K} \sum_{k\in\mathcal{K}} \frac{f_{\mathrm{GM}}(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)})}{R_k(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)})} R_k(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) \tag{12a}$$

$$\text{s. t.} \quad \text{(10b),(10c),(10d), and (10e)}, \tag{12b}$$

which is equivalent to the following problem of adaptive weighted sum rate maximization

$$\max_{\mathbf{F}_{\mathrm{RF}},\, \mathbf{F}_{\mathrm{BB}},\, \mathbf{\Theta}} f(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) \triangleq \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} R_k(\mathbf{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) \tag{13a}$$

$$\text{s. t.} \quad \text{(10b),(10c),(10d), and (10e)}, \tag{13b}$$

where $\eta_k^{(\kappa)}$ is the adaptive weight. For computational stability, a scaled version of $\eta_k^{(\kappa)}$ may be formulated as

$$\eta_k^{(\kappa)} = \frac{\max\limits_{k'\in\mathcal{K}} R_{k'}\left(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right)}{R_k\left(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right)}, \ \forall k \in \mathcal{K}. \tag{14}$$

However, the objective function (13a) of the transformed problem (13) is now tractable, but it is still non-convex due to the presence of multiple fractional SINR terms. Furthermore, the optimization variables $\mathbf{F}_{\mathrm{BB}}$, $\mathbf{F}_{\mathrm{RF}}$, and $\mathbf{\Theta}$ are tightly coupled both in the objective function and constraints of (13), which exacerbates the challenges of solving (13). To handle this issue, we adopt the popular BCD approach for optimizing the BB TPC $\mathbf{F}_{\mathrm{BB}}$, RF TPC $\mathbf{F}_{\mathrm{RF}}$, and the NDRIS phase shift matrix $\mathbf{\Theta}$ with a fixed value $\eta_k^{(\kappa)}, \forall k \in \mathcal{K}$ at the κth iteration. Moreover, at each stage of optimizing these variables, we transform the non-convex objective function (13a) to a convex one via the MM method, which seeks suitable lower-bound surrogate functions.

A. Optimizing the BB TPC $\mathbf{F}_{\mathrm{BB}}$

Given $\left(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right)$, we seek to optimize the next feasible $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$ that satisfies the following criterion

$$f^{(\kappa)}\left(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right) > f^{(\kappa)}\left(\mathbf{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right), \tag{15}$$

by considering the following sub-problem for BB TPC

$$\mathcal{P}_1: \quad \max_{\mathbf{F}_{\mathrm{BB}}} \quad \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \ln\left(1+\gamma_k^{(\kappa)}(\mathbf{F}_{\mathrm{BB}})\right) \tag{16a}$$

$$\text{s. t.} \quad \sum_{k\in\mathcal{K}} \mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k} \geq \Gamma_l, \forall l, \tag{16b}$$

$$\sum_{k\in\mathcal{K}} \mathbf{f}_{\mathrm{BB},k}^H \Omega_p^{(\kappa)} \mathbf{f}_{\mathrm{BB},k} \leq P_{\mathrm{t}}, \tag{16c}$$

where $\gamma_k^{(\kappa)}(\mathbf{F}_{\mathrm{BB}}) = \frac{\left|\left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},k}\right|^2}{\sum_{j\neq k}^{K}\left|\left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},j}\right|^2 + \sigma_k^2}$ with $\left(\mathbf{h}_k^{(\kappa)}\right)^H = \mathbf{h}_{\mathrm{br},k}^H\left(\mathbf{\Theta}^{(\kappa)}\right)\mathbf{F}_{\mathrm{RF}}^{(\kappa)} \in \mathbb{C}^{1\times N_{\mathrm{t}}}$, $\Omega_l^{(\kappa)} = \left(\mathbf{F}_{\mathrm{RF}}^{(\kappa)}\right)^H \mathbf{a}_{\mathrm{B}}(\varphi_l)\mathbf{a}_{\mathrm{B}}^H(\varphi_l)\mathbf{F}_{\mathrm{RF}}^{(\kappa)} \in \mathbb{C}^{N_{\mathrm{RF}}\times N_{\mathrm{RF}}}$ and $\Omega_p^{(\kappa)} = \left(\mathbf{F}_{\mathrm{RF}}^{(\kappa)}\right)^H \mathbf{F}_{\mathrm{RF}}^{(\kappa)} \in \mathbb{C}^{N_{\mathrm{RF}}\times N_{\mathrm{RF}}}$. Then, by rearranging the terms of (44) as given in Appendix A and neglecting the constant quantities, one can equivalently express problem $\mathcal{P}_1$ as follows

$$\min_{\mathbf{F}_{\mathrm{BB}}} \quad \sum_{k\in\mathcal{K}} \mathbf{f}_{\mathrm{BB},k}^H \Delta^{(\kappa)} \mathbf{f}_{\mathrm{BB},k} - 2\sum_{k\in\mathcal{K}} \Re\left\{\left(\mathbf{d}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},k}\right\} \tag{17a}$$

$$\text{s. t.} \quad \text{(16b), (16c)}, \tag{17b}$$

where $\Delta^{(\kappa)} = \sum_{j=1}^{K} \eta_j^{(\kappa)} b_j^{(\kappa)} \|\mathbf{h}_j^{(\kappa)}\|_2^2$ and $\left(\mathbf{d}_k^{(\kappa)}\right)^H = \eta_k^{(\kappa)} a_k^{(\kappa)}\left(\mathbf{h}_k^{(\kappa)}\right)^H, \forall k \in \mathcal{K}$. Similarly, we transform the non-concave constraint (16b) by approximating it with a suitable surrogate function via the MM technique. Specifically, a convex lower bound of the concave function $\mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k}$ around the point $\mathbf{f}_{\mathrm{BB},k}^{(\kappa)}$ can be constructed as follows [40]

$$\mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k} \geq 2\Re\left\{\mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k}\right\} - \mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k}. \tag{18}$$

Following the above transformations, problem (17) can be reformulated as follows

$$\min_{\mathbf{F}_{\mathrm{BB}}} \quad \sum_{k\in\mathcal{K}} \mathbf{f}_{\mathrm{BB},k}^H \Delta^{(\kappa)} \mathbf{f}_{\mathrm{BB},k} - 2\sum_{k\in\mathcal{K}} \Re\left\{\left(\mathbf{d}_k^{(\kappa)}\right)^H \mathbf{v}_k\right\} \tag{19a}$$

$$\text{s. t.} \quad 2\Re\left\{\mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k}\right\} - \mathbf{f}_{\mathrm{BB},k}^H \Omega_l^{(\kappa)} \mathbf{f}_{\mathrm{BB},k} \geq \Gamma_l, \forall l \in \mathcal{L}, \text{ and (16c)}. \tag{19b}$$

Observe that the Problem (19) above is convex, and we solved it via the interior-point method [41].

B. Optimizing the RF TPC $\mathbf{F}_{\mathrm{RF}}$

Given $\left(\boldsymbol{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right)$, we now seek to optimize the next feasible RF TPC $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ that satisfies the criterion of

$$f^{(\kappa)}\left(\boldsymbol{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right) > f^{(\kappa)}\left(\boldsymbol{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right), \tag{20}$$

by considering the following sub-problem for the RF TPC

$$\begin{aligned}
\mathcal{P}_2: \quad & \max_{\mathbf{F}_{\mathrm{RF}}} \quad \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \ln\left(1+\gamma_k^{(\kappa)}(\mathbf{F}_{\mathrm{RF}})\right) && \text{(21a)}\\
& \text{s. t.} \quad \mathrm{tr}\left(\mathbf{F}_{\mathrm{RF}}^H \mathbf{A}_l \mathbf{F}_{\mathrm{RF}} \mathbf{B}\right) \geq \Gamma_l, \forall l \in \mathcal{L}, && \text{(21b)}\\
& \quad\quad \mathrm{tr}\left(\mathbf{F}_{\mathrm{RF}}^H \mathbf{I}_{N_{\mathrm{t}}} \mathbf{F}_{\mathrm{RF}} \mathbf{B}\right) \leq P_{\mathrm{t}}, && \text{(21c)}\\
& \quad\quad |\mathbf{F}_{\mathrm{RF}}(i,j)| = 1, \forall i,j, && \text{(21d)}
\end{aligned}$$

where $\gamma_k^{(\kappa)}(\mathbf{F}_{\mathrm{RF}}) = \gamma_k\left(\boldsymbol{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right)$, $\mathbf{B} = \left(\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right)^H \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$ and $\mathbf{A}_l = \mathbf{a}_{\mathrm{B}}(\varphi_l)\mathbf{a}_{\mathrm{B}}^H(\varphi_l)$. Furthermore, by defining $\boldsymbol{\phi} = \mathrm{vec}\left(\mathbf{F}_{\mathrm{RF}}\right) \in \mathbb{C}^{N_{\mathrm{t}}N_{\mathrm{RF}}\times 1}$ and using the transformation $\mathbf{pQr} = \left(\mathbf{r}^T \otimes \mathbf{p}\right)\mathrm{vec}(\mathbf{Q})$, the quantity $\gamma_k^{(\kappa)}(\mathbf{F}_{\mathrm{RF}})$ in terms of $\boldsymbol{\phi}$ can be rewritten as $\gamma_k^{(\kappa)}(\boldsymbol{\phi}) = \frac{\left|\left(\widetilde{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \boldsymbol{\phi}\right|^2}{\sum_{j\neq k}^{K}\left|\left(\widetilde{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \boldsymbol{\phi}\right|^2 + \sigma_k^2}$ with $\left(\widetilde{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \triangleq (\mathbf{f}_{\mathrm{BB},j}^{(\kappa+1)})^T \otimes \mathbf{h}_{\mathrm{br},k}^H(\boldsymbol{\Theta}^{(\kappa)}) \in \mathbb{C}^{1\times N_{\mathrm{t}}N_{\mathrm{RF}}}$. Subsequently, following the transformations (48) and (50) given in Appendix B, the problem $\mathcal{P}_2$ can be recast as follows

$$\begin{aligned}
\mathcal{P}_{2-A}: \quad & \min_{\boldsymbol{\phi}} \quad \mathcal{F}^{(\kappa)}(\phi) = \boldsymbol{\phi}^H \mathbf{T}^{(\kappa)} \boldsymbol{\phi} - 2\Re\{\mathbf{q}^{(\kappa)}\boldsymbol{\phi}\} && \text{(22a)}\\
& \text{s. t.} \quad \boldsymbol{\phi}^H \boldsymbol{\Lambda}_l^{(\kappa)} \boldsymbol{\phi} \geq \Gamma_l, \ \forall l \in \mathcal{L}, && \text{(22b)}\\
& \quad\quad \boldsymbol{\phi}^H \boldsymbol{\Pi}^{(\kappa)} \boldsymbol{\phi} \leq P_{\mathrm{t}}, && \text{(22c)}\\
& \quad\quad |\boldsymbol{\phi}(n)| = 1, \ \forall n \in \{1,\ldots,N_{\mathrm{t}}N_{\mathrm{RF}}\}, && \text{(22d)}
\end{aligned}$$

where we have $\boldsymbol{\Lambda}_l^{(\kappa)} = \left(\mathbf{B}^{(\kappa)}\right)^T \otimes \mathbf{A}_l^{(\kappa)}$ and $\boldsymbol{\Pi}^{(\kappa)} = \left(\mathbf{B}^{(\kappa)}\right)^T \otimes \mathbf{I}_{N_{\mathrm{t}}}$. Observe that the above problem $\mathcal{P}_{2-A}$ is challenging to solve in the Euclidean space due to the non-convex UM constraint (22d). It is worth noting that the constraint (22d) represents a $(N_{\mathrm{t}}N_{\mathrm{RF}})$-dimensional complex circle Riemannian manifold $\mathcal{R} = \{\boldsymbol{\phi} \in \mathbb{C}^{N_{\mathrm{t}}N_{\mathrm{RF}}\times 1} : |\boldsymbol{\phi}(n)| = 1, \forall 1 \leq n \leq N_{\mathrm{t}}N_{\mathrm{RF}}\}$. Thus, we propose a more efficient penalized Riemannian manifold optimization (PRMO) algorithm that solves the problem $\mathcal{P}_{2-A}$ in the Riemannian space by using the penalized method. Specifically, we first incorporate the constraints (22b) and (22c) into the objective function as penalty terms, resulting in a problem with only the UM constraint. Subsequently, we employ the Riemannian conjugate gradient (RCG) framework [29], [30] to handle the problem on the surface of the Riemannian manifold $\mathcal{R}$. Finally, we adjust the penalty term by updating the penalty parameter to satisfy the constraints (22b) and (22c). Next, we discuss the PRMO algorithm in detail.

1) *Transformation to a penalized manifold problem:*

To transform the problem $\mathcal{P}_{2-A}$ to an unconstrained problem on the surface of a Riemannian manifold $\mathcal{R}$, we add (22b) and (22c) into the objective function as a penalty term. Consequently, the resultant problem is given by

$$\begin{aligned}
\min_{\boldsymbol{\phi}} \quad & \mathcal{G}^{(\kappa)}(\phi) = \mathcal{F}^{(\kappa)}(\phi) + \lambda\left(\sum_{l=1}^{L} \Upsilon_l^{(\kappa)}(\boldsymbol{\phi}) + \chi^{(\kappa)}(\boldsymbol{\phi})\right) && \text{(23a)}\\
\text{s. t.} \quad & (22d), && \text{(23b)}
\end{aligned}$$

where $\lambda \geq 0$ is a penalty parameter and the penalty terms $\Upsilon_l^{(\kappa)}(\boldsymbol{\phi}), \chi^{(\kappa)}(\boldsymbol{\phi})$ are defined as

$$\Upsilon_l^{(\kappa)}(\boldsymbol{\phi}) \triangleq \left(\max\left\{0, \Gamma_l - \boldsymbol{\phi}^H \boldsymbol{\Lambda}_l^{(\kappa)} \boldsymbol{\phi}\right\}\right)^2, \ \forall l \in \mathcal{L}, \tag{24a}$$

$$\chi^{(\kappa)}(\boldsymbol{\phi}) \triangleq \left(\max\left\{0, \boldsymbol{\phi}^H \boldsymbol{\Pi}^{(\kappa)} \boldsymbol{\phi} - P_{\mathrm{t}}\right\}\right)^2. \tag{24b}$$

2) *RCG framework:*

For the given λ, problem (23) represents an unconstrained manifold optimization problem with the non-convex UM constraint (22d). Thus, we adopt the RCG method, which differs from the conventional gradient descent method performed in Euclidean space. The following are the key steps performed in each iteration of the RCG method:

Riemannian gradient: The Riemannian gradient at point ϕ denoted as $\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\phi)$ is obtained by orthogonally projecting the Euclidean gradient onto the tangent space of the manifold $T_\phi\mathcal{R}$ using a projection operator. Mathematically, $\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\phi)$ is given by

$$\begin{aligned}
\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\phi) &= \mathrm{Proj}_{\phi}\nabla\mathcal{G}^{(\kappa)}(\phi)\\
&= \nabla\mathcal{G}^{(\kappa)}(\phi) - \Re\left\{\nabla\mathcal{G}^{(\kappa)}(\phi) \odot \phi^*\right\} \odot \boldsymbol{\phi},
\end{aligned} \tag{25}$$

where Proj is the projection operator and $\nabla\mathcal{G}^{(\kappa)}(\phi)$ represents the Euclidean gradient, which is evaluated as

$$\nabla\mathcal{G}^{(\kappa)}(\phi) = 2\mathbf{T}^{(\kappa)}\boldsymbol{\phi} - 2\mathbf{q}^{(\kappa)H} + \lambda\left(\sum_{l=1}^{L}\boldsymbol{\xi}_l^{(\kappa)} + \boldsymbol{\Xi}^{(\kappa)}\right), \tag{26}$$

where $\boldsymbol{\xi}_l^{(\kappa)}$ and $\boldsymbol{\Xi}^{(\kappa)}$ are given by

$$\boldsymbol{\xi}_l^{(\kappa)} = \begin{cases} 4\left(\boldsymbol{\phi}^H \boldsymbol{\Lambda}_l^{(\kappa)} \boldsymbol{\phi} - \Gamma_l\right)\boldsymbol{\Lambda}_l^{(\kappa)}\boldsymbol{\phi}, & \Gamma_l \geq \boldsymbol{\phi}^H \boldsymbol{\Lambda}_l^{(\kappa)} \boldsymbol{\phi},\\ 0, & \Gamma_l < \boldsymbol{\phi}^H \boldsymbol{\Lambda}_l^{(\kappa)} \boldsymbol{\phi}, \end{cases} \tag{27a}$$

$$\boldsymbol{\Xi}^{(\kappa)} = \begin{cases} 4\left(\boldsymbol{\phi}^H \boldsymbol{\Pi}^{(\kappa)} \boldsymbol{\phi} - P_{\mathrm{t}}\right)\boldsymbol{\Pi}^{(\kappa)}\boldsymbol{\phi}, & \boldsymbol{\phi}^H \boldsymbol{\Pi}^{(\kappa)} \boldsymbol{\phi} \geq P_{\mathrm{t}}\\ 0, & \boldsymbol{\phi}^H \boldsymbol{\Pi}^{(\kappa)} \boldsymbol{\phi} < P_{\mathrm{t}}. \end{cases} \tag{27b}$$

Furthermore, the tangent space $T_\phi\mathcal{R}$ at a point $\boldsymbol{\phi}$ on the manifold $\mathcal{R}$ is a tangent space comprising vectors tangential to the smooth curves on the manifold $\mathcal{R}$, which is given by

$$T_\phi\mathcal{R} = \{\mathbf{z} \in \mathbb{C}^{N_t N_{RF}\times 1} | \Re\{\mathbf{z} \odot \phi^*\} = \mathbf{0}_{N_t N_{RF}\times 1}\}. \tag{28}$$

Steepest search direction: To accelerate convergence, we

obtain the efficient steepest search direction $\boldsymbol{\zeta}_{\iota+1}$ at the $(\iota+1)$st step of the RCG update, which aligns in the direction of minimizing the objective function $\mathcal{G}^{(\kappa)}(\phi)$. To achieve this, the search direction $\boldsymbol{\zeta}_{\iota}$ at $\boldsymbol{\phi}_{\iota}$ is combined with the Riemannian gradient $\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\boldsymbol{\phi}_{\iota+1})$, indicating the descent direction at point $\boldsymbol{\phi}_{\iota+1}$. However, these vectors cannot be added directly, since they lie in different tangential spaces. Therefore, we perform the following evolutionary operation given by $\mathcal{T}_{\boldsymbol{\phi}_{\iota}}\mathcal{R} \rightarrow \mathcal{T}_{\boldsymbol{\phi}_{\iota+1}}\mathcal{R}$, which maps the suitable search direction. Mathematically, the above evolutionary operation is given by

$$\mathcal{T}_{\phi_{\iota}\mapsto\phi_{\iota+1}}(\zeta_{\iota}) = \zeta_{\iota} - \Re\left\{\zeta_{\iota} \odot \phi^{*}_{\iota+1}\right\} \odot \phi_{\iota+1}. \tag{29}$$

Thereby, the steepest search direction in the $(\iota+1)$st step is given by

$$\boldsymbol{\zeta}_{\iota+1} = -\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\boldsymbol{\phi}_{\iota+1}) + \nu_{\iota+1}\,\mathcal{T}_{\phi_{\iota}\mapsto\phi_{\iota+1}}(\zeta_{\iota}), \tag{30}$$

where $\nu_{\iota+1}$ represents the Polak-Ribiére's conjugate parameter [29].

Retraction: Similar to the classic gradient descent approach, we evaluate the next point $\boldsymbol{\phi}_{\iota+1}$ by using the step size τ_{ι} and search direction ζ_{ι}. Yet, it is very likely that the next point $\boldsymbol{\phi}_{\iota+1} = \boldsymbol{\phi}_{\iota} + \tau_{\iota}\zeta_{\iota}$ does not fall on the manifold $\mathcal{R}$, but rather lies on $T_{\phi_{\iota+1}}\mathcal{R}$. Therefore, we perform retraction mapping for retracting back the point from $T_{\phi_{\iota+1}}\mathcal{R}$ to the manifold $\mathcal{R}$ as follows

$$\text{Retr}_{\boldsymbol{\phi}} : \mathcal{T}_{\boldsymbol{\phi}}\mathcal{R} \rightarrow \mathcal{R}:$$

$$\boldsymbol{\phi}_{\iota+1} = \left[\frac{(\boldsymbol{\phi}_{\iota}+\tau_{\iota}\zeta_{\iota})_1}{|(\phi_{\iota}+\tau_{\iota}\zeta_{\iota})_1|}, \ldots, \frac{(\phi_{\iota}+\tau_{\iota}\zeta_{\iota})_{N_{\mathrm{t}}N_{\mathrm{RF}}}}{|(\phi_{\iota}+\tau_{\iota}\zeta_{\iota})_{N_{\mathrm{t}}N_{\mathrm{RF}}}|}\right]^{T}, \tag{31}$$

where the step size τ_{ι} is obtained by the Armijo backtracking line search algorithm [29].

Update of penalty factor: It is worth noting that the penalty parameter λ plays a crucial role in achieving an optimal feasible $\mathbf{V}_{\mathrm{RF}}$. When λ is too small, then the resultant solution may fall far outside the feasible region, potentially leading to constraint violations. Thus, we increase the penalty factor λ as $\lambda = \lambda/w$, where we have $w \in (0,1)$ for ensuring that constraint violations are penalized sufficiently. Therefore, we iteratively optimize the RF TPC $\mathbf{F}_{\mathrm{RF}}$ via employing the PRMO algorithm, whose key steps are summarized in Algorithm 1.

Algorithm 1 PRMO algorithm for solving (22)

Input: $\boldsymbol{\Theta}_{\mathrm{RF}}^{(\kappa)}$, $\mathbf{F}_{\mathrm{RF}}^{(\kappa)}$, $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$, $\lambda \geq 1$, $0 < w < 1$, thresholds $\epsilon_1 > 0$, $\epsilon_2 > 0$

Output: optimal RF TPC $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$

initialize: $\iota = 0$, $\boldsymbol{\phi}_{\iota} = \mathrm{vec}(\mathbf{F}_{\mathrm{RF}}^{(\kappa)})$, $\zeta_{\iota} = -\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\boldsymbol{\phi}_{\iota})$

while $\left(\|\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\boldsymbol{\phi}_{\iota})\|_2^2 \geq \epsilon_1\right)$ **do**
  find τ_{ι} using Armijo backtracking line search algorithm
  update the next point $\phi_{\iota+\mathbf{1}}$ via retraction operation (31)
  evaluate the Riemannian gradient $\nabla_{\mathcal{R}}\mathcal{G}^{(\kappa)}(\phi_{\iota+\mathbf{1}})$ using (25).
  compute the search direction $\boldsymbol{\zeta}_{\iota+1}$ according to (30)
  set $\iota \leftarrow \iota + 1$
end while
end while
if $\left(\Upsilon_{l}^{(\kappa)}(\boldsymbol{\phi}_{\iota}) \leq \epsilon_2,\ \forall l \in \mathcal{L}\ \&\&\ \boldsymbol{\Pi}^{(\kappa)}(\boldsymbol{\phi}_{\iota}) \leq \epsilon_2\right)$
  return $\phi^{(\kappa+\mathbf{1})} = \phi_{\iota}$ **stop**
else
  update $\lambda = \frac{\lambda}{w}$ and **go to step 2**
end if
return: restore $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ from $\phi^{(\kappa+\mathbf{1})}$

C. Optimizing the NDRIS phase shift matrix $\boldsymbol{\Theta}$

Given $\left(\boldsymbol{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right)$, we further seek to optimize $\boldsymbol{\Theta}^{(\kappa+1)}$ that satisfies the following criterion

$$f^{(\kappa)}\left(\boldsymbol{\Theta}^{(\kappa+1)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right) > f^{(\kappa)}\left(\boldsymbol{\Theta}^{(\kappa)}, \mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right) \tag{32}$$

by considering the following sub-problem for the RIS phase shift matrix

$$\mathcal{P}_3: \quad \max_{\boldsymbol{\Theta}} \quad \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \ln\left(1+\gamma_k^{(\kappa)}(\boldsymbol{\Theta})\right) \tag{33a}$$

$$\text{s. t.} \quad \left|\theta_{f(m),m}\right| = 1, \forall m \in \mathcal{M}, \tag{33b}$$

where $\gamma_k^{(\kappa)}(\boldsymbol{\Theta}) = \gamma_k\left(\boldsymbol{\Theta}, \mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}, \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\right)$. It is worth noting that for the available CSI of $\mathbf{h}_{\mathrm{r},k}^{H}$ and $\mathbf{G}$, and fixed RF TPC $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ and BB TPC $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$, the beamforming gain of the kth CU via RIS is given by $\left|\mathbf{h}_{\mathrm{r},k}^{H}\boldsymbol{\Theta}\mathbf{G}\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}\mathbf{f}_{\mathrm{BB},k}^{(\kappa+1)}\right|^2$. To maximize this beamforming gain via the MRC criterion, one has to match the phases of $\boldsymbol{\Theta}$ to that of the ordered elements of the row vector $\mathbf{h}_{\mathrm{r},k}^{H}$ and column vector $\mathbf{G}\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}\mathbf{f}_{\mathrm{BB},k}^{(\kappa+1)}$ [12], [14]. To achieve this ordering, we define the permutation matrices $\mathbf{J}_{\mathrm{t}}$ and $\mathbf{J}_{\mathrm{r}}$, which, when applied to the respective vectors, arrange the elements of $\mathbf{h}_{\mathrm{r},k}^{H}$ and $\mathbf{G}\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}\mathbf{f}_{\mathrm{BB},k}^{(\kappa+1)}$ in ascending order by multiplying them on the right and left, respectively. As a result, the beamforming gain of the kth CU can be written as $\left|\mathbf{h}_{\mathrm{r},k}^{H}\mathbf{J}_{\mathrm{t}}\widetilde{\boldsymbol{\Theta}}\mathbf{J}_{\mathrm{r}}\mathbf{G}\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}\mathbf{f}_{\mathrm{BB},k}^{(\kappa+1)}\right|^2$, where $\widetilde{\boldsymbol{\Theta}} = \mathcal{D}\left(\tilde{\boldsymbol{\theta}}\right)$ is the conventional diagonal phase shift matrix with phase vector $\tilde{\boldsymbol{\theta}} = \left[\tilde{\theta}_1, \ldots, \tilde{\theta}_M\right]^T \in \mathbb{C}^{M\times 1}$. Therefore, we optimize the diagonal phase shift matrix $\widetilde{\boldsymbol{\Theta}}$ and subsequently obtain the phase shift matrix of the NDRIS as $\boldsymbol{\Theta} = \mathbf{J}_{\mathrm{t}}\widetilde{\boldsymbol{\Theta}}\mathbf{J}_{\mathrm{r}}$. Consequently, the resultant optimization problem in terms of $\widetilde{\boldsymbol{\Theta}}$ is given by

$$\mathcal{P}_{3-A}: \quad \max_{\widetilde{\boldsymbol{\Theta}}} \quad \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \ln\left(1+\gamma_k^{(\kappa)}(\widetilde{\boldsymbol{\Theta}})\right) \tag{34a}$$

$$\text{s. t.} \quad \left|\tilde{\theta}_m\right| = 1, \forall m \in \mathcal{M}. \tag{34b}$$

Furthermore, due to the non-convex SINR terms in the objective function and non-convex UM constraint imposed on the elements of $\widetilde{\boldsymbol{\Theta}}$ in (34), optimizing the diagonal phase shift matrix $\widetilde{\boldsymbol{\Theta}}$ is still challenging. To handle these issues, let us define $\mathbf{H}_k^{(\kappa)} = \mathcal{D}(\widetilde{\mathbf{h}}_{\mathrm{r},k}^H)\widetilde{\mathbf{G}} \in \mathbb{C}^{M\times N_\mathrm{t}}$, where $\widetilde{\mathbf{h}}_{\mathrm{r},k}^H = \mathbf{h}_{\mathrm{r},k}^H\mathbf{J}_\mathrm{t}$ and $\widetilde{\mathbf{G}} = \mathbf{J}_\mathrm{r}\mathbf{G}$. Thereby, the SINR term $\gamma_k^{(\kappa)}(\boldsymbol{\Theta})$ can be written in terms of $\tilde{\boldsymbol{\theta}}$ as follows

$$\gamma_k^{(\kappa)}(\tilde{\boldsymbol{\theta}}) = \frac{\left|\left(\widehat{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H\tilde{\boldsymbol{\theta}}\right|^2}{\sum\limits_{j\neq k}^{K}\left|\left(\widehat{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H\tilde{\boldsymbol{\theta}}\right|^2 + \sigma_k^2}, \tag{35}$$

where $\widehat{\mathbf{h}}_{k,j}^{(\kappa)} \triangleq \mathbf{H}_k^{(\kappa)}\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}\mathbf{f}_{\mathrm{BB},j}^{(\kappa+1)} \in \mathbb{C}^{M\times 1}$. As a further advance, we compactly rewrite (51) as follows

$$2\Re\left\{\left(\mathbf{p}^{(\kappa)}\right)^H\tilde{\boldsymbol{\theta}}\right\} - \tilde{\boldsymbol{\theta}}^H\mathbf{E}^{(\kappa)}\tilde{\boldsymbol{\theta}} + \hat{c}^{(\kappa)}, \tag{36}$$

where $\mathbf{E}^{(\kappa)}, \mathbf{p}^{(\kappa)}$ and $\hat{c}^{(\kappa)}$ are given by

$$\mathbf{E}^{(\kappa)} = \sum_{k\in\mathcal{K}}\eta_k^{(\kappa)}\hat{b}_k^{(\kappa)}\left(\sum_{j=1}^{K}\left(\widehat{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H\widehat{\mathbf{h}}_{k,j}^{(\kappa)}\right), \tag{37a}$$

$$\mathbf{p}^{(\kappa)} = \sum_{k\in\mathcal{K}}\eta_k^{(\kappa)}\hat{a}_k^{(\kappa)}\widehat{\mathbf{h}}_{k,k}^{(\kappa)}, \tag{37b}$$

$$\hat{c}^{(\kappa)} = \sum_{k\in\mathcal{K}}\eta_k^{(\kappa)}\hat{c}_k^{(\kappa)}. \tag{37c}$$

Upon following the modification given by Appendix C , the Problem $\mathcal{P}_3$ can be recast as follows

$$\mathcal{P}_{3-A}: \min_{\tilde{\boldsymbol{\theta}}} \quad \mathcal{I}^{(\kappa)}(\tilde{\boldsymbol{\theta}}) = \tilde{\boldsymbol{\theta}}^H\mathbf{E}^{(\kappa)}\tilde{\boldsymbol{\theta}} - 2\Re\left\{\left(\mathbf{p}^{(\kappa)}\right)^H\tilde{\boldsymbol{\theta}}\right\} \tag{38a}$$

$$\text{s .t. } |\tilde{\theta}_m| = 1, \ \forall m \in \mathcal{M}, \tag{38b}$$

Observe that the above problem (38) is still non-convex due to the UM constraint (38b). To handle this problem, we again adopt the RCG framework, where the Euclidean gradient of $\mathcal{I}^{(\kappa)}(\tilde{\boldsymbol{\theta}})$ is evaluated as follows

$$\nabla\mathcal{I}^{(\kappa)}(\tilde{\boldsymbol{\theta}}) = 2\mathbf{E}^{(\kappa)}\tilde{\boldsymbol{\theta}} - 2\mathbf{p}^{(\kappa)}. \tag{39}$$

Furthermore, the optimal solution to problem $\mathcal{P}_{3-A}$ is obtained via the PRMO approach presented in Algorithm 1, where the steps required for updating the penalty parameter are omitted. Algorithm 2 presents the pseudocode for designing joint HBF and RIS phase shift matrices by solving $\mathcal{P}_0$ utilizing the MM-BCD method.

D. Complexity analysis

This section evaluates the complexity of the MM-BCD optimization. In Algorithm 2, the interior-point method is employed for optimizing the BB TPC $\mathbf{F}_{\mathrm{BB}}$, with the variables involved being N_{RF} and K. Thus, the computational complexity of solving the convex Problem (19) to obtain $\mathbf{F}_{\mathrm{BB}}$ is given as $\mathcal{O}(N_{RF}^{3.5}K^{3.5})$. Algorithm 1 focuses on optimizing the sub-problems $\mathcal{P}_2$ and $\mathcal{P}_3$ associated with the RF TPC $\mathbf{F}_{\mathrm{RF}}$ and NDRIS phase shift $\boldsymbol{\Theta}$, respectively. The significant

Algorithm 2 Proposed MM-BCD algorithm for solving (10)

Input: P_t, $\Gamma_l \ \forall l \in \mathcal{L}$, convergence threshold

1: **initialize:** $\kappa = 0$, feasible $\boldsymbol{\Theta}^{(\kappa)}$, $\mathbf{F}_{\mathrm{RF}}^{(\kappa)}$ and $\mathbf{F}_{\mathrm{BB}}^{(\kappa)}$
2: **repeat**
3: compute $\eta_k^{(\kappa)}, \ \forall k \in \mathcal{K}$ using (14)
4: obtain $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$ for given $\boldsymbol{\Theta}^{(\kappa)}$ and $\mathbf{F}_{\mathrm{RF}}^{(\kappa)}$ by solving (19).
5: calculate $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ for given $\boldsymbol{\Theta}^{(\kappa)}$ and $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$ by solving (23) via Algorithm 1.
6: find $\boldsymbol{\Theta}_{\mathrm{RF}}^{(\kappa+1)}$ for given $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ and $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$ by solving (38) via RCG framework.
7: **until** the value of objective function $f_{\mathrm{GM}}(\boldsymbol{\Theta}, \mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ in (10) converges

contribution of the complexity of Algorithm 1 stems from the computation of the Euclidean gradient (27) and (39), which has the complexity order of approximately $\mathcal{O}(N_t^2N_{RF}^2)$ and $\mathcal{O}(M^2)$. Hence, the total computational complexity of solving sub-problems $\mathcal{P}_2$ and $\mathcal{P}_3$ using Algorithm 1 is given as $\mathcal{O}(\mathcal{I}_\mathrm{r}N_t^2N_{RF}^2)$ and $\mathcal{O}(M^2)$, where $\mathcal{I}_\mathrm{r}$ represents the combined number of iterations required for performing RCG and updating the penalty factor λ for sub-problem $\mathcal{P}_2$. From the above analysis, the overall computational complexity of Algorithm 2 for solving the GMR-max problem (10) is $\mathcal{O}\left[I_\mathrm{o}\left(N_{RF}^{3.5}K^{3.5} + \mathcal{I}_\mathrm{r}N_t^2N_{RF}^2 + M^2\right)\right]$, where I_o denotes the number of iterations required for convergence.

IV. SIMULATION RESULTS

In this section, we evaluate the communication and sensing performance using simulations for characterizing the proposed NDRIS-aided mmWave ISAC system. Throughout the simulations, we use the following settings, unless explicitly stated otherwise. As shown in Fig. 3, mmWave ISAC BS is located at the origin (0, 0), configured with a ULA array of $N_\mathrm{t} = 32$ transceiver antennas and $N_{\mathrm{RF}} = \{4, 6\}$ RFCs. The NDRIS structure comprises $M = 64$ elements, and it is deployed at the coordinates $(d_{\mathrm{RIS}}, 5)$ m, where d_{RIS} represents the distance between the RIS and the ISAC BS. Furthermore, we consider $K = 4$ CUs to be randomly placed within a circle of 5 m radius centered at $(30, 0)$ m. Following [14], [29], [30], the mmWave channel gains $\{\alpha^{(\ell)}, \beta_k\}, \forall l, k$ obey the distribution $\mathcal{CN}(0, 10^{-0.1PL(d_m)})$, where $PL(d_m)$ is the path loss, which is modeled as [14], [29], [30]

$$PL(d_i)\ [\mathrm{dB}] = \varepsilon_1 + \varepsilon_2 10\log_{10}(d_i) + \varpi, \tag{40}$$

where $\varpi \in \mathcal{N}(0, \sigma_\varpi^2)$ with $\varepsilon_1 = 72$, $\varepsilon_2 = 2.92$ and $\sigma_\varpi = 8.7$ dB. Additionally, $L = 3$ RTs of interest are located at angles of $-60°$, $-40°$, and $-20°$, and the minimum beam pattern gain requirement of each RT is assumed to be the same $\Gamma_l = \Gamma$ and set as $\Gamma = 20$ dBm. Furthermore, the system operates at 28 GHz with a bandwidth of 251.1886 MHz and total power budget of P_t = 30dBm. The bandwidth chosen reflects the typical value used in mmWave-aided

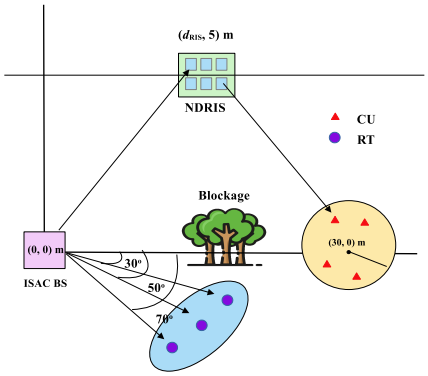

FIGURE 3. Simulation setup for the NDRIS-aided mmWave ISAC system.

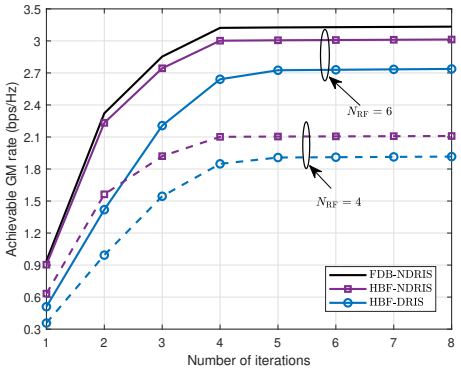

FIGURE 4. Achievable GM rate versus number of iterations.

systems, and it is consistent with the existing literature [9], [10], [14], [15], [25], ensuring realistic modeling of high-frequency system characteristics. Thus, the noise variance N_0 at each CU is set as $\sigma_k^2 = -174 + 10\log_{10} B = -90$ dBm. Finally, all results are multiplied by $\log_2 e$ to convert unit nats/sec to unit bps/Hz. Furthermore, we compare the proposed HBF-NDRIS scheme for jointly optimizing the HBF TPCs and RM schemes to the following benchmark schemes to demonstrate the efficacy of the algorithms proposed for NDRIS-aided mmWave ISAC systems:

- *Scheme 1* (FDB-NDRIS): This scheme considers an NDRIS-aided mmWave ISAC system relying on fully digital beamforming at the ISAC BS. Thus, we employ the MM-BCD algorithm with $N_{\mathrm{RF}} = N_{\mathrm{t}}$ for optimizing the hybrid TPCs and RM.
- *Scheme 2* (HBF-DRIS): In this scheme, we consider a conventional DRIS instead of NDRIS and employ the MM-BCD algorithm for optimizing the hybrid TPCs and RM of the DRIS.
- *Scheme 3* (Random NDRIS): This scheme involves setting the reflecting coefficients of the NDRIS RM to random values uniformly distributed over the interval $[0, 2\pi)$.

Although sparse precoding techniques such as Sparse Bayesian Learning (SBL) [42], [43] and codebook-based methods like Orthogonal Matching Pursuit (OMP) [44], [45] could be employed for HTPC design by reformulating the weighted sum-rate maximization problem in (13) into a weighted beamforming error minimization framework, as shown in [46], their adoption is avoided in this work. This is primarily due to the substantial additional complexity introduced by these methods and their sensitivity to algorithm-specific parameters. Given the already significant computational burden of optimizing the NDRIS phase shift matrix, incorporating SBL or OMP would render the overall system optimization impractically complex for the NDRIS-aided mmWave ISAC system considered.

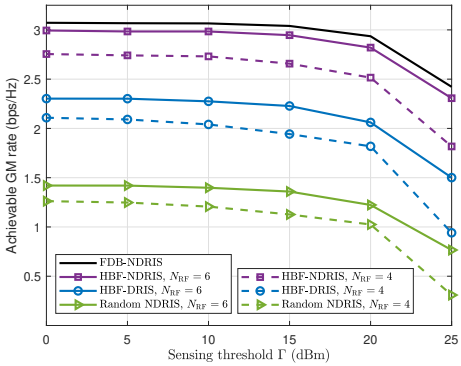

FIGURE 5. Achievable GM rate versus sensing threshold Γ.

A. Convergence behavior

Fig. 4 depicts the convergence performance of the proposed MM-BCD algorithm with the sensing and power termination thresholds of $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-2}$, respectively, for the PRMO algorithm. As shown in the figure, the GM rate objective function of the system for both $N_{\mathrm{RF}} = 4$ and 6 increases and then saturates within a few iterations, which shows the convergence behavior of the MM-BCD algorithm. Furthermore, the NDRIS-HBF scheme requires a similar number of iterations as the DRIS-HBF for its saturation. This demonstrates that despite employing switches, the computational efficiency of the NDRIS-HBF is close to that of its DRIS-HBF counterpart.

B. Communication performance

Fig. 5 plots the achievable GM rate versus the sensing threshold Γ of the RTs for $N_{\mathrm{RF}} = 4$ and 6. As seen from the figure, the achievable GM rate of the system degrades upon increasing Γ from 0 dBm to 25 dBm. This is intuitive because a higher value of Γ results in the radiation of higher transmit power towards the RTs with the aim of meeting the higher sensing requirements, thereby compromising the communication performance. Furthermore, the HBF-NDRIS scheme with only $N_{\mathrm{RF}} = 6$ performs close to the FDB-NDRIS scheme, which evidences the efficacy of the pro-

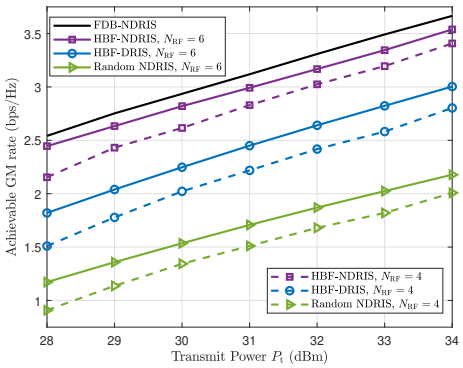

FIGURE 6. **Achievable GM rate versus transmit power P_{t}.**

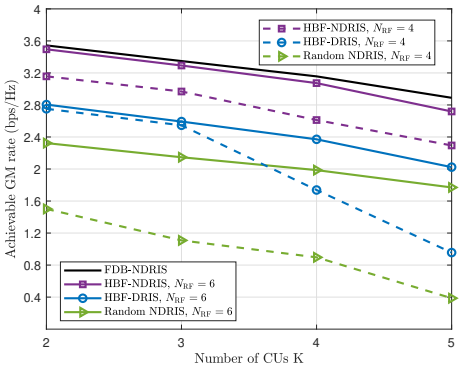

FIGURE 8. **Achievable GM rate versus number of CUs K.**

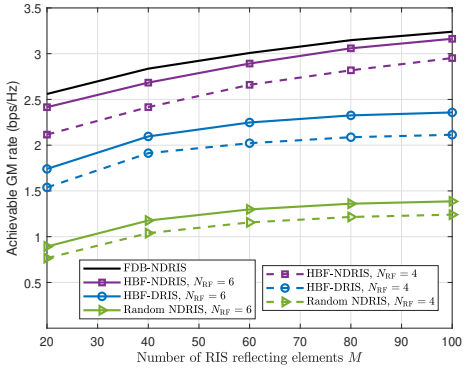

FIGURE 7. **Achievable GM rate versus reflective elements M.**

posed PRMO algorithm in achieving the optimization of the hybrid TPCs. Moreover, observe that the HBF-NDRIS scheme outperforms the conventional HBF-DRIS scheme at low as well as at higher values of Γ. This is due to the fact that the NDRIS yields a higher beamforming gain than DRIS architecture, rendering the former better suited for ISAC mmWave-aided systems. Furthermore, the Random NDRIS scheme performs poorly in contrast to the HBF-NDRIS and HBF-DRIS schemes, which reveals the efficacy of the proposed RCG method for the optimization of the RIS phase shift matrix.

Fig. 6 illustrates the achievable GM rate versus the transmit power of the system under consideration. As expected, the achievable GM rate of the system increases upon increasing the transmit power from 28 dBm to 34 dBm. This is due to the fact that the sensing constraint is rendered less effective at higher values of P_{t}, hence leading to the radiation of higher power towards the CUs, which results in an improvement in the communication performance. Furthermore, the performance of the HBF-NDRIS scheme is enhanced in comparison to the benchmark schemes, clearly at all P_{t} values, which demonstrates the efficacy of the NDRIS array in mmWave ISAC systems.

To further investigate the effect of the reflective elements on system performance, we plot the achievable GM rate versus the number of reflective elements M in Fig. 7. The achievable GM rate of the system increases upon increasing the number of reflective elements because of the higher passive beamforming gain provided by the RIS. Observe that the performance of the HBF-DRIS and Random schemes saturates at a higher number of reflective elements for both $N_{\mathrm{RF}} = 6$ and 4. By contrast, the performance of the HBF-NDRIS scheme increases linearly upon increasing the reflective elements. This is because the HBF-NDRIS scheme enjoys a higher beamforming gain as a benefit of the optimal MRC criterion for the optimization of the RIS phases. By contrast, the HBF-DRIS scheme uses the EGC criterion for phase optimization.

Fig. 8 reveals the effect of CUs on the achievable GM rate of the system for a fixed sensing threshold and transmit power. As expected, the GM rate decreases upon increasing the number of CUs for a fixed transmit power due to the resultant higher MUI and the availability of lower power per CU. Furthermore, observe that the HBF-NDRIS scheme using $N_{\mathrm{RF}} = 6$ performs close to that of the FDB-NDRIS at a higher number of CUs, which shows the efficacy of the PRMO algorithm in handling the MUI. The performance of the HBF-NDRIS scheme with $N_{\mathrm{RF}} = 4$ degrades upon increasing the number of CUs, but it still outperforms the benchmark schemes, interestingly also for $N_{\mathrm{RF}} = 6$. This demonstrates the improved MUI handling capability of the HBF-NDRIS scheme in comparison to the benchmark schemes.

To examine the effect of the RIS position, we plot the achievable GM rate versus the horizontal distance of the RIS by varying d_{RIS} from 10 m to 90 m in Fig. 9. Observed from the figure, the GM rate first decreases upon increasing d_{RIS} from 10 m to 50, but beyond its minimum, it increases again upon increasing d_{RIS} from 50 m to 90 m. This reveals the fact that placing the RIS near the ISAC BS or near the CUs is beneficial for achieving higher communication performance. Moreover, the HBF-NDRIS scheme with both $N_{\mathrm{RF}} = 6$ and 4 outperforms the benchmark scheme at each value of d_{RIS}, which is a benefit of the higher passive beamforming gain of the HBF-NDRIS scheme.

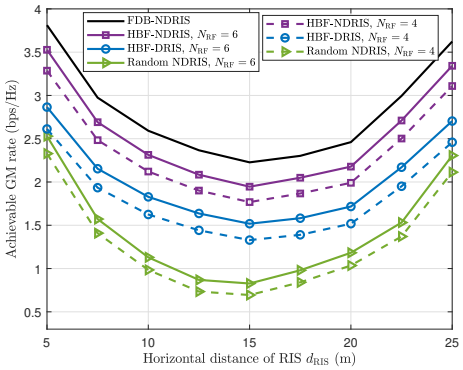

FIGURE 9. Achievable GM rate versus horizontal distance of RIS d_{RIS}.

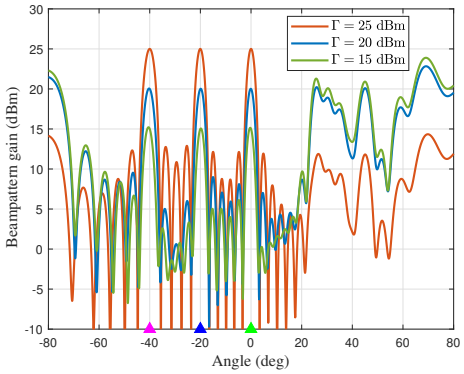

FIGURE 11. Transmit beampattern.

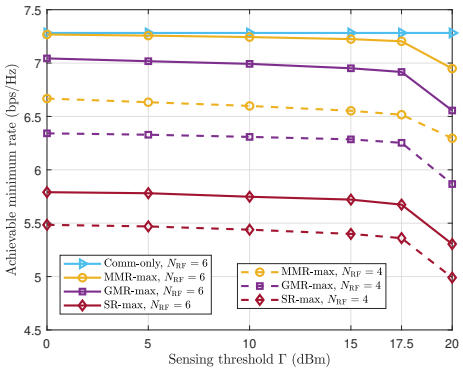

FIGURE 10. Achievable minimum rate versus sensing threshold Γ.

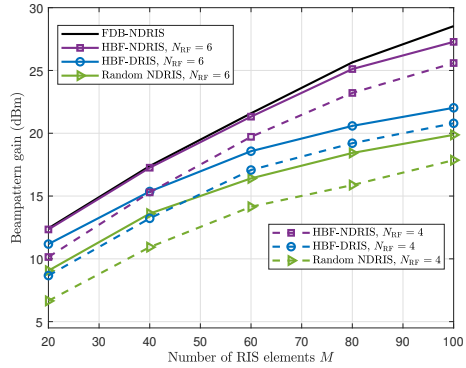

FIGURE 12. Sum beampattern gain versus the number of reflective elements M.

Furthermore, to evaluate the efficiency of the proposed fairness-based GM-rate maximization scheme, we compare it against the following benchmark schemes:

- **Scheme 1 (Comm-only)**: The available transmit power is allocated exclusively to the CUs. Thus, we set $\Gamma = 0$ and optimize the BB TPC $\mathbf{F}_{\mathrm{BB}}$, RF TPC $\mathbf{F}_{\mathrm{RF}}$, and the NDRIS phase shift matrix $\mathbf{\Phi}$ via solving problem $\mathcal{P}_0$.
- **Scheme 3 (MMR-max)**: This scheme aims to maximize the minimum CU rate by solving the following optimization problem:

$$\max_{\mathbf{\Phi},\mathbf{F}_{\mathrm{RF}},\ \mathbf{F}_{\mathrm{BB}}} \quad \min_{k=1,\ldots,K} R_k(\mathbf{\Phi},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}) \tag{41a}$$

$$\text{s. t.} \quad \text{(10b),(10c),(10d), and (10e).} \tag{41b}$$

  Problem (41) is solved by reformulating it as a feasibility problem via the introduction of an auxiliary variable for $R_k(\mathbf{\Phi},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}})$, and subsequently applying the MM-BCD algorithm for the joint optimization of the HTPCs and the NDRIS phase shift matrix, along with the auxiliary variable, as discussed in [46].
- **Scheme 4 (SR-max)**: This scheme optimizes the HPTCs and NDRIS phase shift matrix to maximize the overall sum-rate of the system. The corresponding optimization problem is formulated as:

$$\max_{\mathbf{\Phi},\mathbf{F}_{\mathrm{RF}},\ \mathbf{F}_{\mathrm{BB}}} \quad \sum_{k=1}^{K} R_k(\mathbf{\Phi},\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}) \tag{42a}$$

$$\text{s. t.} \quad \text{(10b),(10c),(10d), and (10e).} \tag{42b}$$

  The SR-max problem is solved using the proposed MM-Alt algorithm by fixing the CU weights as $\eta_k^{(\kappa)} = 1,\ \forall k$.

In Fig. 10, we plot the achievable minimum rate of the system. It can be observed that the MMR-max scheme achieves the highest minimum rate, as it explicitly maximizes the minimum CU rate through targeted power allocation. Importantly, the proposed GM-rate maximization achieves a minimum rate that is very close to that of the optimal MMR-max scheme, while significantly outperforming the SR-max scheme for both $N_{\mathrm{RF}} = \{6,\ 4\}$. These results verify the effectiveness of the GM-rate maximization framework for achieving rate fairness in mmWave ISAC systems.

C. Sensing performance

To examine the sensing performance of the RT, Fig. 11 depicts the transmit beampattern of the proposed NDRIS-HBF scheme having a fixed transmit power and fixed number of reflective elements for the different values of sensing

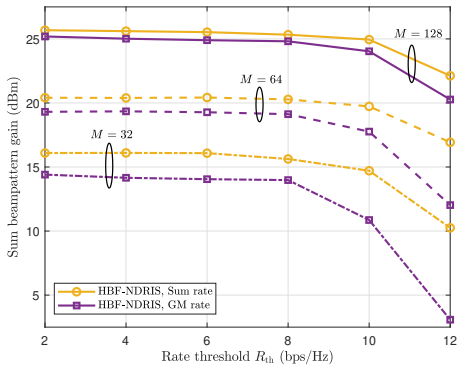

FIGURE 13. **Sum beampattern gain versus rate threshold.**

thresholds $\Gamma = \{15, 20, 25\}$ dBm. As seen from the figure, the main lobes of the transmit beampattern point towards the RTs, and thereby the proposed scheme is able to clearly sense the three RTs located at $-40°, -20°$, and $0°$. Furthermore, the uncertainty observed between $20°$ and $80°$ corresponds to the angular region containing randomly placed CUs, leading to variation in the beampattern gain. Moreover, observe that the gain of the main lobes pointing towards the RTs increases upon increasing the sensing threshold Γ. This is because of the higher power radiated towards the RTs required for meeting the increased sensing requirements.

To further investigate the sensing performance of the system, we evaluate the sum beampattern gain of the RTs by considering the following optimization problem

$$\mathcal{P}_4: \quad \max_{\mathbf{F}_{\rm RF},\, \mathbf{F}_{\rm BB},\, \boldsymbol{\Theta}} \sum_{l=1}^{L} P(\varphi_l) \tag{43a}$$

$$\text{s. t.} \quad \left(\prod_{k \in \mathcal{K}} R_k(\boldsymbol{\Theta}, \mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}) \right)^{\frac{1}{K}} \geq R_{\rm th}, \tag{43b}$$

$$\text{(10c), (10d), and (10e)}, \tag{43c}$$

where $R_{\rm th}$ denotes the rate thresholdfor the communication performance of the system. To solve $\mathcal{P}_4$, we begin by transforming the intractable GM rate constraint (43b) to the weighted sum rate constraint. Subsequently, we adopt the MM-BCD approach to optimize the hybrid TPCs and passive NDRIS phase shift matrix, wherein three subproblems are formulated corresponding to the BB TPC, RF TPC, and NDRIS phase shift matrix, which are then solved via the BCD method. Furthermore, the MM method transforms the non-convex rate expression involved in the weighted sum rate constraint to an equivalent convex form followed by exploiting the PRMO and RCG-based algorithms to design the RF TPC and the RIS phase shift matrices, respectively.

In Fig. 12, we plot the sum beampattern gain versus the number of reflective elements for a fixed value of the rate threshold $R_{\rm th} = 10$ bps/Hz. Although the RIS is not involved in sensing the RTs, the sum beampattern gain of the system increases upon increasing the number of reflective elements, since a lower portion of the available power is used for the CUs due to the the RIS. As a result, a large fraction of power is radiated in the direction of the RTs, which increases the sum beampattern gain of the system. Furthermore, the HBF-NDRIS scheme outperforms the benchmark schemes, thanks to the greater passive beamforming gain offered by the NDRIS.

Fig. 13 depicts the sum beampattern gain versus the rate threshold of the HBF-NDRIS scheme for fixed values of $N_{\rm RF} = 6$ and transmit power $P_{\rm t} = 30$ dBm. Furthermore, we compare the proposed GM rate-based approach to the sum rate-based method, wherein we obtain the HTPCs and NDRIS phase shift matrix for the sum rate by considering the weighing factor $\eta^{(\kappa)}k$ as unity in each iteration. As seen from the figure, the sum beampattern gain decreases upon increasing rate threshold $R_{\rm th}$, since more transmit power is allocated towards satisfying the higher rate requirements of the CUs, thereby leading to a reduction in the power radiated towards the RTs. This observation illustrates the inherent trade-off between sensing accuracy and communication rate. The GM rate-based approach ensures rate fairness by allocating a larger fraction of power to CUs with weaker channel conditions, which further limits the power available for sensing, resulting in a lower sum beampattern gain compared to the sum rate-based method. Nevertheless, the GM rate-based approach attains comparable sensing performance to the sum rate-based design when the number of RIS elements is increased to $M = 128$, highlighting the advantage of incorporating NDRIS. These results confirm that the proposed HBF-NDRIS scheme is capable of simultaneously achieving rate fairness among the CUs and near-optimal sensing performance, thereby effectively balancing the trade-off between communication and sensing objectives.

V. CONCLUSIONS

We proposed a framework for joint HBF and RIS coefficient matrix design in an RIS-aided mmWave ISAC system using the NDRIS phase shift matrix, enabling multi-user communication via reflected links, while achieving target sensing. In contrast to existing communication-centric designs focused on sum rate maximization, we proposed maximizing the GM rate of CUs, while meeting the radar sensing requirements for each RT, adhering to the transmit power budget, the UM constraints of the RF TPC, and the NDRIS phase shift matrix. To obtain a solution to this challenging problem, we first converted the intractable GM rate problem to an equivalent tractable form via the MM framework. Thereafter, we successfully developed an efficient iterative algorithm based on the BCD framework via the construction of the corresponding minorant quadratic function for the objective function at each stage of optimization of the BB TPC, RF TPC, and RIS coefficients. Finally, simulation results were presented, which vindicated the fact that employing an NDRIS provides more control over the propagation paths, hence leading to higher communication performance

than the conventional RIS, all while meeting the sensing requirements. Furthermore, achieving both balanced sum and minimum rates by maximizing the GM rate without the need to impose additional requirements on the CU rates.

Appendix A
DERIVATION OF TRANSFORMATION (16)

Upon defining the quantities $\bar{u} = \left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},k}^{(\kappa)}$ and $\bar{v} = \sum_{j\neq k}^{K} \left|\left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},j}^{(\kappa)}\right|^2 + \sigma_k^2$, a quadratic minorizing function at point $\mathbf{F}_{\mathrm{BB}}^{(\kappa)}$ corresponding to the objective function (16a) is constructed as follows

$$\sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \Big(2\Re\Big\{a_k^{(\kappa)} \left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},k}\Big\} - b_k^{(\kappa)} \sum_{j=1}^{K} \left|\left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},j}\right|^2 + c_k^{(\kappa)}\Big), \tag{44}$$

where

$$a_k^{(\kappa)} = \frac{\gamma_k^{(\kappa)}\left(\mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right)}{\left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},k}^{(\kappa)}}, \tag{45a}$$

$$b_k^{(\kappa)} = \frac{\gamma_k^{(\kappa)}\left(\mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right)}{\sum_{j=1}^{K}\left|\left(\mathbf{h}_k^{(\kappa)}\right)^H \mathbf{f}_{\mathrm{BB},j}\right|^2 + \sigma_k^2}, \tag{45b}$$

$$c_k^{(\kappa)} = \ln\left(1+\gamma_k^{(\kappa)}\left(\mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right)\right) - \gamma_k^{(\kappa)}\left(\mathbf{F}_{\mathrm{BB}}^{(\kappa)}\right) - b_k^{(\kappa)}\sigma_k^2. \tag{45c}$$

Appendix B
DERIVATION OF TRANSFORMATION (21)

The quadratic minorizing function of $\gamma_k^{(\kappa)}(\boldsymbol{\phi})$ at point $\boldsymbol{\phi}^{(\kappa)}$ is formed by considering $\bar{u} = \left(\widetilde{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \boldsymbol{\phi}^{(\kappa)}$ and $\bar{v} = \sum_{j\neq k}^{K}\left|\left(\widetilde{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \boldsymbol{\phi}^{(\kappa)}\right|^2 + \sigma_k^2$ as follows

$$\sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \Big(2\Re\Big\{\tilde{a}_k^{(\kappa)} \left(\widetilde{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \boldsymbol{\phi}\Big\} - \tilde{b}_k^{(\kappa)} \sum_{j=1}^{K} \left|\left(\widetilde{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \boldsymbol{\phi}\right|^2 + \tilde{c}_k^{(\kappa)}\Big), \tag{46}$$

where

$$\tilde{a}_k^{(\kappa)} = \frac{\gamma_k^{(\kappa)}\left(\boldsymbol{\phi}^{(\kappa)}\right)}{\left(\widetilde{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \boldsymbol{\phi}^{(\kappa)}}, \tag{47a}$$

$$\tilde{b}_k^{(\kappa)} = \frac{\gamma_k^{(\kappa)}\left(\boldsymbol{\phi}^{(\kappa)}\right)}{\sum_{j=1}^{K}\left|\left(\widetilde{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \boldsymbol{\phi}^{(\kappa)}\right|^2 + \sigma_k^2} \tag{47b}$$

$$\tilde{c}_k^{(\kappa)} = \ln\left(1+\gamma_k^{(\kappa)}\left(\boldsymbol{\phi}^{(\kappa)}\right)\right) - \gamma_k^{(\kappa)}\left(\boldsymbol{\phi}^{(\kappa)}\right) - \tilde{b}_k^{(\kappa)}\sigma_k^2. \tag{47c}$$

Subsequently, we rewrite (46) in a compact form as follows

$$2\Re\Big\{\left(\mathbf{q}^{(\kappa)}\right)^H \boldsymbol{\phi}\Big\} - \boldsymbol{\phi}^H \mathbf{T}^{(\kappa)} \boldsymbol{\phi} + \tilde{c}^{(\kappa)}, \tag{48}$$

where

$$\mathbf{T}^{(\kappa)} = \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \tilde{b}_k^{(\kappa)} \sum_{j=1}^{K} \widetilde{\mathbf{h}}_{k,j}^{(\kappa)} \left(\widetilde{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H, \tag{49a}$$

$$\mathbf{q}^{(\kappa)} = \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \tilde{a}_k^{(\kappa)} \widetilde{\mathbf{h}}_{k,k}^{(\kappa)} \tag{49b}$$

$$\tilde{c}^{(\kappa)} = \sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \tilde{c}_k^{(\kappa)}. \tag{49c}$$

In a similar fashion, applying the matrix transformation $\mathrm{tr}\left(\mathbf{X}^H\mathbf{Y}\mathbf{X}\mathbf{Z}\right) = \mathrm{vec}(\mathbf{X})^H\left(\mathbf{Z}^T\otimes\mathbf{Y}\right)\mathrm{vec}(\mathbf{X})$, the constraints (21b) and (21c) can be transformed as follows

$$\boldsymbol{\phi}^H \left(\left(\mathbf{B}^{(\kappa)}\right)^T \otimes \mathbf{A}_l^{(\kappa)}\right) \boldsymbol{\phi} \geq \Gamma_l, \ \forall l \in \mathcal{L}, \tag{50a}$$

$$\boldsymbol{\phi}^H \left(\left(\mathbf{B}^{(\kappa)}\right)^T \otimes \mathbf{I}_{N_{\mathrm{t}}}\right) \boldsymbol{\phi} \leq P_{\mathrm{t}}. \tag{50b}$$

Appendix C
DERIVATION OF TRANSFORMATION (34)

The quadratic minorizing function of (34a) at point $\tilde{\boldsymbol{\theta}}^{(\kappa)}$ is constructed by considering $\bar{u} = \left(\widehat{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \tilde{\boldsymbol{\theta}}^{(\kappa)}$ and $\bar{v} = \sum_{j\neq k}^{K}\left|\left(\widehat{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \tilde{\boldsymbol{\theta}}^{(\kappa)}\right|^2 + \sigma_k^2$ as follows

$$\sum_{k\in\mathcal{K}} \eta_k^{(\kappa)} \Big(2\Re\Big\{\hat{a}_k^{(\kappa)} \left(\widehat{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \tilde{\boldsymbol{\theta}}\Big\} - \hat{b}_k^{(\kappa)} \sum_{j=1}^{K} \left|\left(\widehat{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \tilde{\boldsymbol{\theta}}\right|^2 + \hat{c}_k^{(\kappa)}\Big), \tag{51}$$

where $\hat{a}_k^{(\kappa)}, \hat{b}_k^{(\kappa)}$, and $\hat{c}_k^{(\kappa)}$ are given by

$$\hat{a}_k^{(\kappa)} = \frac{\gamma_k^{(\kappa)}\left(\tilde{\boldsymbol{\theta}}^{(\kappa)}\right)}{\left(\widehat{\mathbf{h}}_{k,k}^{(\kappa)}\right)^H \tilde{\boldsymbol{\theta}}^{(\kappa)}}, \tag{52a}$$

$$\hat{b}_k^{(\kappa)} = \frac{\gamma_k^{(\kappa)}\left(\tilde{\boldsymbol{\theta}}^{(\kappa)}\right)}{\sum_{j=1}^{K}\left|\left(\widehat{\mathbf{h}}_{k,j}^{(\kappa)}\right)^H \tilde{\boldsymbol{\theta}}^{(\kappa)}\right|^2 + \sigma_k^2}, \tag{52b}$$

$$\hat{c}_k^{(\kappa)} = \ln\left(1+\gamma_k^{(\kappa)}\left(\tilde{\boldsymbol{\theta}}^{(\kappa)}\right)\right) - \gamma_k^{(\kappa)}\left(\tilde{\boldsymbol{\theta}}^{(\kappa)}\right) - \hat{b}_k^{(\kappa)}\sigma_k^2. \tag{52c}$$

REFERENCES

[1] J. Singh *et al.*, "Pareto optimal hybrid beamforming for short-packet millimeter-wave integrated sensing and communication," *IEEE Trans. Commun.*, pp. 1–1, 2024.

[2] Z. Wang *et al.*, "Multi-vehicle tracking and ID association based on integrated sensing and communication signaling," *IEEE Wireless Commun. Lett.*, vol. 11, no. 9, pp. 1960–1964, 2022.

[3] J. Singh *et al.*, "Multi-beam object-localization for millimeter-wave ISAC-aided connected autonomous vehicles," *IEEE Trans. Veh. Technol.*, pp. 1–6, 2024.

[4] Y. Cui, H. Ding, S. Ke, and L. Zhao, "Integrated sensing and communication in mmWave wireless backhaul networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 5, pp. 6455–6469, 2024.

[5] Z. Cheng, Z. He, and B. Liao, "Hybrid beamforming for multi-carrier dual-function radar-communication system," *IEEE Trans. Cognitive Commun. and Networking*, vol. 7, no. 3, pp. 1002–1015, 2021.

[6] F. Gao, L. Xu, and S. Ma, "Integrated sensing and communications with joint beam-squint and beam-split for mmWave/THz massive MIMO," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2963–2976, 2023.

[7] I. Ahmed *et al.*, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 4, pp. 3060–3097, 2018.

[8] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, 2014.

[9] S. Gong, C. Xing, P. Yue, L. Zhao, and T. Q. S. Quek, "Hybrid analog and digital beamforming for RIS-assisted mmWave communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1537–1554, 2023.

[10] L. Chen, X. Yuan, and Y.-J. A. Zhang, "Coverage analysis of RIS-assisted mmwave cellular networks with 3D beamforming," *IEEE Trans. Commun.*, vol. 72, no. 6, pp. 3618–3633, 2024.

[11] S. Zargari, A. Khalili, and R. Zhang, "Energy efficiency maximization via joint active and passive beamforming design for multiuser MISO IRS-aided SWIPT," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 557–561, 2021.

[12] Q. Li *et al.*, "Reconfigurable intelligent surfaces relying on non-diagonal phase shift matrices," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6367–6383, 2022.

[13] H. Li, S. Shen, M. Nerini, and B. Clerckx, "Reconfigurable intelligent surfaces 2.0: Beyond diagonal phase shift matrices," *IEEE commun. magazine*, 2023.

[14] J. Singh, S. Srivastava, A. K. Jagannatham, and L. Hanzo, "Joint transceiver and reconfigurable intelligent surface design for multiuser mmWave MIMO systems relying on non-diagonal phase shift matrices," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2897–2912, 2023.

[15] R. Li, B. Guo, M. Tao, Y.-F. Liu, and W. Yu, "Joint design of hybrid beamforming and reflection coefficients in RIS-aided mmWave MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2404–2416, 2022.

[16] J. Zhu, Q. Li, H. Chen, and H. V. Poor, "Statistical CSI based hybrid mmWave MIMO-NOMA with max-min fairness," in *ICC 2021 - IEEE International Conf. Commun.*, 2021, pp. 1–6.

[17] T. Yan, C. Li, and J. Shao, "Design of hybrid precoding for millimeter-wave MIMO-NOMA systems based on user fairness," *IEEE Commun. Lett.*, vol. 26, no. 11, pp. 2794–2798, 2022.

[18] H. Yu *et al.*, "Maximizing the geometric mean of user-rates to improve rate-fairness: Proper vs. improper Gaussian signaling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 295–309, 2021.

[19] W. Zhu *et al.*, "Low-complexity Pareto-optimal 3D beamforming for the full-dimensional multi-user massive MIMO downlink," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 8869–8885, 2023.

[20] Y. Chen *et al.*, "Maximizing the geometric mean of user-rates to improve rate-fairness in double RIS-assisted system," in *2022 11th Int. Conf. Control Autom. Inf. Sci. (ICCAIS)*. IEEE, 2022, pp. 203–208.

[21] Y. Liu *et al.*, "A shared cluster-based stochastic channel model for integrated sensing and communication systems," *IEEE Trans. Veh. Technol.*, vol. 73, no. 5, pp. 6032–6044, 2024.

[22] Z. Xiao, S. Chen, and Y. Zeng, "Simultaneous multi-beam sweeping for mmWave massive MIMO integrated sensing and communication," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 8141–8152, 2024.

[23] C. Qi, W. Ci, J. Zhang, and X. You, "Hybrid beamforming for millimeter wave MIMO integrated sensing and communications," *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1136–1140, 2022.

[24] J. Gong, W. Cheng, F. Chen, and J. Wang, "Hybrid beamforming for millimeter wave integrated sensing and communications," in *ICC 2023 - IEEE International Conf. Commun.*, 2023, pp. 2958–2963.

[25] X. Wang, Z. Fei, J. A. Zhang, and J. Xu, "Partially-connected hybrid beamforming design for integrated sensing and communication systems," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6648–6660, 2022.

[26] B. Liu *et al.*, "Optimal beamformer design for millimeter wave dual-functional radar-communication based V2X systems," *IEEE J. Sel. Areas Commun. Communications*, vol. 40, no. 10, pp. 2980–2993, 2022.

[27] Z. Du, F. Liu, W. Yuan, C. Masouros, Z. Zhang, S. Xia, and G. Caire, "Integrated sensing and communications for V2I networks: Dynamic predictive beamforming for extended vehicle targets," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3612–3627, 2023.

[28] A. Chowdary, A. Bazzi, and M. Chafii, "On hybrid radar fusion for integrated sensing and communication," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8984–9000, 2024.

[29] J. Singh *et al.*, "Joint hybrid transceiver and reflection matrix design for RIS-aided mmWave MIMO cognitive radio systems," *IEEE Trans. Cognitive Commun. and Networking*, pp. 1–1, 2024.

[30] J. Singh, S. Srivastava, S. P. Yadav, A. K. Jagannatham, and L. Hanzo, "Energy efficiency optimization in reconfigurable intelligent surface aided hybrid multiuser mmWave MIMO systems," *IEEE Open J. Veh. Technol.*, 2023.

[31] Q. Li, M. El-Hajjar, and L. Hanzo, "Ergodic spectral efficiency analysis of intelligent omni-surface aided systems suffering from imperfect CSI and hardware impairments," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 5073–5086, 2024.

[32] Q. Li *et al.*, "Achievable rate analysis of the STAR-RIS-aided NOMA uplink in the face of imperfect CSI and hardware impairments," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 6100–6114, 2023.

[33] ——, "Performance analysis of active RIS-aided systems in the face of imperfect CSI and phase shift noise," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8140–8145, 2023.

[34] S. Shen, B. Clerckx, and R. Murch, "Modeling and architecture design of reconfigurable intelligent surfaces using scattering parameter network analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1229–1243, 2021.

[35] G. Rebeiz and J. Muldavin, "RF MEMS switches and switch circuits," *IEEE Microwave Magazine*, vol. 2, no. 4, pp. 59–71, 2001.

[36] G. Lavigne and C. Caloz, "Nonreciprocal phase gradient metasurface: Principle and transistor implementation," in *2019 Thirteenth Int. Congr. Artif. Mater. Novel Wave Phenomena (Metamaterials)*, 2019, pp. X–212–X–214.

[37] W. Lyu, S. Yang, Y. Xiu, Y. Li, H. He, C. Yuen, and Z. Zhang, "CRB minimization for RIS-aided mmWave integrated sensing and communications," *IEEE Internet Things J.*, 2024.

[38] W. Hao, Y. Qu, S. Zhou, F. Wang, Z. Lu, and S. Yang, "Joint beamforming design for hybrid RIS-assisted mmWave ISAC system relying on hybrid precoding structure," *IEEE Internet Things J.*, 2024.

[39] A. Bazzi and M. Chafii, "Low dynamic range for RIS-aided bistatic integrated sensing and communication," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 3, pp. 912–927, 2025.

[40] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, 2016.

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

[42] A. Bazzi *et al.*, "A comparative study of sparse recovery and compressed sensing algorithms with application to AoA estimation," in *Proc. 2016 IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2016, pp. 1–5.

[43] A. Garg *et al.*, "Angularly sparse channel estimation in dual- wideband tera-hertz (THz) hybrid MIMO systems relying on bayesian learning," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4384–4400, 2024.

[44] O. E. Ayach *et al.*, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[45] J. Singh *et al.*, "Hybrid transceiver design and optimal power allocation for the cognitive mmWave multiuser MIMO downlink relying on limited feedback," *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 241–256, 2023.

[46] W. Zhu, H. D. Tuan, E. Dutkiewicz, Y. Fang, and L. Hanzo, "Low-complexity pareto-optimal 3D beamforming for the full-dimensional multi-user massive MIMO downlink," *IEEE Trans. Veh. Technol.*, 2023.

Jitendra Singh (Graduate Student Member, IEEE) received the Integrated Dual Degree (B.Tech. and M.Tech.) in Electronics and Communication Engineering, with a specialization in Wireless Communication and Networks, from Gautam Buddha University, Greater Noida, India, in 2017. He is currently pursuing a Ph.D. in the Department of Electrical Engineering at the Indian Institute of Technology (IIT) Kanpur, India. His research interests lie at the intersection of emerging wireless technologies and signal processing, including Cognitive Radio Networks, millimeter-Wave (mmWave) Communication, Intelligent Reflecting Surfaces (IRS), and Integrated Sensing and Communication (ISAC) systems. As part of his doctoral work, he has contributed to several projects focusing on next-generation wireless system design and optimization. He is also a member of the research team that won the prestigious Qualcomm 6G University Research India Program in 2023, recognizing innovative contributions to 6G wireless technologies.

Aditya K. Jagannatham (Senior Member, IEEE) received the bachelor's degree from the Indian Institute of Technology, Bombay, and the M.S. and Ph.D. degrees from the University of California at San Diego, San Diego, CA, USA. From April 2007 to May 2009, he was employed as a Senior Wireless Systems Engineer with Qualcomm Inc., San Diego, where he was a part of the Qualcomm CDMA Technologies Division. He is currently a Professor with the Department of Electrical Engineering, IIT Kanpur, where he also holds the Arun Kumar Chair Professorship. His research interests include next-generation wireless cellular and WiFi networks, with a special emphasis on various 5G technologies, such as massive MIMO, mmWave MIMO, FBMC, NOMA, as well as emerging 6G technologies, such as OTFS, IRS, THz systems, and VLC. He has been twice awarded the P. K. Kelkar Young Faculty Research Fellowship for excellence in research, received multiple Qualcomm Innovation Fellowships in 2018, 2022 and 2024, the IIT Kanpur Excellence in Teaching Award, IIT Kanpur Distinguished Teacher Award, the CAL(IT)2 Fellowship at the University of California at San Diego, the Upendra Patel Achievement Award at Qualcomm San Diego and the Qualcomm 6G UR India Gift.

Lajos Hanzo ((FIEEE'04)) received Honorary Doctorates from the Technical University of Budapest (2009) and Edinburgh University (2015). He is a Foreign Member of the Hungarian Science-Academy, Fellow of the Royal Academy of Engineering (FREng), of the IET, of EURASIP. He holds the IEEE Edwin Howard Armstrong Award and the IEEE Eric Sumner Technical Field Award. For further details, please see http://www-mobile.ecs.soton.ac.uk/,https://en.wikipedia.org/wiki/Lajos_Hanzo.