An Overview of Signal Processing Techniques for RIS/IRS-aided Wireless Systems

Cunhua Pan, Gui Zhou, Kangda Zhi, Sheng Hong, Tuo Wu, Yijin Pan, Hong Ren, Marco Di Renzo, A. Lee Swindlehurst, Rui Zhang, et al.

An Overview of Signal Processing Techniques for RIS/IRS-aided Wireless Systems

Cunhua Pan, Gui Zhou, Kangda Zhi, Sheng Hong, Tuo Wu, Yijin Pan, Hong Ren, Marco Di Renzo, *Fellow, IEEE*, A. Lee Swindlehurst, *Fellow, IEEE*, Rui Zhang, *Fellow, IEEE*, and Angela Yingjun Zhang, *Fellow, IEEE*

***Abstract*—In the past as well as present wireless communication systems, the wireless propagation environment is regarded as an uncontrollable black box that impairs the received signal quality, and its negative impacts are compensated for by relying on the design of various sophisticated transmission/reception schemes. However, the improvements through applying such schemes operating at two endpoints (i.e., transmitter and receiver) only are limited even after five generations of wireless systems. Reconfigurable intelligent surface (RIS) or intelligent reflecting surface (IRS) have emerged as a new and revolutionary technology that can configure the wireless environment in a favorable manner by properly tuning the phase shifts of a large number of quasi passive and low-cost reflecting elements, thus standing out as a promising candidate technology for the next-/sixth-generation (6G) wireless system. However, to reap the performance benefits promised by RIS/IRS, efficient signal processing techniques are crucial, for a variety of purposes such as channel estimation, transmission design, radio localization, and so on. In this paper, we provide a comprehensive overview of recent advances on RIS/IRS-aided wireless systems from the signal processing perspective. We also highlight promising research directions that are worthy of investigation in the future.**

***Index Terms*—Reconfigurable Intelligent Surface (RIS), Intelligent Reflecting Surface (IRS), channel estimation, transmission design, localization, wireless systems, 6G.**

I. Introduction

While the fifth-generation (5G) wireless communication system is under deployment worldwide, research interest has shifted to the future sixth-generation (6G) wireless system [1]–[3], which targets supporting not only cutting-edge applications like multisensory augmented/virtual reality applications, wireless brain computer interactions, and fully autonomous systems, but also the wireless evolution from "connected things" to "connected intelligence". The required key performance indicators (KPIs), including data rates, reliability, latency, spectrum/energy efficiency, and connection density, will be superior to those for 5G. For example, the energy and spectrum efficiency for 6G are expected to be 10-100 times and 5 times better than those of 5G, respectively. These KPIs, however, cannot be fully achieved by the existing three-pillar 5G physical layer techniques [4], which include massive multiple-input multiple-output (MIMO), millimeter wave (mmWave) communications, and ultra-dense heterogeneous networks. In particular, a large number of antennas along with active radio frequency (RF) chains are needed for massive MIMO to achieve high spectrum efficiency, which leads to high energy consumption and hardware cost. Moreover, moving to the mmWave frequency band renders the electromagnetic waves more susceptible to blockage by obstacles such as furniture and walls in indoor scenarios. In addition, more costly RF chains and sophisticated hybrid precoding are necessary for mmWave communication systems. The dense deployment of small base stations (BSs) also incurs high maintenance cost, network energy consumption, and hardware cost due to high-speed backhaul links. Furthermore, sophisticated interference management techniques are necessary in ultra-dense networks.

Conventionally, the wireless environment is perceived as a randomly varying entity that impairs the signal quality due to uncontrolled reflections, refractions and unexpected interference. Although a plethora of physical layer techniques such as advanced modulation/demodulation and precoding/decoding schemes have been developed at the endpoints of communication links to compensate for these negative impacts, it is undeniable that a certain level of saturation has been reached in terms of achievable data rate and performance reliability. Huge performance gains are expected when regarding the wireless environment as an additional variable to optimize. This is made possible by exploiting the new and revolutionary idea of reconfigurable intelligent surfaces (RISs) or intelligent reflecting surfaces (IRSs) [5]–[9], which are capable of reconfiguring the wireless propagation environment into a transmission medium with more desirable characteristics.

An RIS/IRS is a planar surface composed of a large number of quasi passive and low-cost reflecting elements, each of which can impose an independent phase shift/amplitude on the impinging electromagnetic signals in a fully customized way. Thanks to recent advances in metamaterials [10], the phase shifts imposed on the incident electromagnetic signals can be adjusted in real-time in reaction to the rapid variations

C. Pan is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. He was with the School of Electronic Engineering and Computer Science at Queen Mary University of London, London E1 4NS, U.K. (Email: cunhuapan21@gmail.com). G. Zhou, K. Zhi and T. Wu are with the School of Electronic Engineering and Computer Science at Queen Mary University of London, London E1 4NS, U.K. (Email: {g.zhou, k.zhi, tuo.wu}@qmul.ac.uk). S. Hong is with Information Engineering School of Nanchang University, Nanchang 330031, China. (Email: shenghong@ncu.edu.cn). Y. Pan and H. Ren are the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. (Email: {panyj, hren}@seu.edu.cn). M. Di Renzo is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, 3 Rue Joliot-Curie, 91192 Gif-sur-Yvette, France. (Email: marco.direnzo@centralesupelec.fr). A. L. Swindlehurst is with the Center for Pervasive Communications and Computing, University of California, Irvine, CA 92697, USA. (Email: swindle@uci.edu). R. Zhang is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (Email: elezhang@nus.edu.sg). Y. J. Zhang is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (Email: yjzhang@ie.cuhk.edu.hk).

in the wireless propagation environment. By judiciously tuning the phase shifts of the RIS/IRS, the signals reradiated from the RIS/IRS can be added constructively with the signals from other paths to enhance the received signal power at the desired users, or can be combined destructively to mitigate the undesired signals at unintended users such as multiuser interference and signal leakage at the eavesdroppers. Due to these functionalities, RIS/IRS can be used to extend the coverage area, improve the channel rank, mitigate the interference, enhance the reliability, and improve the positioning accuracy. Unlike conventional relaying techniques, an RIS/IRS is free from RF chains and amplifiers, and thus entails much reduced power consumption and hardware cost. Furthermore, due to their quasi passive nature, RIS/IRS can be fabricated with a low profile, light weight, and limited thickness, which enables them to be readily layered on surfaces available in the environment, including building facades, ceilings, street lamps, and so on.

The appealing advantages of RIS/IRS have led to extensive research focused on its fundamental performance limits [11]–[20], channel modeling [21]–[25], and prototype design [26]–[29]. It was shown in [28] that, in an indoor realistic propagation environment, 26 dB power gain can be achieved by an RIS/IRS prototype consisting of 1100 controllable elements operating in the 5.8 GHz band. In addition, several tutorial/overview papers have summarized recent developments in this area, including the technical challenges [30], operation principles [31], transmission design and applications [32], electromagnetic modeling [33], practical design issues [34] and channel estimation [35].

In this paper, we provide a systematic overview of existing works on IRS/RIS mainly from the signal processing point of view, by focusing on channel estimation, transmission design and radio localization issues. Specifically, in Section II, we overview existing contributions on channel estimation based on the structure of the channels, including unstructured channels that model low-frequency rich-scattering scenarios and structured channels that are appropriate for high-frequency sparse channels. In Section III, we overview existing works on transmission design from two aspects, namely, optimization techniques and the availability of channel state information (CSI). The existing contributions on transmission design are classified into three cases, based on fully instantaneous CSI, two-timescale CSI and fully long-term CSI, respectively. Then, in Section IV, we introduce RIS/IRS-aided localization techniques by differentiating between far-field and near-field channel models. Promising research directions for future work are highlighted in Section V and conclusions are drawn in Section VI.

Notations: $\mathbf{1}_M$ and $\mathbf{0}_M$ are column vectors of all ones and all zeros, respectively. The Hadamard, Kronecker and Khatri-Rao products between two matrices $\mathbf{A}$ and $\mathbf{B}$ are denoted by $\mathbf{A} \odot \mathbf{B}$, $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{A} \diamond \mathbf{B}$, respectively. $\|\mathbf{A}\|_2$ denote the 2-norm of matrix $\mathbf{A}$. $\mathbf{A}_{(:,n)}$ and $\mathbf{A}_{(m,:)}$ denote the n-th column and the m-th row of matrix $\mathbf{A}$, respectively. $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote the conjugate, transpose and Hermitian operators, respectively. Given two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, we define $\langle \mathbf{X}_1, \mathbf{X}_2 \rangle \triangleq \mathrm{Re}\left\{\mathrm{tr}\left(\mathbf{X}_1^{\mathrm{H}}\mathbf{X}_2\right)\right\}$, where $\mathrm{Re}\{\cdot\}$ is the real-value operator. $\mathrm{Im}\{\cdot\}$ is the imaginary-value operator. The symbol $\mathbb{C}$ denotes the complex field, $\mathbb{R}$ represents the real field, and $j \triangleq \sqrt{-1}$ is the imaginary unit.

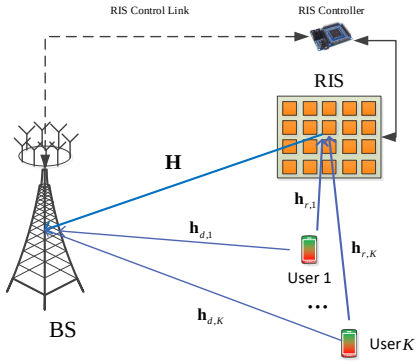

Fig. 1. Illustration of an RIS-aided uplink communication system.

II. Channel Estimation

In this section, we overview existing methods for channel estimation. For illustrative purposes, we focus our attention on the uplink, but similar considerations can be made for the downlink. The uplink channel estimation scenario for a narrow-band RIS-aided[1] multi-user communication system is shown in Fig. 1. The system model consists of one BS with N antennas, one RIS with M reflecting elements, and K single-antenna users. Let $\mathbf{H} \in \mathbb{C}^{N\times M}$ denote the RIS-BS channel, $\mathbf{h}_{r,k} \in \mathbb{C}^{M\times 1}$ the user-RIS channel of user k, and $\mathbf{h}_{d,k} \in \mathbb{C}^{N\times 1}$ the user-BS channel of user k. We assume that all the channels are subject to quasi-static fading and hence the channel coefficients remain constant within one channel coherence interval.

At time slot t, the received baseband signal at the BS is given by

$$\mathbf{y}_t = \sum\nolimits_{k=1}^{K} \sqrt{P_k}(\mathbf{h}_{d,k} + \mathbf{H}\boldsymbol{\Theta}_t\mathbf{h}_{r,k})x_{k,t} + \mathbf{n}_t, \qquad (1)$$

where $x_{k,t} \in \mathbb{C}$ is the pilot signal transmitted by user k satisfying the constraint $|x_{k,t}| = 1$, P_k is the transmit power, $\mathbf{n}_t \sim \mathcal{CN}\left(0, \sigma^2\mathbf{I}_N\right)$ is the additive white Guassian noise (AWGN), and $\boldsymbol{\Theta}_t$ is the reflection coefficient matrix of the RIS. In general, $\boldsymbol{\Theta}_t$ is a diagonal matrix given by $\boldsymbol{\Theta}_t = \mathrm{diag}\left(\boldsymbol{\theta}_t\right)$, where $\boldsymbol{\theta}_t = [\theta_{t,1}, \cdots, \theta_{t,M}]^{\mathrm{T}}$ is the corresponding reflection coefficients at the RIS with $\theta_{t,m} = \alpha_m e^{j\varphi_{t,m}}$ being the reflection coefficient corresponding to the m-th reflecting element. Here, α_m and $\varphi_{t,m} \in [0, 2\pi)$ are the amplitude and the phase shift of the m-th element of the RIS, respectively. For simplicity, the reflection amplitude is assumed to be $\alpha_m = 1, \forall m$.[2]

A key property for channel estimation in RIS-aided systems is that there is a scaling ambiguity issue that prevents the RIS-BS channel and the user-RIS channel from being individually

[1] In the remaining parts, we use RIS to represent RIS/IRS for notational simplicity.

[2] Reflection amplitude variation can be exploited for further enhancing the multiuser communication performance, especially when the CSI is not perfectly estimated in practice [36].

identifiable. Specifically, for any invertible $M \times M$ diagonal matrix $\mathbf{\Lambda}$, we have

$$\mathbf{H}\mathbf{\Theta}_t\mathbf{h}_{r,k} = \mathbf{H}\mathbf{\Lambda}\mathbf{\Theta}_t\mathbf{\Lambda}^{-1}\mathbf{h}_{r,k} = \tilde{\mathbf{H}}\mathbf{\Theta}_t\tilde{\mathbf{h}}_{r,k}, \tag{2}$$

where $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{\Lambda}$ and $\tilde{\mathbf{h}}_{r,k} = \mathbf{\Lambda}^{-1}\mathbf{h}_{r,k}$. Hence, even if $\mathbf{H}\mathbf{\Theta}_t\mathbf{h}_{r,k}$ can be estimated based on (1), one is still not able to extract the individual channels $\mathbf{H}$ and $\mathbf{h}_{r,k}$. Fortunately, there is generally no need to address this ambiguity issue when designing the phase shift matrix at the RIS for data transmission without loss of optimality.

Denote the cascaded channel for user k as $\mathbf{G}_k = \mathbf{H}\mathrm{diag}\left(\mathbf{h}_{r,k}\right) \in \mathbb{C}^{N\times M}$. Then, the received signal in (1) can be rewritten as

$$\mathbf{y}_t = \sum\nolimits_{k=1}^{K} \sqrt{P_k}(\mathbf{h}_{d,k} + \mathbf{G}_k\boldsymbol{\theta}_t)x_{k,t} + \mathbf{n}_t. \tag{3}$$

Hence, as will be seen in Section III, the cascaded channel $\mathbf{G}_k$ and the direct channel $\mathbf{h}_{d,k}$ are sufficient for designing RIS-aided communications. As a result, most of the existing contributions have focused on designing algorithms to estimate the cascaded channels and the direct channel $\mathbf{h}_{d,k}$ separately[3].

In the following, we review the state-of-art techniques based on the structure of the channel models, namely, unstructured channel models and structured channel models.

A. Unstructured Channel Models

In this subsection, we consider the case when the channels are characterized by rich scattering. This is often the case in sub-6 GHz communication systems, where the propagation environment cannot be efficiently parameterized. In the following, we first consider the simple single-user case, and then address the multiuser case.

*1) **Single-user Case**:* Since a single user is considered, the user index is omitted in the following derivations and the received signal model in (3) becomes

$$\mathbf{y}_t = \sqrt{P}(\mathbf{h}_d + \mathbf{G}\boldsymbol{\theta}_t)x_t + \mathbf{n}_t. \tag{4}$$

The m-th column of $\mathbf{G}$ is denoted by $\mathbf{g}_m$. Define $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_M]$ and $\mathbf{h}_r = \left[h_r^1; \cdots; h_r^M\right]$. Then, $\mathbf{g}_m = h_r^m\mathbf{h}_m$. By defining the overall channel as $\mathbf{c} = \left[\mathbf{h}_d^{\mathrm{T}}, \mathbf{g}_1^{\mathrm{T}}, \cdots, \mathbf{g}_M^{\mathrm{T}}\right]^{\mathrm{T}}$, (4) can be rewritten as

$$\mathbf{y}_t = \sqrt{P}\left(x_t\left[1, \boldsymbol{\theta}_t^{\mathrm{T}}\right] \otimes \mathbf{I}_N\right)\mathbf{c} + \mathbf{n}_t. \tag{5}$$

Assume that T time slots are used for channel training, i.e, T time slots are reserved for estimating the end-to-end channel, and define

$$\mathbf{\Phi} = \begin{bmatrix} 1, \boldsymbol{\theta}_1^{\mathrm{T}} \\ \cdots \\ 1, \boldsymbol{\theta}_T^{\mathrm{T}} \end{bmatrix} \in \mathbb{C}^{T\times(M+1)}, \mathbf{\Xi} = \mathbf{\Phi} \otimes \mathbf{I}_N. \tag{6}$$

Stacking the T training time slots together, the overall received signal vector can be expressed as

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{bmatrix} = \sqrt{P}\begin{bmatrix} x_1\left[1, \boldsymbol{\theta}_1^{\mathrm{T}}\right] \otimes \mathbf{I}_N \\ \vdots \\ x_T\left[1, \boldsymbol{\theta}_T^{\mathrm{T}}\right] \otimes \mathbf{I}_N \end{bmatrix}\mathbf{c} + \begin{bmatrix} \mathbf{n}_1 \\ \vdots \\ \mathbf{n}_T \end{bmatrix}$$

$$= \sqrt{P}\mathbf{X}\mathbf{\Xi}\mathbf{c} + \mathbf{n}, \tag{7}$$

where we have defined $\mathbf{X} = \mathrm{diag}\left(\left[x_1\mathbf{1}_N; \cdots; x_T\mathbf{1}_N\right]\right)$ and $\mathbf{n} = \left[\mathbf{n}_1^{\mathrm{T}}, \cdots, \mathbf{n}_T^{\mathrm{T}}\right]^{\mathrm{T}}$. By defining $\mathbf{y} = \left[\mathbf{y}_1^{\mathrm{T}}, \cdots, \mathbf{y}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{Z} = \mathbf{X}\mathbf{\Xi}$, the overall received signal vector in (7) can be written as

$$\mathbf{y} = \sqrt{P}\mathbf{Z}\mathbf{c} + \mathbf{n}. \tag{8}$$

Our aim is to estimate $\mathbf{c}$ based on the pilot matrix $\mathbf{Z}$ that is assumed to be known and depends on the pilot sequence being used for channel estimation. To ensure that $\mathbf{c}$ can be uniquely estimated, $\mathbf{Z}$ must be a full rank matrix. Hence, the number of time slots for channel training need to satisfy the condition $T \geq (M+1)$.

In general, based on (8), there are two common methods for estimating $\mathbf{c}$.

Method I: Least Squares (LS) Estimator

The simplest method to estimate $\mathbf{c}$ is the LS estimator, which is formulated as

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\|\mathbf{y} - \sqrt{P}\mathbf{Z}\mathbf{c}\right\|^2, \tag{9}$$

for which the solution is

$$\hat{\mathbf{c}} = \frac{1}{\sqrt{P}}\left(\mathbf{Z}^{\mathrm{H}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{H}}\mathbf{y}. \tag{10}$$

Assume that the noise vectors are uncorrelated, that is $\mathbf{n} \sim \mathcal{CN}\left(0, \sigma^2\mathbf{I}_{NT}\right)$. Then, the LS channel estimate is equivalent to the maximum likehood (ML) estimate, and it is an unbiased estimator, e.g., $\mathbb{E}\left\{\hat{\mathbf{c}}\right\} = \mathbf{c}$. The error covariance matrix of the estimated channel is equal to the Cramér-Rao bound (CRB):

$$\begin{aligned}\mathbf{R}_e =& \mathbb{E}\left\{\left(\mathbf{c} - \hat{\mathbf{c}}\right)\left(\mathbf{c} - \hat{\mathbf{c}}\right)^{\mathrm{H}}\right\} \\ =& \frac{1}{P}\mathbb{E}\left\{\left(\mathbf{Z}^{\mathrm{H}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{H}}\mathbf{n}\mathbf{n}^{\mathrm{H}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{H}}\mathbf{Z}\right)^{-1}\right\} \\ =& \frac{\sigma^2}{P}\left(\mathbf{Z}^{\mathrm{H}}\mathbf{Z}\right)^{-1} = \frac{\sigma^2}{P}\left(\mathbf{\Xi}^{\mathrm{H}}\mathbf{\Xi}\right)^{-1} = \frac{\sigma^2}{P}\left(\mathbf{\Phi}^{\mathrm{H}}\mathbf{\Phi}\right)^{-1} \otimes \mathbf{I}_N.\end{aligned} \tag{11}$$

Note that the error covariance matrix in (11) is not related to the training signals sent by the users, e.g., $\mathbf{X}$. Therefore, the optimization of the channel estimation boils down to the design of the training phase shift matrix $\mathbf{\Phi}$. In the following, two commonly used schemes are introduced.

- ***On-off Scheme [43], [44]:*** The simplest scheme is the on-off scheme adopted in [43], [44]. The main idea is to switch each reflecting element on and off [43], or to switch groups of reflecting elements on and off [44] to reduce the training overhead. Specifically, in the first time slot, all the reflecting elements are switched off, and the BS estimates the direct channel $\mathbf{h}_d$. Then, in the remaining time slots, only one element (group) is switched on while keeping the others off. In this scheme, the number of time slots for channel training is equal to $T = M+1$. Note that $\theta_{t,m} = 0, 1$ corresponds to the case when the reflecting element is off and on, respectively. Hence, the training phase shift matrix is given by

$$\mathbf{\Phi} = \begin{bmatrix} 1 & \mathbf{0}_M^{\mathrm{T}} \\ \mathbf{1}_M & \mathbf{I}_M \end{bmatrix}. \tag{12}$$

[3]Due to the page limit, we do not discuss methods where the individual channels $\mathbf{H}$ and $\mathbf{h}_{r,k}$ are estimated individually. Interested readers can refer to [37]–[42] for more details.

Then, based on (11), the error covariance matrix is equal to

$$\mathbf{R}_e = \frac{\sigma^2}{P}\left(\mathbf{\Xi}^{\mathrm{H}}\mathbf{\Xi}\right)^{-1} = \frac{\sigma^2}{P}\left(\mathbf{\Phi}^{\mathrm{H}}\mathbf{\Phi}\right)^{-1} \otimes \mathbf{I}_N = \frac{\sigma^2}{P}\begin{bmatrix} 1 & -\mathbf{1}_M^{\mathrm{T}} \\ -\mathbf{1}_M & \mathbf{E}_M \end{bmatrix} \otimes \mathbf{I}_N, \quad (13)$$

where $\mathbf{E}_M = \mathbf{1}_M\mathbf{1}_M^{\mathrm{T}} + \mathbf{I}_M$. Hence, the error variance of the channels is obtained as

$$\mathrm{var}\left(\hat{\mathbf{h}}_d\right) = \frac{\sigma^2}{P}\mathbf{I}_N, \mathrm{var}\left(\hat{\mathbf{g}}_m\right) = \frac{2\sigma^2}{P}\mathbf{I}_N, m = 1, \cdots, M. \quad (14)$$

The factor 2 in the numerator of $\mathrm{var}\left(\hat{\mathbf{g}}_m\right)$ is due to the error propagation of the estimation error of $\mathbf{h}_d$. The main drawback of the on-off scheme is that switching off all the reflecting elements except one would reduce the reflected power, which degrades the received signal-to-noise ratio (SNR).

- ***Discrete Fourier Transform (DFT) Scheme [45]–[48]:*** To enhance the reflected signal power, the discrete DFT training scheme was proposed in [45]. Specifically, the training phase shifts at the RIS are optimized to minimize the mean squared error, and it is demonstrated that the DFT training scheme achieved the optimal performance. Based on the DFT method, the training phase shift matrix $\mathbf{\Phi}$ is equal to the first $M+1$ columns of a $T \times T$ DFT matrix, which is given by

$$[\mathbf{\Phi}]_{t,m} = e^{-j\frac{2\pi(t-1)(m-1)}{T}}, t = 1, \cdots, T, m = 1, \cdots, M.$$

Then, the error covariance matrix is equal to

$$\mathbf{R}_e = \frac{\sigma^2}{P}\left(\mathbf{\Xi}^{\mathrm{H}}\mathbf{\Xi}\right)^{-1} = \frac{\sigma^2}{P}\left(\mathbf{\Phi}^{\mathrm{H}}\mathbf{\Phi}\right)^{-1} \otimes \mathbf{I}_N = \frac{\sigma^2}{TP}\mathbf{I}_{N(M+1)}. \quad (15)$$

Hence, the error variance of the channels is given by

$$\mathrm{var}\left(\hat{\mathbf{h}}_d\right) = \mathrm{var}\left(\hat{\mathbf{g}}_m\right) = \frac{\sigma^2}{TP}\mathbf{I}_N, m = 1, \cdots, M+1. \quad (16)$$

By comparing (16) with (14), the DFT training scheme reduces the channel error by a factor equal to $1/T$ for the direct channel $\hat{\mathbf{h}}_d$ and by a factor equal to $1/(2T)$ for the RIS-reflected channels $\hat{\mathbf{g}}_m, \forall m$. Recently, the DFT training scheme has been extended to the multiple-input multiple-output (MIMO) case in [47] and to the orthogonal frequency division multiplexing (OFDM) system in [48].

- ***Hadamard Matrix [47], [49]:*** The training phase shift matrix can also be designed using the first $M+1$ columns of the $T \times T$ Hadamard matrix [47], where $T = 2$ or T is a multiple of 4. Specifically, one can construct a T-dimensional Hadamard matrix as follows

$$\mathbf{D}_T = \begin{bmatrix} \mathbf{D}_{T/2} & \mathbf{D}_{T/2} \\ \mathbf{D}_{T/2} & -\mathbf{D}_{T/2} \end{bmatrix}, \mathbf{D}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

with $T = 2^n, n = 1, 2, \cdots$. It can be readily verified that $\mathbf{D}_{2^B}^{\mathrm{H}}\mathbf{D}_{2^B} = T\mathbf{I}$. Then, the training phase shift matrix $\mathbf{\Phi}$ can be set equal to the first $M+1$ columns of the matrix $\mathbf{D}_{2^B}$. Then, the error covariance matrix is calculated as in (16), which means that setting the training phase shift matrix as the Hadamard matrix results in the minimum MSE of the estimator. Compared to the DFT matrix, the main advantage is that only two discrete phase shifts, namely $\{0, \pi\}$, are required for channel training, which can reduce the hardware complexity, and is thus an appealing solution from the implementation standpoint.

Method II: Linear Minimum Mean-Squared-Error (LMMSE)

The LS estimators do not exploit prior knowledge of the channel distributions. When such information is available, the optimal estimate of $\mathbf{c}$ that minimizes the MSE, which is defined as $\mathbb{E}\left[\|\mathbf{c} - \hat{\mathbf{c}}\|^2\right]$, is the MMSE estimate, which is given by

$$\hat{\mathbf{c}}_{\mathrm{mmse}} = \mathbb{E}\left[\mathbf{c} | \mathbf{y}\right]. \quad (17)$$

Since $\mathbf{c}$ depends on the cascaded channel that is the product of the user-RIS channel and the RIS-BS channel, the unknown vector $\mathbf{c}$ is, in general, not Gaussian distributed. This means that the posterior distribution $p(\mathbf{c}|\mathbf{y})$ is difficult to obtain, and thus $\hat{\mathbf{c}}_{\mathrm{mmse}}$ cannot be readily calculated in a closed form expression. To address this issue, the LMMSE method was proposed in [50].

Before introducing the LMMSE method, we first provide the distributions of various channels. The most common approach is to assume that the channel coefficients are correlated. Specifically, the three channels in the RIS-aided system model in Fig. 1 can be written as

$$\mathbf{H} = \mathbf{R}_{\mathrm{HB}}^{\frac{1}{2}}\tilde{\mathbf{H}}\mathbf{R}_{\mathrm{HR}}^{\frac{1}{2}}, \mathbf{h}_r = \mathbf{R}_{\mathrm{h_rR}}^{\frac{1}{2}}\tilde{\mathbf{h}}_r, \mathbf{h}_d = \mathbf{R}_{\mathrm{h_dB}}^{\frac{1}{2}}\tilde{\mathbf{h}}_d, \quad (18)$$

where $\mathbf{R}_{\mathrm{HB}}$ and $\mathbf{R}_{\mathrm{h_dB}}$ are the spatial correlation matrices with unit diagonal elements at the BS for channel $\mathbf{H}$ and $\mathbf{h}_r$, respectively, $\mathbf{R}_{\mathrm{h_rR}}$ and $\mathbf{R}_{\mathrm{HR}}$ are the spatial correlation matrices with unit diagonal elements at the RIS for channel $\mathbf{h}_r$ and $\mathbf{H}$, respectively. Then, $\tilde{\mathbf{H}}$, $\tilde{\mathbf{h}}_r$, and $\tilde{\mathbf{h}}_d$ have probability distributions equal to $\tilde{\mathbf{H}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M \otimes \mathbf{I}_N)$, $\tilde{\mathbf{h}}_r \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ and $\tilde{\mathbf{h}}_d \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$, respectively.

Based on the above definitions, using the linear model in (8), the LMMSE of $\mathbf{c}$ is given by [50]

$$\hat{\mathbf{c}} = \mathbb{E}[\mathbf{c}] + \sqrt{P}\mathbf{C}_{\mathbf{cc}}\mathbf{Z}^{\mathrm{H}}\left(P\mathbf{Z}\mathbf{C}_{\mathbf{cc}}\mathbf{Z}^{\mathrm{H}} + \sigma^2\mathbf{I}_{TN}\right)^{-1}(\mathbf{y} - \bar{\mathbf{y}}), \quad (19)$$

where $\mathbf{C}_{\mathbf{cc}} = \mathbb{E}\left[\mathbf{c}\mathbf{c}^{\mathrm{H}}\right]$ and $\bar{\mathbf{y}} = \mathbb{E}\{\mathbf{y}\}$. $\bar{\mathbf{y}}$ can be readily shown to be the zero vector. Since $\mathbf{H}$ and $\mathbf{h}_r$ are independent and have zero mean, we have $\mathbb{E}[\mathbf{c}] = \mathbf{0}$. Hence, the matrix $\mathbf{C}_{\mathbf{cc}}$ can be formulated as

$$\mathbf{C}_{\mathbf{cc}} = \begin{bmatrix} \mathbf{R}_{\mathrm{h_dB}} & \mathbf{0}_{N \times MN} \\ \mathbf{0}_{MN \times N} & (\mathbf{R}_{\mathrm{h_rR}} \odot \mathbf{R}_{\mathrm{HR}}) \otimes \mathbf{R}_{\mathrm{HB}} \end{bmatrix}. \quad (20)$$

Then, the error covariance matrix for the channel estimate is given by [50]

$$\mathbf{R}_e = \left(\mathbf{C}_{\mathbf{cc}}^{-1} + \frac{P}{\sigma^2}\mathbf{Z}^{\mathrm{H}}\mathbf{Z}\right)^{-1} = \left(\mathbf{C}_{\mathbf{cc}}^{-1} + \frac{P}{\sigma^2}\mathbf{\Phi}^{\mathrm{H}}\mathbf{\Phi} \otimes \mathbf{I}_N\right)^{-1}. \quad (21)$$

To minimize the MSE, the majorization-minimization (MM) algorithm was proposed in [50] to optimize the training phase shift matrix $\mathbf{\Phi}$. However, the simulation results in [50] show that the performance of the optimized phase shift matrix is similar to that of the DFT matrix. For this reason, we

assume the DFT-based training phase shift matrix, and the corresponding error covariance matrix of the LMMSE channel estimator is given by

$$\mathbf{R}_e = \left(\mathbf{C}_{\mathbf{cc}}^{-1} + \frac{TP}{\sigma^2} \mathbf{I}_{N(M+1)} \right)^{-1}. \quad (22)$$

As for the special case when all the channels undergo uncorrelated Rayleigh fading, we obtain $\mathbf{C}_{\mathbf{cc}} = \mathbf{I}_{N(M+1)}$ and

$$\mathbf{R}_e = \frac{\sigma^2}{\sigma^2 + TP} \mathbf{I}_{N(M+1)}. \quad (23)$$

Hence, the error variance of the channels is given by

$$\text{var}\left(\hat{\mathbf{h}}_d\right) = \text{var}\left(\hat{\mathbf{g}}_m\right) = \frac{\sigma^2}{\sigma^2 + TP} \mathbf{I}_N, m = 1, \cdots, M. \quad (24)$$

By comparing (24) with (16), we observe that the LMMSE has a smaller estimation error as compared to the LS method.

Training Overhead Analysis: To ensure that $\mathbf{c}$ can be uniquely estimated using the LS or the LMMSE estimators, the number of time slots for channel training needs to fulfill the condition $T \geq (M+1)$. In typical setups, this results in an excessive training overhead, e.g.,when the number of reflecting elements M is large. In this case, the remaining time slots for data transmission are, in fact, significantly reduced.

To reduce the pilot overhead, the authors of [44] and [45] proposed the element grouping (EG) method. In practical RIS-aided communication systems, the channel associated with adjacent RIS reflecting elements may be highly correlated. In such cases, the EG method groups the adjacent elements and assigns the same reflection pattern to them. This scheme is effective when the reflecting elements are installed closely together. Assume that the group size is J. Thus, the number of groups is $M' = M/J$, which is assumed to be an integer. Then, define $\boldsymbol{\theta}_t = \boldsymbol{\theta}'_t \otimes \mathbf{1}_J$, where $\boldsymbol{\theta}'_t \in \mathbb{C}^{M' \times 1}$, so that

$$\mathbf{G}\boldsymbol{\theta}_t = \mathbf{G}\left(\boldsymbol{\theta}'_t \otimes \mathbf{1}_J\right) = \mathbf{G}\left(\mathbf{I}_{M'} \otimes \mathbf{1}_J\right)\boldsymbol{\theta}'_t = \mathbf{G}'\boldsymbol{\theta}'_t, \quad (25)$$

where $\mathbf{G}' = \mathbf{G}\left(\mathbf{I}_{M'} \otimes \mathbf{1}_J\right) \in \mathbb{C}^{N \times M'}$, and each column of $\mathbf{G}'$ is the unit-coefficient combination of the columns of $\mathbf{G}$ corresponding to the group of reflecting elements. Then, the model in (25) has the same form as the original models. Hence, the above mentioned channel estimation methods can be directly applied. In this case, to estimate the direct channel $\mathbf{h}_d$ and the reflected channel $\mathbf{G}'$, the minimum training pilot overhead is $J+1$, which is lower than $M+1$ for the original model. When $\mathbf{G}'$ is estimated, the original channel $\mathbf{G}$ can be approximately recovered as $\mathbf{G} \approx \left(\mathbf{G}' \otimes \mathbf{1}_J^{\mathrm{T}}\right)/J$.

Simulation Results: For illustration, we consider a single-user scenario and compare the MSE performance of various algorithms in Fig. 2. For simplicity, the exponential spatial correlation model is adopted, namely $[\mathbf{R}_{\mathrm{HB}}]_{i,j} = \rho_1^{|i-j|}, [\mathbf{R}_{\mathrm{HR}}]_{i,j} = \rho_2^{|i-j|}, [\mathbf{R}_{\mathrm{h_r R}}]_{i,j} = \rho_3^{|i-j|}, [\mathbf{R}_{\mathrm{h_d B}}]_{i,j} = \rho_4^{|i-j|}$. The number of antennas and reflecting elements are $N = 8$ and $M = 63$, respectively. The SNR is defined as $\gamma = P/\sigma^2$. The simulation results confirm the theoretical results. The LS-on-off scheme provides the worst MSE performance, while the LMMSE DFT approach offers the best estimation accuracy. Note that except for the LS EG DFT approach, the MSE performance of the other LS methods does not depend on the correlation coefficients since they do not exploit the prior knowledge of the spatial correlation. In contrast, the MSE performance of the LMMSE DFT scheme decreases as the correlation increases. It is interesting to observe that the MSE of the LS EG DFT method decreases by increasing the spatial correlation and approaches the MSE of the LS DFT method without grouping, which demonstrates that the EG scheme is effective when the reflecting elements are strongly correlated.

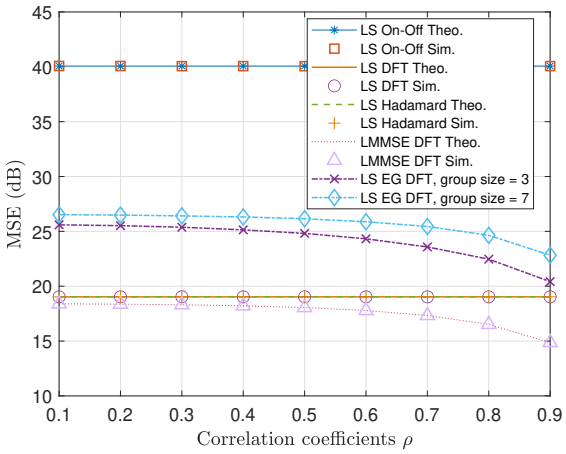

Fig. 2. MSE comparison of different channel estimation schemes under the single-user setup, where $\rho = \rho_1 = \rho_2 = \rho_3 = \rho_4$, $T = M + 1 = 64$, $\gamma = -10$ dB.

*2) **Multi-user Case**:* Now, we consider the multi-user case, and briefly review several existing methods for channel estimation.

Method I: Direct Channel Estimation Method [51]

In [51], the authors proposed a channel estimation algorithm that is a direct extension of the single-user case. Specifically, the channel estimation period is divided into T sub-phases, and the RIS applies the phase shift vector $\boldsymbol{\theta}_t$ in the t-th sub-phase. In each sub-phase, the users transmit orthogonal pilot sequences with length K, and the k-th user's pilot sequence is defined as $\mathbf{x}_k = [x_{k,1}, \cdots, x_{k,K}]^{\mathrm{T}} \in \mathbb{C}^{K \times 1}$, such that $\mathbf{x}_k^{\mathrm{H}}\mathbf{x}_l = 0$ for $k \neq l, \forall l, k$, and $\mathbf{x}_k^{\mathrm{H}}\mathbf{x}_k = K$. In different sub-phases, the users adopt the same set of pilot sequences while the RIS uses different phase shift vectors. In the t-th sub-phase, the received training signal $\mathbf{Y}_t \in \mathbb{C}^{N \times K}$ at the BS is given by

$$\mathbf{Y}_t = \sum_{k=1}^{K} \sqrt{P_k}\left(\mathbf{h}_{d,k} + \mathbf{G}_k\boldsymbol{\theta}_t\right)\mathbf{x}_k^{\mathrm{T}} + \mathbf{N}_t, t = 1, \cdots, T, \quad (26)$$

where $\mathbf{N}_t \in \mathbb{C}^{N \times K}$ is the noise matrix at the BS, whose columns are independent and have probability distribution $\mathcal{CN}\left(\mathbf{0}, \sigma^2\mathbf{I}_N\right)$. By right multiplying both sides of (26) with $\mathbf{x}_k^*$ and by taking into account that the users' pilot sequences are orthonormal, we have

$$\mathbf{y}_{t,k} = \sqrt{P_k}\left(\mathbf{h}_{d,k} + \mathbf{G}_k\boldsymbol{\theta}_t\right) + \mathbf{n}_{t,k}, t = 1, \cdots, T, \quad (27)$$

where $\mathbf{y}_{t,k} = \mathbf{Y}_t\mathbf{x}_k^*$ and $\mathbf{n}_{t,k} = \mathbf{N}_t\mathbf{x}_k^*$. The obtained expression in (27) is similar to the single-user case in (4), and the already described single-user channel estimation algorithms can be directly applied. For this scheme, the minimum number of sub-phases is equal to $T = M + 1$ [51]. Hence, the pilot overhead is equal to $K(M+1)$.

Method II: Exploiting the Common RIS-BS Channel [52]

The above channel estimation scheme fails to exploit the inherent structure of RIS-aided communication systems, and requires a large amount of pilot training overhead, especially when K and/or M are large. Specifically, all users' cascaded channels share the same RIS-BS channel and, therefore, it is possible to reduce the channel training overhead since the number of independent complex variables to be estimated is $KN + NM + MK$ rather than $(M + 1)NK$.

In [52], the authors exploited the common RIS-BS channel and proposed a novel three-stage channel estimation algorithm. In the first stage, the RIS is switched off, and the direct user-BS channels can be readily estimated. In the second stage, one reference user is selected and only this user can transmit its pilot signals. For ease of notation, let us denote this selected user as user 1. Since the direct channel $\mathbf{h}_{d,1}$ has been estimated in stage one, its impact when estimating the cascaded channel of user 1, i.e., $\mathbf{G}_1$, can be subtracted from the received signal at the BS, and the cascaded channel $\mathbf{G}_1$ can be estimated using the same channel estimation methods as for the single-user case.

Now, we focus on the third stage. For simplicity, we assume that the channels estimated in the first two stages are perfect. By substituting $\mathbf{G}_1 = \mathbf{H}\text{diag}\left(\mathbf{h}_{r,1}\right)$ into (1), the received signal can be rewritten as

$$\begin{aligned}\mathbf{y}_t = &\textstyle\sum_{k=1}^{K} \sqrt{P_k}\mathbf{h}_{d,k}x_{k,t} + \sqrt{P_1}\mathbf{G}_1\boldsymbol{\theta}_t x_{1,t} \\ &+ \textstyle\sum_{k=2}^{K} \sqrt{P_k}\mathbf{G}_1\text{diag}\left(\boldsymbol{\theta}_t\right)\tilde{\mathbf{h}}_k x_{k,t} + \mathbf{n}_t,\end{aligned} \tag{28}$$

where $\tilde{\mathbf{h}}_k = \text{diag}(\mathbf{h}_{r,1})^{-1}\mathbf{h}_{r,k}, k = 2, \cdots, K$. If $\mathbf{h}_{r,k}$ can be estimated, the cascaded channel $\mathbf{G}_k$ for $k = 2, \cdots, K$ can be calculated as $\mathbf{G}_k = \mathbf{G}_1\text{diag}(\tilde{\mathbf{h}}_k)$. Hence, the remaining task is to estimate $\mathbf{h}_{r,k}, k = 2, \cdots, K$. In [52], the authors assume that one user sends its pilot signals $x_{k,t} = 1, P_k = 1$ at each time and all the other users keep silent. Thus, we have $x_{j,t} = 0$ for all $j \neq k$. Then, the received signal is given by

$$\mathbf{y}_t = \mathbf{h}_{d,k} + \mathbf{G}_1\text{diag}\left(\boldsymbol{\theta}_t\right)\tilde{\mathbf{h}}_k + \mathbf{n}_t. \tag{29}$$

In the following, we discuss the two cases where M is less and greater than N, respectively.

Case I: $M \leq N$. By setting $\boldsymbol{\theta}_t = \mathbf{1}$, $\tilde{\mathbf{h}}_k$ can be directly estimated as

$$\tilde{\mathbf{h}}_k = \mathbf{G}_1^{-1}\left(\mathbf{y}_t - \mathbf{h}_{d,k}\right). \tag{30}$$

Accordingly, a single time slot is required to estimate $\tilde{\mathbf{h}}_k$ in this case.

Case II: $M > N$. In this case, $\tilde{\mathbf{h}}_k$ cannot be estimated using (30) since the system of equations in (29) is under-determined. To address this issue, a larger number of time slots for training is required. In the first training time slot, the first N reflecting elements are switched on and the reflection coefficients are set equal to $[\theta_{t,1}, \theta_{t,2} \cdots, \theta_{t,N}]^{\mathrm{T}} = \mathbf{1}$. Then, the first N coefficients of $\tilde{\mathbf{h}}_k$ can be estimated as follows

$$\begin{bmatrix} \tilde{h}_{k,1} \\ \tilde{h}_{k,2} \\ \vdots \\ \tilde{h}_{k,N} \end{bmatrix} = \begin{bmatrix} G_{1,11} & \cdots & G_{1,1N} \\ \vdots & \ddots & \vdots \\ G_{1,N1} & \cdots & G_{1,NN} \end{bmatrix}^{-1} \left(\mathbf{y}_t - \mathbf{h}_{d,k}\right), \tag{31}$$

where $\tilde{h}_{k,i}$ is the i-th element of $\tilde{\mathbf{h}}_k$ and $G_{1,ij}$ is the (i,j)-th entry of the matrix $\mathbf{G}_1$. In the second time slot, the next N reflecting elements are switched on, i.e., $[\theta_{t,N+1}, \theta_{t,2} \cdots, \theta_{t,2N}]^{\mathrm{T}} = \mathbf{1}$, while the others are switched off. The corresponding N coefficients in $\tilde{\mathbf{h}}_k$ can be estimated using again (31). This procedure is repeated until all the coefficients in $\tilde{\mathbf{h}}_k$ are estimated. For each user, at least $\lceil \frac{M}{N} \rceil$ time slots are required to estimate $\tilde{\mathbf{h}}_k$.

By combining Case I with Case II, the minimum number of required pilots is equal to $K + M + \max\left(K-1, (K-1)\lceil M/N \rceil\right)$, which is much less than the number of pilots that is needed for the channel estimation algorithm in [51]. The idea of this approach has been extended to multiuser OFDMA in [53].

The main issue of the method introduced in [52] is the error propagation. The authors of [54] and [55] have proposed channel estimation protocols to address this issue, where the direct channel and the reflected channels are estimated at the same time. The amount of pilot overhead required by the solutions proposed in [54] and [55] is the same as that of the method proposed in [52].

B. Structured Channel Models

A geometric channel model or the Saleh-Valenzuela (SV) channel model [56] is usually used to characterize the channels in the mmWave and THz frequency bands, where multipath scattering is sparse and the number of channel parameters is small. Assuming a uniform linear array (ULA) at the BS and that the RIS elements are arranged in uniform planar array (UPA), the SV channel model of the RIS-BS link in the presence of L_{BR} spatial paths is given by [56]

$$\begin{aligned}\mathbf{H} &= \textstyle\sum_{l=1}^{L_{\mathrm{BR}}} \alpha_l \mathbf{a}_{\mathrm{B}}\left(\omega_{\mathrm{BH},l}\right)\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}\left(\boldsymbol{\omega}_{\mathrm{RH},l}\right) \\ &= \mathbf{A}_{\mathrm{B}}(\boldsymbol{\omega}_{\mathrm{BH}})\boldsymbol{\Lambda}_{\mathrm{BR}}\mathbf{A}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH}}),\end{aligned} \tag{32}$$

where the diagonal matrix $\boldsymbol{\Lambda}_{\mathrm{BR}} = \text{diag}(\alpha_1, \ldots, \alpha_{L_{\mathrm{BR}}})$ contains the complex path gains $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{L_{\mathrm{BR}}}]^{\mathrm{T}}$. The columns of $\mathbf{A}_{\mathrm{R}}(\omega_{\mathrm{RH}}) = [\mathbf{a}_{\mathrm{R}}\left(\omega_{\mathrm{RH},1}\right), \ldots, \mathbf{a}_{\mathrm{R}}\left(\omega_{\mathrm{RH},L_{\mathrm{BR}}}\right)]$ denote the steering vectors of an $M_x \times M_z$ UPA, namely the RIS, on the xz-plane, which are expressed as

$$\mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{RH},l}\right) = \mathbf{a}_x(\omega_{x,l}) \otimes \mathbf{a}_z(\omega_{z,l}), \forall l, \tag{33}$$

where $\mathbf{a}_x(\omega_{x,l}) = [1, e^{j\omega_{x,l}}, \ldots, e^{j(M_x-1)\omega_{x,l}}]^{\mathrm{T}}$ is the horizontal array response with spatial frequency $\omega_{x,l} = 2\pi\frac{d}{\lambda_c}\sin(\phi_{el,l})\cos(\phi_{az,l})$, and $\mathbf{a}_z(\omega_{z,l}) = [1, e^{j\omega_{z,l}}, \ldots, e^{j(M_z-1)\omega_{z,l}}]^{\mathrm{T}}$ is the vertical array response with spatial frequency $\omega_{z,l} = 2\pi\frac{d}{\lambda_c}\cos(\phi_{el,l})$. Here, d denotes the antenna spacing, λ_c is the carrier wavelength, and $\phi_{el,l}$ and $\phi_{az,l}$ denote the elevation and the azimuth, respectively, angles of departures (AoDs) from the RIS. The two-dimensional (2D) spatial frequency vector $\boldsymbol{\omega}_{\mathrm{RH},l} = [\omega_{x,l}, \omega_{z,l}]^{\mathrm{T}}$ corresponds to the 2D spatial angles $\{\phi_{el,l}, \phi_{az,l}\}$.

The steering vectors of an N-element ULA $\mathbf{A}_{\mathrm{B}}(\omega_{\mathrm{BH}}) = [\mathbf{a}_{\mathrm{B}}\left(\omega_{\mathrm{BH},1}\right), \ldots, \mathbf{a}_{\mathrm{B}}\left(\omega_{\mathrm{BH},L_{\mathrm{BR}}}\right)]$ are the same as (33), but the angles of elevation are $\phi_{el,l} = \frac{\pi}{2}, \forall l$. Hence, $\mathbf{a}_{\mathrm{B}}(\omega_{\mathrm{BH},l})$ is

$$\mathbf{a}_{\mathrm{B}}(\omega_{\mathrm{BH},l}) = [1, e^{j\omega_{\mathrm{BH},l}}, \ldots, e^{j(N-1)\omega_{\mathrm{BH},l}}]^{\mathrm{T}}, \forall, l, \tag{34}$$

where $\boldsymbol{\omega}_{\mathrm{BH}}$ is the angle of arrival (AoA) spatial frequency vector at the BS, which is given by $\boldsymbol{\omega}_{\mathrm{BH}} = [\omega_{\mathrm{BH},1}, \ldots, \omega_{\mathrm{BH},L_{\mathrm{BR}}}]$.

The SV channel model for the user-RIS link in the presence of L_{RU} spatial paths has a similar structure. Precisely, denoting by $\mathbf{h}_{r,k} \in \mathbb{C}^{M\times 1}$ the channel from user k to the RIS, it can be formulated as

$$\begin{aligned}\mathbf{h}_{r,k} &= \sum\nolimits_{l=1}^{L_{\mathrm{RU}}} \beta_{k,l}\mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{RH}_r,k,l}\right)\\ &= \mathbf{A}_{\mathrm{R}}(\boldsymbol{\omega}_{\mathrm{RH}_r,k})\boldsymbol{\beta}_k,\end{aligned} \tag{35}$$

where $\boldsymbol{\beta}_k = [\beta_{k,l}, \ldots, \beta_{k,L_{\mathrm{RU}}}]^{\mathrm{T}}$ is the vector of complex path gains, and the matrix $\mathbf{A}_{\mathrm{R}}(\boldsymbol{\omega}_{\mathrm{RH}_r,k}) = [\mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{RH}_r,k,1}\right), \ldots, \mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{RH}_r,k,L_{\mathrm{RU}}}\right)]$ collects the steering vectors of the propagation paths whose AoA spatial frequency vector is $\boldsymbol{\omega}_{\mathrm{RH}_r,k} = [\boldsymbol{\omega}_{\mathrm{RH}_r,k,1}^{\mathrm{T}}, \ldots, \boldsymbol{\omega}_{\mathrm{RH}_r,k,L_{\mathrm{RU}}}^{\mathrm{T}}]^{\mathrm{T}}$.

Accordingly, the received baseband signal in (1) is rewritten as[4]

$$\begin{aligned}\mathbf{y}_t &= \sum\nolimits_{k=1}^{K} \sqrt{P_k}\mathbf{H}\boldsymbol{\Theta}_t\mathbf{h}_{r,k}x_{k,t} + \mathbf{n}_t\\ &= \sum\nolimits_{k=1}^{K} \sqrt{P_k}\mathbf{G}_k\boldsymbol{\theta}_t x_{k,t} + \mathbf{n}_t.\end{aligned} \tag{36}$$

1) Single-user case: Let us start with the analysis of the single-user case. By omitting the user index for simplicity, (36) reduces to

$$\mathbf{y}_t = \sqrt{P}\mathbf{G}\boldsymbol{\theta}_t x_t + \mathbf{n}_t. \tag{37}$$

The overarching idea of estimating a channel with an inherent geometric structure is to reconstruct the channel by estimating a small number of angles and gains instead of directly estimating all the channel coefficients. We overview several existing channel estimation methods based on this approach.

Method I: One-Stage Channel Estimation [57], [58]

In this method, the estimation of the cascaded channel is reformulated in terms of an AoA estimation problem

$$\mathbf{y}_t = \mathbf{Z}\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\gamma} + \mathbf{n}_t, \tag{38}$$

where $\mathbf{Z}$ is the known sensing matrix that contains the training pilots, the matrix $\mathbf{A}(\boldsymbol{\omega})$ is the array response that depends on the spatial frequency vector $\boldsymbol{\omega}$ and $\boldsymbol{\gamma}$ is the vector with the unknown channel gains. The detailed definitions of $\mathbf{Z}$, $\mathbf{A}(\boldsymbol{\omega})$ and $\boldsymbol{\gamma}$ will be given later.

Accordingly, any known AoA estimation algorithms, including the multiple signal classification (MUSIC) [59], the estimation of signal parameters via rational invariance techniques (ESPRIT) [60] and compressed sensing (CS) based techniques, can be directly applied. Specifically, by utilizing the identity $\mathbf{AB}\diamond\mathbf{CD} = (\mathbf{A}\otimes\mathbf{C})(\mathbf{B}\diamond\mathbf{D})$, the cascaded channel $\mathbf{G}$ can be formulated as

$$\begin{aligned}\mathbf{G} &= \mathbf{H}\mathrm{diag}(\mathbf{h}_r) = \mathbf{H}\diamond\mathbf{h}_r^{\mathrm{T}}\\ &= \mathbf{A}_{\mathrm{B}}(\boldsymbol{\omega}_{\mathrm{BH}})\left(\boldsymbol{\Lambda}_{\mathrm{BR}}\otimes\boldsymbol{\beta}^{\mathrm{T}}\right)\left(\mathbf{A}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH}})\diamond\mathbf{A}_{\mathrm{R}}^{\mathrm{T}}(\boldsymbol{\omega}_{\mathrm{RH}_r})\right).\end{aligned} \tag{39}$$

Denoting by $\boldsymbol{\omega} = [\boldsymbol{\omega}_{\mathrm{BH}}^{\mathrm{T}}, \boldsymbol{\omega}_{\mathrm{RH}}^{\mathrm{T}}, \boldsymbol{\omega}_{\mathrm{RH}_r}^{\mathrm{T}}]^{\mathrm{T}}$, using the identity $\mathrm{vec}(\mathbf{A}\left(\mathrm{diag}(\mathbf{b})\otimes\mathbf{d}^{\mathrm{T}}\right)\mathbf{C}) = (\mathbf{C}^{\mathrm{T}}\diamond\mathbf{A})(\mathbf{b}\otimes\mathbf{d})$ and vectorizing both sides of (39), we arrive at

$$\mathrm{vec}(\mathbf{G}) = \mathbf{A}(\boldsymbol{\omega})\boldsymbol{\gamma}, \tag{40}$$

[4]For simplicity, we ignore the direct BS-user channels since they are likely blocked by obstacles at high frequency bands.

where

$$\mathbf{A}(\boldsymbol{\omega}) = \left(\mathbf{A}_{\mathrm{R}}^{\mathrm{T}}(\boldsymbol{\omega}_{\mathrm{RH}_r})\diamond\mathbf{A}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH}})\right)^{\mathrm{T}}\diamond\mathbf{A}_{\mathrm{B}}(\boldsymbol{\omega}_{\mathrm{BH}}), \tag{41}$$

$$\boldsymbol{\gamma} = \boldsymbol{\alpha}\otimes\boldsymbol{\beta}. \tag{42}$$

The matrix $\left(\mathbf{A}_{\mathrm{R}}^{\mathrm{T}}(\boldsymbol{\omega}_{\mathrm{RH}_r})\diamond\mathbf{A}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH}})\right)^{\mathrm{T}}$ is the cascaded array response at the RIS [61], whose k-th column is

$$\begin{aligned}\left[\left(\mathbf{A}_{\mathrm{R}}^{\mathrm{T}}(\boldsymbol{\omega}_{\mathrm{RH}_r})\diamond\mathbf{A}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH}})\right)_{k,:}\right]^{\mathrm{T}} &= \left[\mathbf{a}_{\mathrm{R}}^{\mathrm{T}}(\boldsymbol{\omega}_{\mathrm{RH}_r,p})\diamond\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH},q})\right]^{\mathrm{T}}\\ &= \mathbf{a}_{\mathrm{R}}(\boldsymbol{\omega}_{\mathrm{RH}_r,p} - \boldsymbol{\omega}_{\mathrm{RH},q}),\end{aligned} \tag{43}$$

where $p = \lfloor k/L_{\mathrm{BR}}\rfloor$ and $q = \mathrm{mod}_{L_{\mathrm{BR}}}(k)$. The vector $\boldsymbol{\gamma}$ depends on the complex path gains of the cascaded channel.

By combining (37) with (40), (37) can be reformulated in a compact form as

$$\begin{aligned}\mathbf{y}_t &= \sqrt{P}\left(x_t\boldsymbol{\theta}_t^{\mathrm{T}}\otimes\mathbf{I}_N\right)\mathrm{vec}(\mathbf{G}) + \mathbf{n}_t\\ &= \sqrt{P}\left(x_t\boldsymbol{\theta}_t^{\mathrm{T}}\otimes\mathbf{I}_N\right)\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\gamma} + \mathbf{n}_t.\end{aligned} \tag{44}$$

Similar to (7), the measurement (pilot) signals received at the BS over T training time slots are stacked into a vector, as

$$\mathbf{y} = \sqrt{P}\mathbf{Z}\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\gamma} + \mathbf{n} \in \mathbb{C}^{NT\times 1}, \tag{45}$$

where

$$\mathbf{Z} = \begin{bmatrix} x_1\boldsymbol{\theta}_1^{\mathrm{T}}\otimes\mathbf{I}_N \\ \vdots \\ x_T\boldsymbol{\theta}_T^{\mathrm{T}}\otimes\mathbf{I}_N \end{bmatrix}.$$

By inspection of (45), the obtained signal model corresponds to a standard AoA estimation problem as formulated in (38). Therefore, the deterministic maximum likehood (DML) criterion can be adopted to estimate $\boldsymbol{\omega}$ [35]

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \mathbf{y}^{\mathrm{H}}\mathbf{P}_{\mathbf{ZA}(\boldsymbol{\omega})}^{\perp}\mathbf{y},$$

where $\mathbf{P}_{\mathbf{ZA}(\boldsymbol{\omega})}^{\perp}$ is the projection orthogonal to the columns of $\mathbf{ZA}(\boldsymbol{\omega})$. Given the estimate, the vector of channel gains is given by $\hat{\boldsymbol{\gamma}} = (\mathbf{ZA}(\hat{\boldsymbol{\omega}}))^{\dagger}\mathbf{y}/\sqrt{P}$.

However, the DML criterion ignores the structural properties of the signal in (45). More precisely, the cascaded gain vector $\boldsymbol{\gamma}$ in (42) is a nonlinear function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and the elements of $\boldsymbol{\gamma}$ are not independent. In addition, the spatial frequency of each cascaded steering vector in (43) is different. In order to exploit the inherent relationship between the columns of the cascaded array response matrix, the authors of [57] and [58] adopted the CS technique, according to which each array response matrix $\mathbf{A}(\boldsymbol{\omega})$ is replaced by a dictionary matrix composed of potential steering vectors. Specifically, (40) is approximated using the virtual angular domain (VAD) representation, i.e., [57]

$$\mathrm{vec}(\mathbf{G}) \approx \mathbf{A}_D\boldsymbol{\gamma}_D, \tag{46}$$

where $\boldsymbol{\gamma}_D$ is an $L_{\mathrm{BR}}L_{\mathrm{RU}}$-sparse vector with $L_{\mathrm{BR}}L_{\mathrm{RU}}$ nonzero elements corresponding to the cascaded channel gains. The matrix $\mathbf{A}_D$ is the composite dictionary of $\mathbf{A}(\boldsymbol{\omega})$ and is defined as $\mathbf{A}_D = \left(\mathbf{A}_{\mathrm{R}D}^{\mathrm{T}}\diamond\mathbf{A}_{\mathrm{R}D}^{\mathrm{H}}\right)^{\mathrm{T}}\otimes\mathbf{A}_{\mathrm{B}D}$. The columns of the overcomplete dictionary matrix $\mathbf{A}_{\mathrm{B}D} \in \mathbb{C}^{N\times G_{\mathrm{B}}}$ ($N \ll G_{\mathrm{B}}$) are the BS steering vectors whose sample spatial frequencies

are chosen from pre-discretized grids. However, the over-complete dictionary matrix $\mathbf{A}_{\mathrm{R}D} \in \mathbb{C}^{M\times G_{\mathrm{R}}}$ ($M \ll G_{\mathrm{R}}$) is specified by a two-dimensional grid and its columns can be written as $\mathbf{a}_x(\omega_{x,q}) \otimes \mathbf{a}_z(\omega_{z,p}), 1 \le q,p \le \sqrt{G_{\mathrm{R}}}$ with $\omega_{x,q}$ and $\omega_{z,p}$ being independently selected from pre-discretized grids. By capitalizing on the sparse approximation in (46), the estimation of $\mathbf{G}$ can be transformed into a sparse signal recovery problem as

$$\min_{\boldsymbol{\gamma}_D} \ ||\boldsymbol{\gamma}_D||_1 \tag{47a}$$

$$\text{s.t. } ||\mathbf{y} - \sqrt{P}\mathbf{Z}\mathbf{A}_D\boldsymbol{\gamma}_D||_2 \le \xi, \tag{47b}$$

where the 1-norm of $\boldsymbol{\gamma}_D$ enforces the sparse structure and $\xi \ge 0$ is an error tolerance parameter that depends on the noise power. Problem (47) can be solved by using several classical CS algorithms, such as the orthogonal matching pursuit (OMP) [57] and the alternating direction method of multipliers (ADMM) [58]. After recovering $\boldsymbol{\gamma}_D$, the cascaded channel $\mathbf{G}$ can be retrieved from (46). However, the main issue with the CS approach used for solving Problem (47) is the large size of $\mathbf{A}_D$, which has $G_{\mathrm{B}}G_{\mathrm{R}}^2$ columns. If each spatial frequency dimension has 100 grids, this matrix has 10^{10} columns. Therefore, a more tractable and practical method is needed. According to [62], $l\log(\frac{m}{l})$ measurements are needed for successful recovering an l-sparse vector of dimension $m\times 1$. Thus, the required pilot overhead for solving Problem (47) is $T \ge \frac{L_{\mathrm{BR}}L_{\mathrm{RU}}}{N}\log(\frac{G_{\mathrm{B}}G_{\mathrm{R}}^2}{L_{\mathrm{BR}}L_{\mathrm{RU}}})$.

Method II: Two-Stage Channel Estimation [63]–[65]

A different approach for channel estimation consists of splitting the cascaded channel into two stages. In the first stage, the AoAs at the BS are estimated. After eliminating these estimated angles from the variables to the estimated, only the cascaded spatial frequencies and the channel gains need to be determined. The estimation of both stages can be formulated in terms of several several AoA estimation problems, thus any AoA estimation algorithm can be employed.

Specifically, it is assumed that the user sends the pilot signal $x_t = 1$ during T time slots. Defining $\boldsymbol{\Xi} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T]$ and stacking the T observations in (37), we obtain

$$\begin{aligned}\mathbf{Y} &= [\mathbf{y}_1, \dots, \mathbf{y}_T] \\ &= \sqrt{P}\mathbf{G}\boldsymbol{\Xi} + \mathbf{N},\end{aligned} \tag{48}$$

where $\mathbf{N} = [\mathbf{n}_1, \dots, \mathbf{n}_T]$. Based on (48), we elaborate the estimation process of both stages.

- **First Stage:** We rewrite (48) as

$$\mathbf{Y} = \sqrt{P}\mathbf{A}_{\mathrm{B}}(\boldsymbol{\omega}_{\mathrm{BH}})\mathbf{B}^{\mathrm{T}}\boldsymbol{\Xi} + \mathbf{N} \in \mathbb{C}^{N\times T}, \tag{49}$$

  where $\mathbf{B} = (\boldsymbol{\Lambda}_{\mathrm{BR}}\mathbf{A}_{\mathrm{R}}^{\mathrm{H}}(\boldsymbol{\omega}_{\mathrm{RH}})\mathrm{diag}(\mathbf{h}_r))^{\mathrm{T}}$. Any AoA estimation algorithms can be used to determine $\boldsymbol{\omega}_{\mathrm{BH}}$ from (49). For example, the authors of [63] and [64] approximated (49) as a sparse signal recovery problem similar to Problem (47), and adopted the OMP technique to solve the formulated problem. Since $N \gg L_{\mathrm{BR}}$, (49) readily satisfies the condition of the number of measurements required for successful recovery, i.e., $N \ge L_{\mathrm{BR}}\log(\frac{G_{\mathrm{B}}}{L_{\mathrm{BR}}})$. The main issue with the on-grid CS method is the mismatch between the estimated and the actual angles. The estimation error can be improved by enlarging the dimension of the dictionary, but this also reduces the orthogonality of the dictionary, and leads to a higher computational complexity. Furthermore, the authors of [66] and [67] first transformed the signal in (49) from the spatial domain to the angle domain by utilizing a DFT matrix, and then estimated the AoAs by searching for the non-zero elements in the angle domain. However, this method is limited to massive MIMO scenarios, where the steering vectors lie approximately in the DFT matrix space. The estimation error caused by the slight power leakage can also be improved using the angle rotation technique [67]. Moreover, to address the mismatch issue, the atomic norm minimization technique was adopted in [65] to estimate the off-grid angles by employing convex optimization tools.

- **Second Stage:** Assuming that the estimate $\hat{\boldsymbol{\omega}}_{\mathrm{BH}}$ of $\boldsymbol{\omega}_{\mathrm{BH}}$ is perfect, we obtain

$$\begin{aligned}\mathbf{Y}_{\mathrm{R}} &= \frac{1}{\sqrt{P}}(\mathbf{A}_{\mathrm{B}}^{\dagger}(\hat{\boldsymbol{\omega}}_{\mathrm{BH}})\mathbf{Y})^{\mathrm{T}} \\ &\approx \boldsymbol{\Xi}^{\mathrm{T}}\mathbf{B} + \mathbf{N}_b \in \mathbb{C}^{T\times L_{\mathrm{BR}}},\end{aligned} \tag{50}$$

  where $\mathbf{N}_b = \frac{1}{\sqrt{P}}(\mathbf{A}_{\mathrm{B}}^{\dagger}(\hat{\boldsymbol{\omega}}_{\mathrm{BH}})\mathbf{N})^{\mathrm{T}}$ and $\mathbf{B}$ is the matrix to be estimated.

  Correlation-ignored Approach [63], [65].

  If one ignores the correlation among the cascaded gains and spatial frequencies, the estimation of $\mathbf{Y}_{\mathrm{R}}$ can be obtained by constructing L_{BR} L_{RU}-dimensional AoA estimation problems [63], [65]. In particular, the l-th column of (50) is given by

$$\begin{aligned}[\mathbf{Y}_{\mathrm{R}}]_{:,l} &= \boldsymbol{\Xi}^{\mathrm{T}}\mathrm{diag}(\mathbf{h}_r)\mathbf{a}_{\mathrm{R}}^{*}(\boldsymbol{\omega}_{\mathrm{RH},l})\alpha_l + [\mathbf{N}_b]_{:,l} \\ &= \boldsymbol{\Xi}^{\mathrm{T}}\mathbf{A}_{\mathrm{R}}(\triangle\boldsymbol{\omega}_{\mathrm{Rh_r},l})\alpha_l\boldsymbol{\beta} + [\mathbf{N}_b]_{:,l},\end{aligned} \tag{51}$$

  which depends on $\triangle\boldsymbol{\omega}_{\mathrm{Rh_r},l} = [\boldsymbol{\omega}_{\mathrm{Rh_r},1}^{\mathrm{T}} - \boldsymbol{\omega}_{\mathrm{RH},l}^{\mathrm{T}}, ..., \boldsymbol{\omega}_{\mathrm{Rh_r},L_{\mathrm{RU}}}^{\mathrm{T}} - \boldsymbol{\omega}_{\mathrm{RH},l}^{\mathrm{T}}]^{\mathrm{T}}$ and $\alpha_l\boldsymbol{\beta}$. Thus, (50) can be decomposed into L_{BR} L_{RU}-dimensional AoA estimation problems along with determining $L_{\mathrm{BR}}L_{\mathrm{RU}}$ cascaded path gains at the RIS. With the estimates of $\triangle\boldsymbol{\omega}_{\mathrm{Rh_r},l}$ and $\alpha_l\boldsymbol{\beta}$ from (51), the l-th column of $\mathbf{B}$ is estimated as

$$[\mathbf{B}]_{:,l} = \mathbf{A}_{\mathrm{R}}(\triangle\widehat{\boldsymbol{\omega}}_{\mathrm{Rh_r},l})\widehat{\alpha_l\boldsymbol{\beta}}. \tag{52}$$

  Correlation-based Approach [67].

  If one considers the fact that the cascaded gains and spatial frequencies are interrelated, the estimation complexity can be further reduced to solve one L_{RU}-dimensional AoA estimation problem plus $L_{\mathrm{BR}} - 1$ one-dimensional AoA estimation problems. In particular, $[\mathbf{Y}_{\mathrm{R}}]_{:,l}$ in (51) is rewritten as

$$[\mathbf{Y}_{\mathrm{R}}]_{:,l} = \triangle\alpha_l\boldsymbol{\Xi}^{\mathrm{T}}\mathrm{diag}\left(\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}(\triangle\boldsymbol{\omega}_{\mathrm{RH},l})\right)[\mathbf{B}]_{:,1} + [\mathbf{N}_b]_{:,l}, \tag{53}$$

  where $\triangle\alpha_l = \frac{\alpha_l}{\alpha_1}$ and $\triangle\boldsymbol{\omega}_{\mathrm{RH},l} = \boldsymbol{\omega}_{\mathrm{RH},l} - \boldsymbol{\omega}_{\mathrm{RH},1}$. The first column of $\mathbf{B}$ is first estimated from (51) using an L_{RU}-dimensional AoA estimation problem. Substituting the estimated $[\mathbf{B}]_{:,1}$ into (53), then $[\mathbf{B}]_{:,l}$ for $l = 2, ..., L_{\mathrm{BR}}$ can be determined by only estimating $\triangle\alpha_l$ and $\triangle\boldsymbol{\omega}_{\mathrm{RH},l}$

through a one-dimensional AoA estimation problem. Therefore, only $L_{\mathrm{BR}}+L_{\mathrm{RU}}-1$ equivalent cascaded paths need to be estimated at the RIS. Compared with the correlation-ignored approach, the correlation-based approach enjoys a much lower computational complexity since it needs to solve the L_{RU}-dimensional AoA estimation problem only once.

We characterize the pilot overhead at this stage based on the condition for successful recovery using the CS technique: $T \geq L_{\mathrm{RU}} \log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})$.

Method III: Beam Training for Channel Estimation [68]–[70].

A different method for estimating sparse channels is beam training, which is suitable for line-of-sight (LoS)-dominant channels where the LoS path is much stronger than the non-LoS (NLoS) components. Specifically, the composite LoS-dominant user-RIS-BS channel can be formulated as

$$\mathbf{H}\boldsymbol{\Theta}_t\mathbf{h}_r = \alpha\beta\mathbf{a}_{\mathrm{B}}\left(\omega_{\mathrm{BH}}\right)\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}\left(\boldsymbol{\omega}_{\mathrm{RH}}\right)\mathrm{diag}(\boldsymbol{\theta}_t)\mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{RH}_r}\right).$$

It is readily seen that the maximum beam gain is obtained when the beams at the BS and the RIS align with the user-RIS-BS channel.

- **BS Beam:** Let the vector $\mathbf{f}_{\mathrm{B}} \in \mathbb{C}^{N\times 1}$ be the combiner at the BS. The maximization of $|\mathbf{f}_{\mathrm{B}}^{\mathrm{H}}\mathbf{a}_{\mathrm{B}}\left(\omega_{\mathrm{BH}}\right)|$ as a function of $\mathbf{f}_{\mathrm{B}}$ returns $\mathbf{f}_{\mathrm{B,opt}} = \mathbf{a}_{\mathrm{B}}^{\mathrm{H}}\left(\omega_{\mathrm{BH}}\right)$.
- **RIS Beam:** The optimal phase shifts at the RIS that maximize $|\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}\left(\boldsymbol{\omega}_{\mathrm{RH}}\right)\mathrm{diag}(\boldsymbol{\theta}_t)\mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{Rh}_r}\right)|$ are given by $\boldsymbol{\theta}_{\mathrm{opt}} = \mathbf{a}_{\mathrm{R}}\left(\boldsymbol{\omega}_{\mathrm{RH}} - \boldsymbol{\omega}_{\mathrm{Rh}_r}\right)$.

To estimate the AoAs and the AoDs of the LoS-dominant user-RIS-BS channel, appropriate training codebooks are first constructed for each beam by discretizing the entire angular domain. Then, an exhaustive beam training method can be used: the RIS applies the different beams of the codebook one-by-one in consecutive time slots. Then, the BS determines the optimal BS and RIS beam directions that provide the maximum received signal power/SNR. However, such an exhaustive beam training method is pilot-consuming and requires at least $T = M_x$ time slots [68]. Therefore, various methods have been proposed to improve the efficiency of beam training, such as the multi-beam training method [68], the ternary-tree hierarchical search [69], and the two-way tree stage search [70].

Simulation Results: As an example, Fig. 3 shows the normalized mean square error (NMSE) of the cascaded channel matrix that is obtained by using the one-stage and two-stage estimation methods. The NMSE is defined as $\mathrm{NMSE} = \mathbb{E}\{\left\|\widehat{\mathbf{G}}-\mathbf{G}\right\|_F^2 / \|\mathbf{G}\|_F^2\}$. The OMP algorithm is adopted to solve the AoA estimation problems in both cases. It can be seen that the two-stage method outperforms the one-stage method. This is because the two-stage method reduces the power leakage among the cascaded multipath components compared with (46) by dividing the cascaded paths into several groups in (51), each of which corresponds to one BS-RIS path and does not leak power to other groups. However, there is a gap between the two methods and the CRB. This is due to the inherent mismatch between the on-grid angle estimated by the OMP and the actual continuous angle.

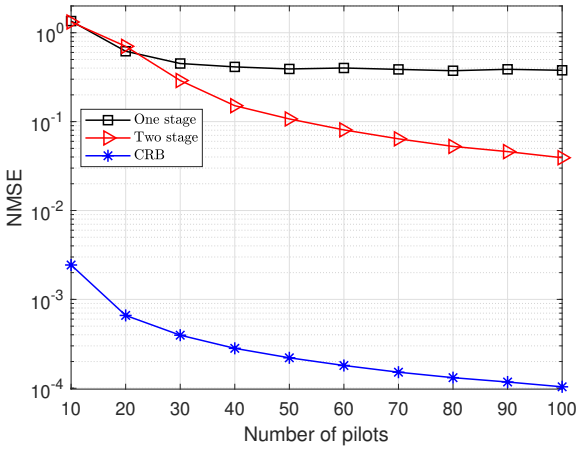

Fig. 3. NMSE versus the pilot overhead, when $N = 100$, $M = 10 \times 10$, $L_{\mathrm{BR}} = L_{\mathrm{RU}} = 4$ and SNR$= 0$ dB.

2) Multi-user case: In the multi-user case, all users share the common channel $\mathbf{H}$. This fact can be utilized to reduce the pilot overhead and the computational complexity as proposed in [61], [66], [67].

Method I: Double-sparse Based Channel Estimation [61], [66].

The authors of [61] investigated the row-column-block sparsity of the multi-user cascaded channels and proposed a joint multi-user channel estimation method:

- **Common column-block sparsity due to the presence of common scatters near the BS:** Note that all users share the same AoA array responses at the BS. Once the common AoAs are estimated from a specific user using the methods introduced in Subsection II-B1, the impact of the corresponding array responses can be eliminated from the received signals of all users.
- **Common row-block sparsity due to scaling property:** Based on [52], the signals from all users propagate through the same RIS, so the cascaded channels of different users are scaled by a diagonal array, as $\mathbf{G}_k = \mathbf{G}_1\mathrm{diag}(\tilde{\mathbf{h}}_k)$ defined in the text right after (28). Then, all cascaded channels are approximated by the common row-block sparse representation as $\mathbf{G}_k = \mathbf{A}_B\mathbf{X}\mathbf{A}_R^{\mathrm{H}}\mathrm{diag}(\tilde{\mathbf{h}}_k)$, where the sparse matrix $\mathbf{X}$ contains the cascaded gains. The optimization of $\mathbf{X}$ and $\tilde{\mathbf{h}}_k$ is formulated into a multi-user joint sparse matrix recovery problem, and solved by using the iterative reweighted algorithm [61], [71].

The pilot overhead of this method is $T \geq K\lceil M/(L_{\mathrm{BR}}L_{\mathrm{RU}})\rceil$ [61].

Method II: Exploiting the Common RIS-BS Channel [67].

The above approach still requires estimation of the unstructured vector $\tilde{\mathbf{h}}_k$, for $k = 2, ..., K$, the dimension of which is proportional to M. In order to reduce the number of parameters and pilot overhead, a geometric version of $\tilde{\mathbf{h}}_k$ is constructed in [67]:

- First, the spatial frequencies and gains of the cascaded channel for user 1 are estimated using the method discussed in Subsection II-B1 with pilot overhead $T_1 \geq L_{\mathrm{RU}} \log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})$.

TABLE I
PILOT OVERHEAD OF VARIOUS CHANNEL ESTIMATION ALGORITHMS.

Algorithm		**Pilot Overhead**	**Pilot Overhead,** $G_{\mathrm{B}}=4N$ **and** $G_{\mathrm{R}}=4M$
Single user	One stage [57], [58]	$\frac{L_{\mathrm{BR}}L_{\mathrm{RU}}}{N}\log(\frac{G_{\mathrm{B}}G_{\mathrm{R}}^2}{L_{\mathrm{BR}}L_{\mathrm{RU}}})$	$\frac{L_{\mathrm{BR}}L_{\mathrm{RU}}}{N}\log(\frac{64NM^2}{L_{\mathrm{BR}}L_{\mathrm{RU}}})$
	Two stage [63]–[65]	$L_{\mathrm{RU}}\log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})$	$L_{\mathrm{RU}}\log(\frac{4M}{L_{\mathrm{RU}}})$
Multiple user	[61]	$KM/(L_{\mathrm{BR}}L_{\mathrm{RU}})$	$KM/(L_{\mathrm{BR}}L_{\mathrm{RU}})$
	[67]	$L_{\mathrm{RU}}\log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})+(K-1)\frac{L_{\mathrm{RU}}}{L_{\mathrm{BR}}}\log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})$	$L_{\mathrm{RU}}\log(\frac{4M}{L_{\mathrm{RU}}})+(K-1)\frac{L_{\mathrm{RU}}}{L_{\mathrm{BR}}}\log(\frac{4M}{L_{\mathrm{RU}}})$

- Then, the cascaded channels for the remaining users are reformulated as $\mathbf{G}_k=\mathbf{H}_{\mathrm{c}}\mathrm{diag}(\mathbf{h}_{\mathrm{c},k})$, for $k=2,...,K$, where $\mathbf{H}_{\mathrm{c}}=\mathbf{H}\mathrm{diag}(\bar{\beta}\mathbf{a}_{\mathrm{R}}(\bar{\boldsymbol{\omega}}))^{-1}$ can be constructed based on the parameters estimated from user 1. Please refer to [67] for the detailed construction of $\mathbf{H}_{\mathrm{c}}$. The remaining unknow vector $\mathbf{h}_{\mathrm{c},k}=\mathrm{diag}(\bar{\beta}\mathbf{a}_{\mathrm{R}}(\bar{\boldsymbol{\omega}}))\mathbf{h}_{r,k}$ belongs to a standard AoA estimation model and can be estimated via various AoA estimation or CS methods. The required pilot overhead here is $T_2\geq(K-1)\frac{L_{\mathrm{RU}}}{L_{\mathrm{BR}}}\log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})$.

The overall pilot overhead for the scheme in [67] is given by $T\geq L_{\mathrm{RU}}\log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})+(K-1)\frac{L_{\mathrm{RU}}}{L_{\mathrm{BR}}}\log(\frac{G_{\mathrm{R}}}{L_{\mathrm{RU}}})$.

Table I summarizes the pilot overhead of various algorithms using CS-based techniques. The dimension of the dictionary is generally set to 2-4 times the number of antennas. Therefore, we can substitute $G_{\mathrm{B}}=4N$ and $G_{\mathrm{R}}=4M$ into the third column to obtain a more intuitive relationship between the pilot overhead and the number of antennas in the fourth column.

III. TRANSMISSION DESIGN

Based on the channels estimated by using the methods introduced in Section II, the reflection coefficients of the RIS and the beamforming vectors at the BS can be jointly optimized to achieve the desired objectives, such as maximizing the sum spectral efficiency or energy efficiency, and minimizing the total energy consumption, symbol-error probability, or transmission delay, etc.

The same system model in Fig. 1 is considered, but we focus our attention on downlink transmission for illustrative purposes. The baseband channels from the BS to the RIS, from the RIS to user k, and from the BS to user k are denoted by $\mathbf{H}^{\mathrm{H}}\in\mathbb{C}^{M\times N}$, $\mathbf{h}_{r,k}^{\mathrm{H}}\in\mathbb{C}^{1\times M}$ and $\mathbf{h}_{d,k}^{\mathrm{H}}\in\mathbb{C}^{1\times N}$, respectively. The transmitted data symbols for user k is denoted by s_k with $\mathbb{E}\{s_k\}=0$ and $\mathbb{E}\{s_k^2\}=1$. The beamforming vector for user k is $\mathbf{w}_k\in\mathbb{C}^{N\times 1}$. Then, the transmitted data at the BS is given by $\mathbf{x}=\sum_{j=1}^{K}\mathbf{w}_j s_j$, and the received signal at user k is expressed as

$$y_k=\left(\mathbf{h}_{d,k}^{\mathrm{H}}+\mathbf{h}_{r,k}^{\mathrm{H}}\boldsymbol{\Theta}\mathbf{H}^{\mathrm{H}}\right)\sum_{j=1}^{K}\mathbf{w}_j s_j+n_k, \tag{54}$$

where $n_k\sim\mathcal{CN}\left(0,\sigma_k^2\right)$ is the AWGN at user k and $\boldsymbol{\Theta}$ is the reflection coefficient matrix defined in Section II. Using $\mathbf{G}_k^{\mathrm{H}}=\mathrm{diag}\left(\mathbf{h}_{r,k}^{\mathrm{H}}\right)\mathbf{H}^{\mathrm{H}}$ and $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_M]^{\mathrm{T}}$, the signal-to-interference-plus-noise-ratio (SINR) at user k is given by

$$\mathrm{SINR}_k=\frac{\left|\left(\mathbf{h}_{d,k}^{\mathrm{H}}+\boldsymbol{\theta}^{\mathrm{T}}\mathbf{G}_k^{\mathrm{H}}\right)\mathbf{w}_k\right|^2}{\sum_{i=1,i\neq k}^{K}\left|\left(\mathbf{h}_{d,k}^{\mathrm{H}}+\boldsymbol{\theta}^{\mathrm{T}}\mathbf{G}_k^{\mathrm{H}}\right)\mathbf{w}_i\right|^2+\sigma_k^2}. \tag{55}$$

Hence, it is observed from (55) that the cascaded channel $\mathbf{G}_k^{\mathrm{H}}$ along with the direct channel $\mathbf{h}_{d,k}^{\mathrm{H}}$ are sufficient for designing the transmission of the BS and the RIS. The data rate of user k is $R_k=\log_2(1+\mathrm{SINR}_k)$.

For ease of writing, the collection of all beamforming vectors is denoted by $\mathbf{W}=[\mathbf{w}_1,\cdots,\mathbf{w}_K]$. The vast majority of the existing transmission design problems can be expressed in the following general form

$$\begin{aligned}\min_{\mathbf{W},\boldsymbol{\Theta}}\ & f(\mathbf{W},\boldsymbol{\Theta})\\ \text{s.t.}\ & \mathrm{C1}: g_i(\mathbf{W},\boldsymbol{\Theta})\geq D_i, i=1,\cdots,I,\\ & \mathrm{C2}: \theta_m\in\mathcal{S}_1 \text{ or } \mathcal{S}_2, \forall m=1,2,\ldots,M,\end{aligned} \tag{56}$$

where $f(\mathbf{W},\boldsymbol{\Theta})$ and $g_i(\mathbf{W},\boldsymbol{\Theta})$ can be any functions that depend on $\mathbf{W}$ and $\boldsymbol{\Theta}$, and I denotes the number of constraints in C1. In constraint C2, $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the set of continuous and discrete phase shifts respectively, which are given by

$$\mathcal{S}_1=\left\{\theta|\,\theta=e^{j\varphi},\varphi\in[0,2\pi)\right\}, \tag{57}$$

$$\mathcal{S}_2=\left\{\theta|\,\theta=e^{j(l-1)\Delta\varphi},l=1,\cdots,L\right\}, \tag{58}$$

where $L=2^b$ with b being the number of bits used to quantize the continuous phase shifts, and $\Delta\varphi=2\pi/L$. The formulation of constraints C1 and C2 implies that only the phase shifts of the reflecting elements are optimized. Some research works, e.g., [72]–[74], studied the case when both the phase shifts and the amplitudes of the reflecting elements can be optimized subject to the constraint $|\theta_m|\leq 1,\forall m$. The possibility of optimizing both the phase shifts and the amplitudes simultaneously usually results in a more complex hardware design. Hence, we limit our overview to design methods in which only the phase shift can be adjusted. It is worth noting that some works considered practical RIS reflection models with phase-dependent amplitude [75]–[77] or reflection models that account for the mutual coupling among the reflecting elements [78]–[80]. For simplicity, we only consider the case study with phase-independent and element-independent model for $\mathcal{S}_1$ and $\mathcal{S}_2$.

In the following, we discuss the existing contributions on transmission design by focusing our attention on two aspects: 1) Optimization techniques; 2) CSI availability.

A. Optimization Techniques

There are two main difficulties in solving the optimization problem in (56): 1) The optimization problem is non-convex/NP-hard due to the unit modulus constraint or the discrete-valued phase shifts; 2) The phase shifts and the beamforming vectors are coupled as shown in the SINR

expression in (55). As a result, a globally optimal solution is difficult to obtain. Instead, most of the existing works have aimed at finding highly efficient and locally optimal solutions with low computational complexity. Fortunately, several works have confirmed that the suboptimal solutions can achieve improved system performance compared with the performance of systems in the absence of RISs.

If the phase shifts are given, the optimization problem reduces to a conventional beamforming design problem, which has been extensively studied in the literature. Inspired by this consideration, alternating optimization (AO) algorithms are usually applied to decouple the optimization variables. Due to the usually complicated expressions for the data rate/SINR, advanced algorithms such as the weighted minimum mean-square error (WMMSE) algorithm [81] or fractional programming (FP) [82] are often used to convert the original intractable problem into a new tractable but approximated problem. Then, the beamforming vectors at the BS are typically obtained by using existing beamforming design methods. In the following, we focus our attention on the optimization of the phase shifts of the RIS.

1) **Optimization Techniques for Continuous Phase Shifts.**

Existing techniques for optimizing the design of the continuous phase shifts can be classified into the following categories.

(1) *Relaxation and projection [83]–[85]*: The unit modulus constraint on the phase shift can be rewritten as $\mathcal{S}_1 = \{\theta \,|\, |\theta| = 1, \theta \in \mathbb{C}\}$. The idea of this technique is first to relax the non-convex constraint $\mathcal{S}_1$ to the convex constraint $\tilde{\mathcal{S}}_1 = \{\theta \,|\, |\theta| \leq 1, \theta \in \mathbb{C}\}$, and then to project the obtained solution onto the unit-modulus constraint $\mathcal{S}_1$. Accordingly, given the solution θ_m of the relaxed problem, the final solution is $\theta_m^\star = e^{j\varphi_m}$, where φ_m is the phase of θ_m.

(2) *Semidefinite relaxation (SDR) [86]–[92]*: The SDR method is the most common method for optimizing the phase shifts under constraint $\mathcal{S}_1$, i.e., for continuous phase shifts. Define $\mathbf{V} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}$. Then, the unit modulus constraint can be equivalently written as $\mathbf{V} \succ \mathbf{0}$ and $\mathrm{rank}\,(\mathbf{V}) = 1$. Because of the rank one constraint, the transformed problem is still non-convex. Based on the SDR method, the non-convex rank one constraint is removed. The obtained relaxed problem is a convex semidefinite program (SDP), which can be readily solved by using CVX [93]. In general, the obtained relaxed problem is not a rank-one solution, i.e., $\mathrm{rank}\,(\mathbf{V}) \neq 1$. In this case, the Gaussian randomization method [94] is utilized to obtain a rank-one solution.

(3) *Majorization-Minimization (MM) algorithm [95]–[100]*: The MM algorithm is another widely used technique for optimizing the phase shifts of the RIS. The MM algorithm is an iterative optimization method that approximates a difficult problem as a series of more tractable subproblems that are solved iteratively. Assume that the solution of the subproblem at the t-th iteration is $\boldsymbol{\theta}^t$ and the corresponding objective function is $f(\boldsymbol{\theta}^t)$ [5]. Based on the MM algorithm, a surrogate objective function $\tilde{f}\,(\boldsymbol{\theta}|\,\boldsymbol{\theta}^t)$ is constructed, which fulfills the following three conditions: 1) $\tilde{f}\,(\boldsymbol{\theta}^t|\,\boldsymbol{\theta}^t) = f\,(\boldsymbol{\theta}^t)$; 2) $\nabla_{\boldsymbol{\theta}} \tilde{f}\,(\boldsymbol{\theta}|\,\boldsymbol{\theta}^t)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t} = \nabla_{\boldsymbol{\theta}} f\,(\boldsymbol{\theta}^t)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t}$; 3) $\tilde{f}\,(\boldsymbol{\theta}|\,\boldsymbol{\theta}^t) \geq f\,(\boldsymbol{\theta})$. If these conditions are fulfilled, the sequence of the solutions obtained by solving each subproblem will converge. By replacing the original objective function with the constructed function $\tilde{f}\,(\boldsymbol{\theta}|\,\boldsymbol{\theta}^t)$ and removing the constant terms, the subproblem to be solved in each iteration is given by

$$\max_{\boldsymbol{\theta}} \ \mathrm{Re}\left\{\boldsymbol{\theta}^{\mathrm{H}}\mathbf{q}^t\right\} \tag{59a}$$

$$\text{s.t.} \quad |\theta_m| = 1, m = 1, \cdots, M, \tag{59b}$$

where $\mathbf{q}^t$ is a constant complex vector at the t-th iteration. The optimal solution to the optimization problem in (59) is

$$\boldsymbol{\theta}^{t+1} = e^{j\arg(\mathbf{q}^t)}. \tag{60}$$

This procedure is iterated until convergence according to any criterion of convergence. If the phase shifts of the RIS appear in the constraints of the optimization problem, the pricing-based method can be utilized [99].

(4) *Manifold approach [101]–[105]*: There exist different kinds of manifold methods. In this paper, we consider the complex circle manifold (CCM) method [101] as an example. The constraint space in $\mathcal{S}_1$ can be regarded as the product of M complex circles, which is a sub-manifold of $\mathbb{C}^M$ given by

$$\mathcal{S}^M \triangleq \left\{\mathbf{x} \in \mathbb{C}^M : |x_l| = 1, l = 1, 2, \cdots, M\right\}, \tag{61}$$

where x_l is the l-th element of vector $\mathbf{x}$. The main idea of the CCM method is to derive the gradient descent algorithm based on the manifold space given in (61). The optimization problem aims at optimizing the phase shifts to minimize the objective function $\hat{f}\,(\boldsymbol{\theta})$. The main steps can be summarized as follows.

(a) Computation of the gradient in Euclidean space: The search direction for the minimization problem is the opposite of the gradient of $\hat{f}\,(\boldsymbol{\theta})$, which is given by $\boldsymbol{\eta}^t = -\nabla_{\boldsymbol{\theta}} \hat{f}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t}$;

(b) Computation of the Riemannian gradients: The Riemannian gradient of $\hat{f}\,(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^t$ should lie in the tangent space $\mathcal{T}_{\boldsymbol{\theta}^t}\mathcal{S}^M$ [101]. Then, the Riemannian gradient of $\hat{f}\,(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^t$ is obtained by projecting $\boldsymbol{\eta}^t$ onto $\mathcal{T}_{\boldsymbol{\theta}^t}\mathcal{S}^M$, which yields $\mathbf{P}_{\mathcal{T}_{\boldsymbol{\theta}^t}\mathcal{S}^M}(\boldsymbol{\eta}^t) = \boldsymbol{\eta}^t - \mathrm{Re}\{\boldsymbol{\eta}^{t*} \odot \boldsymbol{\theta}^t\} \odot \boldsymbol{\theta}^t$;

(c) Update over the tangent space: Update the point $\boldsymbol{\theta}^t$ on the tangent space $\mathcal{T}_{\boldsymbol{\theta}^t}\mathcal{S}^M$ as $\bar{\boldsymbol{\theta}}^t = \boldsymbol{\theta}^t + \beta \mathbf{P}_{\mathcal{T}_{\boldsymbol{\theta}^t}\mathcal{S}^M}(\boldsymbol{\eta}^t)$, where β is a constant step size;

(d) Retraction operator: This step aims to map $\bar{\boldsymbol{\theta}}^t$ onto the manifold $\mathcal{S}^M$ using the retraction operator $\boldsymbol{\theta}^{t+1} = \bar{\boldsymbol{\theta}}^t \odot \frac{1}{|\bar{\boldsymbol{\theta}}^t|}$. Through iterating steps (a) to (d) until convergence, the final solution is obtained.

(5) *Element-wise block coordinate descent (BCD) [106]–[109]*: The idea of the element-wise BCD algorithm is simple. At the m-th iteration, one reflection coefficient θ_m is optimized by keeping fixed the other reflecting coefficients $\theta_{m'}, m' \neq m, m = 1, \cdots, M$. The algorithm ends after M iterations when all the reflection coefficients are optimized one-by-one while keeping the other fixed. The element-wise BCD algorithm is simple since it is simpler to optimize a single variable rather than optimizing M variables simultaneously.

[5]When the beamforming vectors in $\mathbf{W}$ are given, the objective function in (56) is denoted by $f(\boldsymbol{\theta})$, and the functions in constraint C1 are denoted by $g_i(\boldsymbol{\theta}), \forall i$.

However, the complexity may be high when the number of reflecting elements is large.

(6) *Rank-one equivalents [110], [111]*: Similar to the SDR method, by defining $\mathbf{V} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}$, the unit modulus constraint can be written as $\mathbf{V} \succ \mathbf{0}$ and $\mathrm{rank}\,(\mathbf{V}) = 1$. The rank-one constraint can be equivalently transformed to

$$\mathrm{tr}\,(\mathbf{V}) - \|\mathbf{V}\|_2 = 0. \tag{62}$$

Also, $\mathrm{tr}\,(\mathbf{V}) = \sum_{m=1}^{M} \lambda_m$ and $\|\mathbf{V}\|_2 = \lambda_1$, where λ_m denotes the m-th largest singular value of $\mathbf{V}$. Since $\mathbf{V} \succ \mathbf{0}$ and $\mathbf{V}$ is a non-zero matrix, the equality $\mathrm{tr}\,(\mathbf{V}) - \|\mathbf{V}\|_2 = 0$ holds only when $\lambda_1 > 0$ and $\lambda_m = 0, m = 2, \cdots, M$. Then, at the $(t+1)$-th iteration of the iterative algorithm, a lower-bound for $\|\mathbf{V}\|_2$ at the point $\mathbf{V}^t$ can be derived as

$$\|\mathbf{V}\|_2 \geq \left\|\mathbf{V}^t\right\|_2 + \left\langle \left(\mathbf{V} - \mathbf{V}^t\right), \partial_{\mathbf{V}} \|\mathbf{V}\|_2|_{\mathbf{V}=\mathbf{V}^t} \right\rangle \triangleq f\left(\mathbf{V}; \mathbf{V}^t\right), \tag{63}$$

where $\partial_{\mathbf{V}} \|\mathbf{V}\|_2|_{\mathbf{V}=\mathbf{V}^t}$ is a subgradient of $\|\mathbf{V}\|_2$ with respect to $\mathbf{V}$ at $\mathbf{V} = \mathbf{V}^t$, which is equal to $\mathbf{u}_1\mathbf{u}_1^{\mathrm{H}}$ with $\mathbf{u}_1$ denoting the eigenvector that corresponds to the largest singular value of $\mathbf{V}^t$.

Based on (63), the constraint in (62) can be approximated with the following convex constraint

$$\mathrm{tr}\,(\mathbf{V}) - f\left(\mathbf{V}; \mathbf{V}^t\right) \leq \varepsilon, \tag{64}$$

where ε is a very small positive constant. Then, using (63) and (64), one has $0 \leq \mathrm{tr}\,(\mathbf{V}) - \|\mathbf{V}\|_2 \leq \mathrm{tr}\,(\mathbf{V}) - f\,(\mathbf{V}; \mathbf{V}^t) \leq \varepsilon$. Hence, when ε tends to zero, $\mathrm{tr}\,(\mathbf{V})$ will approach $\|\mathbf{V}\|_2$, which ensures that the rank-one constraint is fulfilled.

(7) *Alternating direction method of multipliers (ADMM) based algorithm [86], [112], [113]*: An auxiliary variable $\boldsymbol{\omega}$ is introduced such that $\boldsymbol{\omega} = \boldsymbol{\theta}$, which can be regarded as a copy of $\boldsymbol{\theta}$. The feasible region of constraint C1 is denoted by $\mathcal{B}$, which, by using the indicator function, can be formulated as follows

$$\mathbb{I}_{\mathcal{B}}\,(\mathbf{W}, \boldsymbol{\theta}) = \begin{cases} f\,(\mathbf{W}, \boldsymbol{\theta}), & \text{if } \{\mathbf{W}, \boldsymbol{\theta}\} \in \mathcal{B} \\ \infty, & \text{otherwise.} \end{cases} \tag{65}$$

Similarly, the feasible region that corresponds to constraint C2, i.e., $\mathcal{S}_1$ can be written as follows

$$\mathbb{I}_{\mathcal{S}_1}\,(\boldsymbol{\omega}) = \begin{cases} 0, & \text{if } \boldsymbol{\omega} \in \mathcal{S}_1 \\ \infty, & \text{otherwise.} \end{cases} \tag{66}$$

Then, the equivalent ADMM reformulation for the optimization problem in (56) is

$$\begin{aligned} \min_{\mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\omega}} \quad & \mathbb{I}_{\mathcal{B}}\,(\mathbf{W}, \boldsymbol{\theta}) + \mathbb{I}_{\mathcal{S}_1}\,(\boldsymbol{\omega}) \\ \text{s.t.} \quad & \boldsymbol{\omega} = \boldsymbol{\theta}. \end{aligned} \tag{67}$$

The augmented Lagrangian of the optimization problem in (67) is

$$\mathcal{L}_{\xi} = \mathbb{I}_{\mathcal{B}}\,(\mathbf{W}, \boldsymbol{\theta}) + \mathbb{I}_{\mathcal{S}_1}\,(\boldsymbol{\omega}) + \frac{\xi}{2} \|\boldsymbol{\theta} - \boldsymbol{\omega} + \boldsymbol{\lambda}\|_2^2, \tag{68}$$

where $\xi > 0$ is a constant penalty parameter, and $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_M]^T$ is the dual variable vector of the constraint $\boldsymbol{\omega} = \boldsymbol{\theta}$. Based on the ADMM algorithm, the variables $\mathbf{W}, \boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are alternately optimized.

The ADMM algorithm is an iterative approach. In the t-th iteration, given $\mathbf{W}^t, \boldsymbol{\theta}^t$ and $\boldsymbol{\omega}^t$, the variables are updated as follows.

(a) Updating $\boldsymbol{\theta}$: The subproblem for updating $\boldsymbol{\theta}$ is

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & f\,(\boldsymbol{\theta}) + \frac{\xi}{2} \left\|\boldsymbol{\theta} - \boldsymbol{\omega}^t + \boldsymbol{\lambda}^t\right\|_2^2 \\ \text{s.t.} \quad & g_i\,(\boldsymbol{\theta}) \geq D_i, i = 1, \cdots, I. \end{aligned} \tag{69}$$

Note that the unit-modulus constraint for $\boldsymbol{\theta}$ is not included in this subproblem, which significantly reduces the complexity of computing $\boldsymbol{\theta}$.

(b) Updating $\mathbf{W}$: The subproblem for updating $\mathbf{W}$ is

$$\begin{aligned} \min_{\mathbf{W}} \quad & f\,(\mathbf{W}) \\ \text{s.t.} \quad & g_i\,(\mathbf{W}) \geq D_i, i = 1, \cdots, I. \end{aligned} \tag{70}$$

(c) Updating $\boldsymbol{\omega}$: The subproblem for updating $\boldsymbol{\omega}$ is

$$\boldsymbol{\omega}^{t+1} = \arg \min_{\boldsymbol{\omega} \in \mathcal{S}_1} \left\|\boldsymbol{\theta}^{t+1} + \boldsymbol{\lambda}^t - \boldsymbol{\omega}\right\|_2^2. \tag{71}$$

The objective of the optimization problem in (71) is to project $\boldsymbol{\theta}^{t+1} + \boldsymbol{\lambda}^t$ onto the feasible set $\mathcal{S}_1$, whose solution is $\boldsymbol{\omega}^{t+1} = e^{j \arg(\boldsymbol{\theta}^{t+1} + \boldsymbol{\lambda}^t)}$.

(d) Updating $\boldsymbol{\lambda}$: The update of $\boldsymbol{\lambda}$ is $\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \boldsymbol{\theta}^{t+1} - \boldsymbol{\omega}^{t+1}$.

(8) *Penalty convex-concave procedure (CCP) [114]–[116]*: The unit modulus constraint can be equivalently rewritten as $1 \leq |\theta_m|^2 \leq 1, \forall m$. Using the successive convex approximation (SCA) method, the non-convex constraint $1 \leq |\theta_m|^2$ can be converted into a series of convex constraints, i.e., $1 \leq 2\mathrm{Re}(\theta_m^* \theta_m^t) - |\theta_m^t|^2$, where θ_m^t is the solution in the t-th iteration. By introducing $2M$ slack variables $\mathbf{b} = [b_1, \cdots, b_{2M}]$, the phase shift optimization problem can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{\theta}, \mathbf{b} \geq \mathbf{0}} \quad & f\,(\boldsymbol{\theta}) - \lambda^t \sum\nolimits_{m=1}^{2M} b_m \\ \text{s.t.} \quad & g_i\,(\boldsymbol{\theta}) \geq D_i, i = 1, \cdots, I, \\ & \left|\theta_m^t\right|^2 - 2\mathrm{Re}(\theta_m^* \theta_m^t) \leq b_m - 1, \forall m, \\ & |\theta_m|^2 \leq 1 + b_{m+M}, \forall m, \end{aligned} \tag{72}$$

where λ^t is the regularization factor to control the feasibility of the constraints in the t-th iteration. After some transformations, the optimization problem in (72) can be solved by CVX, and the detailed procedure to solve this problem, which can be found in references [114]–[116], is omitted here for brevity.

(9) *Barrier function penalty [117], [118]*: The unit modulus constraint can be equivalently written as $\mathrm{tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}\right) = M$ and $\|\boldsymbol{\theta}\|_\infty \leq 1$. Since $\|\boldsymbol{\theta}\|_\infty$ is non-differentiable, the l_p norm with large p can be used to approximate it, i.e., $\|\boldsymbol{\theta}\|_\infty = \lim_{p \to \infty} \|\boldsymbol{\theta}\|_p$. To deal with the constraint $\|\boldsymbol{\theta}\|_p \leq 1$, the logarithmic barrier function $F(x)$ can be used to approximate the penalty of violating the l_p constraint, as

$$F(x) = \begin{cases} -\frac{1}{\kappa} \ln(x), & x > 0, \\ \infty, & x \leq 0, \end{cases} \tag{73}$$

where $\kappa > 0$ is the barrier function penalty factor. For simplicity, constraint C1 is ignored. Accordingly, the phase shift optimization problem can be reformulated as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \ & G(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + F\left(1 - \|\boldsymbol{\theta}\|_p\right) \\ \text{s.t.} \ & \operatorname{tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}\right) = M. \end{aligned} \tag{74}$$

Due to the non-convex constraint, the optimization problem in (74) is still non-convex. To circumvent this issue, a possible solution is to utilize the gradient and projection method, which provides a low complexity but suboptimal solution. Specifically, the gradient of the objective function $G(\boldsymbol{\theta})$ can be formulated as

$$\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) = \frac{\|\boldsymbol{\theta}\|_p^{1-p}}{2\kappa\left(1 - \|\boldsymbol{\theta}\|_p\right)}\boldsymbol{\xi} + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}), \tag{75}$$

where $\boldsymbol{\xi} = \left[\theta_1|\theta_1|^{p-2}, \cdots, \theta_M|\theta_M|^{p-2}\right]^T$.

Since the problem formulation in (74) is a minimization problem, the search direction is opposite to the direction of the gradient in (75). Let $\boldsymbol{\theta}^{(i)}$ denote $\boldsymbol{\theta}$ at the i-th iteration, the search direction in the i-th iteration is $\mathbf{d}_{\mathrm{gd}}^{(i)} = -\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}$ Then, this search direction $\mathbf{d}_{\mathrm{gd}}^{(i)}$ is projected onto the tangent plane of $\operatorname{tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}\right) = M$, as

$$\mathbf{d}_p^{(i)} = \mathbf{d}_{\mathrm{gd}}^{(i)} - \frac{\left(\mathbf{d}_{\mathrm{gd}}^{(i)}\right)^T \left(\boldsymbol{\theta}^{(i)}\right)^* \boldsymbol{\theta}^{(i)}}{\left\|\boldsymbol{\theta}^{(i)}\right\|^2}. \tag{76}$$

Then, the update of $\boldsymbol{\theta}$ in the $(i+1)$-th iteration is

$$\boldsymbol{\theta}^{(i+1)} = (1 - \alpha^\star)\boldsymbol{\theta}^{(i)} + \alpha^\star\sqrt{M}\frac{\boldsymbol{d}_p^{(i)}}{\left\|\boldsymbol{d}_p^{(i)}\right\|^2}, \tag{77}$$

where the parameter $\alpha^\star$ is obtained by

$$\alpha^\star = \arg\max_{\alpha} f\left((1-\alpha)\boldsymbol{\theta}^{(i)} + \alpha\sqrt{M}\frac{\boldsymbol{d}_p^{(i)}}{\left\|\boldsymbol{d}_p^{(i)}\right\|^2}\right). \tag{78}$$

(10) *Accelerated projected gradient (APG) [119]–[123]*: For simplicity, constraint C1 is ignored and only the optimization of the phase shifts is considered. A projection operator $\mathbb{P}_{\mathcal{S}_1}$ is defined as

$$\hat{\boldsymbol{\theta}} = \mathbb{P}_{\mathcal{S}_1}(\boldsymbol{\theta}) \Leftrightarrow \hat{\theta}_m = \begin{cases} \theta_m/|\theta_m|, \text{if } \theta_m \neq 0 \\ 1, \qquad \text{otherwise.} \end{cases} \tag{79}$$

Then, the update of the phase shifts in the $(i+1)$-th iteration is given by

$$\boldsymbol{\theta}_{i+1} = \mathbb{P}_{\mathcal{S}_1}\left(\mathbf{z}_i - \frac{1}{\gamma_i}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{z}_i}\right), \tag{80}$$

where $\mathbf{z}_i = \boldsymbol{\theta}_i + \alpha_i(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1})$, and α_i is updated as

$$\alpha_i = \frac{\xi_{i-1} - 1}{\xi_i}, \xi_i = \frac{1 + \sqrt{1 + 4\xi_{i-1}^2}}{2}. \tag{81}$$

In (80), γ_i is obtained by using the backtracking line search method [124].

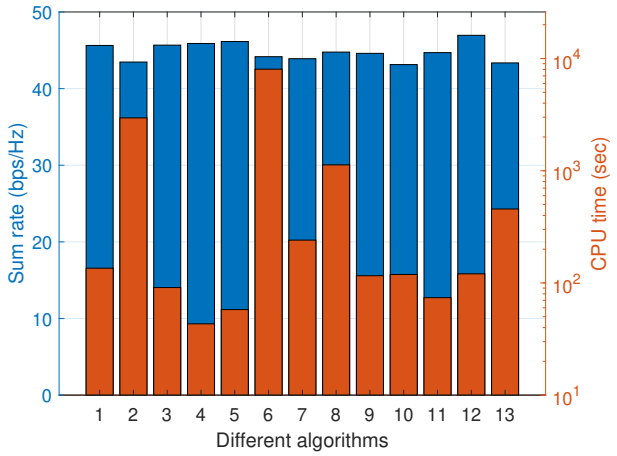

Fig. 4. Sum rate and CPU time consumption of different algorithms, when $N = 10$, $M = 100$, $K = 4$ and SNR=5 dB.

(11) *Gradient descent approach [95], [125]–[127]*: When the objective function $f(\boldsymbol{\theta})$ is differentiable, the optimization problem can be solved by using the gradient descent method. Specifically, let $\boldsymbol{\theta}^t$ be the phase shift vector at the t-th iteration. Then, the optimization variable $\boldsymbol{\theta}$ at the $(t+1)$-th iteration is updated as

$$\boldsymbol{\theta}^{t+1} = \exp\left(j\arg\left(\boldsymbol{\theta}^t - \mu\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t}\right)\right), \tag{82}$$

where μ is the step size and the arg operator is used for satisfying the unit-modulus constraint.

(12) *Heuristic methods [128]–[131]*: When the objective function is analytically involving, the above-mentioned algorithms may not be applicable or the computation of the gradient may be time-consuming. Possible solutions to circumvent this issue include the use of heuristic methods such the genetic algorithms (GA) or the particle swarm optimization (PSO) methods. More details can be found in [128]–[131].

(13) *Deep reinforcement learning [132]–[134]*: Machine learning methods can also be applied to optimize the phase shifts of the RIS. A suitable approach is the use of deep reinforcement learning. In fact, unlike supervised learning methods that require a large number of training labels, deep reinforcement learning based methods do not need training labels and can learn and operate in an online manner. Examples of application of deep reinforcement learning to the optimization of RIS-aided communications can be found in [132]–[134].

Simulation results: Fig. 4 illustrates the performance of the different algorithms discussed in this article in terms of sum rate and CPU run time. All algorithms are represented by the numbers they are introduced above. It can be seen that most of the algorithms for which a closed-form solution for the phase shifts can be found at each iteration (algorithms 3-5, 7, 9-11) provide a high sum rate with a low CPU time (around 100 seconds). However, the time-consuming algorithms (algorithms 1,2,6,8), which are implemented by using CVX, are more flexible to address optimization problems with complex constraints, such as quality of service (QoS) constraints.

2) **Optimization Techniques for Discrete Phase Shifts**

Due to hardware limitations, it is challenging to implement continuous-valued phase shifts in practice [135], [136]. Hence, it is important to study the optimization of RISs subject to discrete-valued phase shifts, which leads to an NP-hard optimization problem. The existing works in this area can be classified into the following categories.

(1) *Rounding method [83], [112], [135], [137], [138]*: The main idea of this method is first to obtain a continuous solution that fulfills the unit modulus constraint. We denote this solution as $\hat{\theta}_m, \forall m$. Then, the obtained solution is rounded to the nearest discrete value in $\mathcal{S}_2$ as

$$\theta_m^\star = \arg\min_{\phi\in\mathcal{S}_2} \left|\hat{\theta}_m - \phi\right|, \forall m. \tag{83}$$

(2) *Binary mode selection method [135], [139]*: Note that $\theta_m = e^{j\varphi_m}, \forall m$, the objective function of the optimization problem can be transformed into a function of φ_m, $\cos(\varphi_m)$ and $\sin(\varphi_m)$ [6]. For each reflecting element m, we introduce the binary vector $\mathbf{x}_m = [x_{1,m}, \cdots, x_{L,m}]^T$ such that $\sum_{l=1}^{L} x_{l,m} = 1, x_{l,m} \in \{0,1\}$. Hence, $\mathbf{x}_m$ can be regarded as a mode selection vector. Define $\mathbf{a} = [0, \Delta\varphi, \cdots, (L-1)\Delta\varphi]^T$, $\mathbf{b} = [1, \cos(\Delta\varphi), \cdots, \cos((L-1)\Delta\varphi)]^T$ and $\mathbf{c} = [0, \sin(\Delta\varphi), \cdots, \sin((L-1)\Delta\varphi)]^T$. Then, we have

$$\varphi_m = \mathbf{a}^T\mathbf{x}_m, \cos(\varphi_m) = \mathbf{b}^T\mathbf{x}_m, \sin(\varphi_m) = \mathbf{c}^T\mathbf{x}_m. \tag{84}$$

Based on (84), the original optimization problem is converted into a binary variable optimization problem, and the branch and bound (BnB) method can be utilized to obtain the globally optimal solution [135]. However, the BnB method has an exponential computational complexity. To reduce the complexity, the authors of [139] proposed to apply the SCA method. Specifically, the binary constraint on $x_{l,m}$ can be equivalently transformed into the following two continuous constraints:

$$x_{l,m} - x_{l,m}^2 \le 0, 0 \le x_{l,m} \le 1. \tag{85}$$

The first constraint in (85) is the difference between two convex functions, and the SCA method can be used.

(3) *Negative square penalty (NSP) [140], [141]*: The NSP method [140] was adopted to solve the discrete phase shift in [141]. Specifically, the discrete phase shift optimization problem can be expressed as

$$\min_{\boldsymbol{\theta}\in\tilde{\mathcal{S}}_2} \; f(\boldsymbol{\theta}) \tag{86}$$

where $\tilde{\mathcal{S}}_2 = \left\{\theta \middle| \theta = e^{j\left(\frac{2\pi}{L}l + \frac{\pi}{L}\right)}, l = 0, \cdots, L-1\right\}$. Based on the NSP methods, we introduce the following problem

$$\min_{\boldsymbol{\theta}\in\bar{\mathcal{S}}_2} \; F(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}) - \lambda\|\boldsymbol{\theta}\|^2, \tag{87}$$

where $\bar{\mathcal{S}}_2 = \text{conv}\tilde{\mathcal{S}}_2$ is the convex hull of $\tilde{\mathcal{S}}_2$, and λ is a penalty parameter. The penalty term $\lambda\|\boldsymbol{\theta}\|^2$ pushes each θ_m to an extreme point of $\bar{\mathcal{S}}_2$. It is important to note that the constraint of the optimization in (87) is convex, which is easier to handle than the original problem formulation in (86). The authors of [140] provided the conditions on λ for the problem in (87) to be equivalent to the original problem in (86). The MM method is utilized to solve the reformulated problem in (87). Let $\boldsymbol{\theta}^i$ denote the value of $\boldsymbol{\theta}$ at the i-th iteration. Then, for any $\boldsymbol{\theta}$, we have

$$F(\boldsymbol{\theta}) \le f(\boldsymbol{\theta}) + \lambda\left\|\boldsymbol{\theta}^i\right\|^2 - 2\text{Re}\left\{\boldsymbol{\theta}^{\text{H}}\boldsymbol{\theta}^i\right\} \triangleq G\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^i\right). \tag{88}$$

Then, at the i-th iteration, the optimization problem is reformulated as follows

$$\min_{\boldsymbol{\theta}\in\bar{\mathcal{S}}_2} \; G\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^i\right). \tag{89}$$

The optimization problem in (89) can be tackled by using the APG method, which can be written as

$$\boldsymbol{\theta}^{i+1} = \mathbb{P}_{\bar{\mathcal{S}}_2}\left(\mathbf{z}^i - \frac{1}{\beta_i}\nabla_{\boldsymbol{\theta}} G\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^i\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^i}\right), \tag{90}$$

where the same notation as for the APG method introduced for the case of continuous phase shifts is used. Specifically, the projection onto the convex set $\bar{\mathcal{S}}_2$ is denoted by $\mathbb{P}_{\bar{\mathcal{S}}_2}(u)$ whose closed-form expression is [140]

$$\mathbb{P}_{\bar{\mathcal{S}}_2}(u) = e^{j\frac{2\pi l}{L}}\left([\text{Re}(\tilde{u})]_0^{\cos(\pi/L)} + j\,[\text{Im}(\tilde{u})]_{-\sin(\pi/L)}^{\sin(\pi/L)}\right),$$

where

$$n = \left\lfloor \frac{\angle u + \pi/L}{2\pi/L} \right\rfloor, \tilde{u} = ue^{-j\frac{2\pi n}{L}}.$$

(4) *Heuristic methods [131]*: Heuristic methods such as the PSO are effective methods to address the discrete phase shift optimization problem. More details can be found in [131].

[6]The objective function may also include the phase differences. For illustration purposes, this is ignored in this paper.

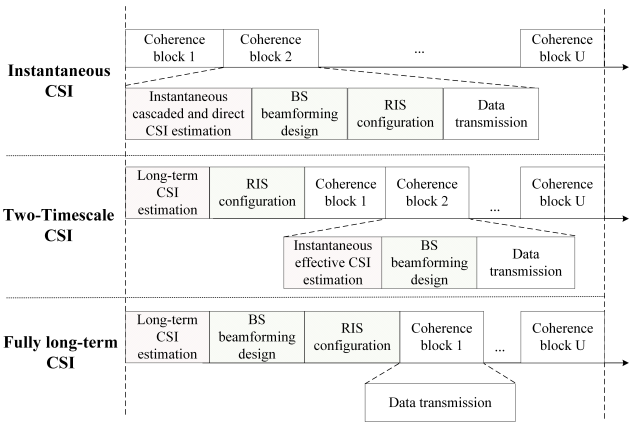

Fig. 5. Illustration of transmission designs based on the various levels of CSI availability. A total of U channel coherence blocks are considered, during which the long-term CSI is fixed.

B. Various Levels of CSI Availability

Depending on the levels of CSI availability, existing contributions on transmission design of RIS-aided communication systems can be classified into three categories: 1) Systems designs based on instantaneous CSI; 2) Systems designs based on the so-called two-timescale CSI; 3) Systems designs that rely on fully long-term CSI. The operation protocol for each category is shown in Fig. 5 and the details for each category are given next.

1) Instantaneous CSI

For the case of instantaneous CSI, the overall CSI of the system is assumed to be available at the BS. The RIS-related channels can be the cascaded channels $\mathbf{G}_k, \forall k$, or the individual channels $\mathbf{H}^{\mathrm{H}}$ and $\mathbf{h}_{r,k}^{\mathrm{H}}, \forall k$. The existing works on the transmission design can be classified into two categories: (1) Perfect instantaneous CSI; (2) Imperfect instantaneous CSI.

(1) Perfect instantaneous CSI: Most of the existing works have considered transmission design based on the assumption that the instantaneous CSI is perfectly available. Based on this assumption, the performance gains provided by introducing an RIS in various wireless applications have been investigated, such as mmWave/terahertz systems [84], [89], [142]–[145], multicell systems [101], [146], [147], physical layer security systems [83], [87], [88], [97], [98], [148], [149], simultaneous wireless information and power transfer (SWIPT) [99], [108], [113], [150]–[154], mobile edge computing networks [74], [111], [155]–[160], multicast networks [96], [161], cognitive radio networks [138], [162], [163], non-orthogonal multiple access [90], [92], [110], [112], [164]–[169], two-way communications [85], [100], and full-duplex (FD) communication [170]. In these works, the AO method was adopted to alternately optimize the beamforming vectors at the BS and the phase shifts at the RIS, and the phase shift optimization problem was addressed using the algorithms summarized in Subsection III-A.

(2) Imperfect instantaneous CSI: As discussed in Section II, channel estimation errors are inevitable. If the estimated CSI is naively regarded as perfect in the transmission design, the obtained solution will likely violate the QoS requirements. This issue is further aggravated in RIS-aided communication systems due to the additional RIS-related channels to be estimated. Hence, it is imperative to consider robust transmission designs by taking into account the channel estimation errors. Most of the early contributions in this area studied the case when the channels from the RIS to the users are imperfect [171]–[176] while regarding the channels from the BS to the RIS as perfect. This approach requires estimation of the BS-RIS and RIS-user channels separately, which is challenging to implement in practice. Instead, the authors of [114] proposed a framework of robust transmission design for RIS-aided multiuser systems based on imperfect CSI of the cascaded channels. Two models were used to characterize the cascaded channel estimation error, namely, the bounded CSI error model and the statistical CSI error model. Specifically, the cascaded and direct channels can be written as

$$\mathbf{h}_{d,k} = \hat{\mathbf{h}}_{d,k} + \tilde{\mathbf{h}}_{d,k}, \mathbf{G}_k = \hat{\mathbf{G}}_k + \tilde{\mathbf{G}}_k, \tag{91}$$

where $\hat{\mathbf{h}}_{d,k}$ and $\hat{\mathbf{G}}_k$ are the estimated channels, and $\tilde{\mathbf{h}}_{d,k}$ and $\tilde{\mathbf{G}}_k$ are the corresponding channel estimation errors. Next, we discuss the two CSI error models.

A. Bounded CSI Error Model: In this case, the channel error is modeled as follows

$$\left\|\tilde{\mathbf{G}}_k\right\|_2 \leq \varepsilon_{c,k}, \left\|\tilde{\mathbf{h}}_{d,k}\right\|_2 \leq \varepsilon_{d,k}, \tag{92}$$

where $\varepsilon_{c,k}$ and $\varepsilon_{d,k}$ quantify the level of channel uncertainty. Under this model, the authors of [114] jointly optimized the active beamforming at the BS and the phase shifts at the RIS so as to minimize the total power consumption under the unit-modulus constraint for the phase shifts and by ensuring that the data rate of each user is above a threshold for all possible channel error realizations. By defining

$$\Omega_{\mathbf{h}_{d,k}} = \left\{\tilde{\mathbf{h}}_{d,k} \in \mathbb{C}^{N\times 1} : \left\|\tilde{\mathbf{h}}_{d,k}\right\|_2 \leq \varepsilon_{d,k}\right\}, \tag{93}$$

$$\Omega_{\mathbf{G}_k} = \left\{\tilde{\mathbf{G}}_k \in \mathbb{C}^{N\times M} : \left\|\tilde{\mathbf{G}}_k\right\|_2 \leq \varepsilon_{c,k}\right\}, \tag{94}$$

the robust design problem can be formulated as

$$\begin{aligned} \min_{\mathbf{W},\boldsymbol{\theta}} \ & \|\mathbf{W}\|_2^2 \\ \text{s.t.} \ & \theta_m \in \mathcal{S}_1 \text{ or } \mathcal{S}_2, \forall m = 1,2,\ldots,M, \\ & R_k(\mathbf{W},\boldsymbol{\theta}) \geq D_k, \forall \tilde{\mathbf{h}}_{d,k} \in \Omega_{\mathbf{h}_{d,k}}, \forall \tilde{\mathbf{G}}_k \in \Omega_{\mathbf{G}_k}, \forall k, \end{aligned} \tag{95}$$

where D_k is the minimum data requirement of user k, $\mathcal{S}_1$ and $\mathcal{S}_2$ are defined in (57) and (58), respectively.

The key difficulty in solving the optimization problem in (95) lies in how to deal with an infinite number of possible channel errors. The most popular technique is to use the $\mathcal{S}$-procedure, which transforms the worst-case constraints into a more tractable form with linear matrix inequalities. The AO and penalty CCP techiques were adopted to solve Problem (95) in [114]. Later, similar techniques were used in RIS-aided cognitive radio networks [115], [138], physical layer security systems [177], and secure cognitive radio communications [178]. In [179], the penalty-based alternating minimization method was proposed, which is guaranteed to converge to a stationary point and is a Karush-Kuhn-Tucker (KKT) solution of the considered problem.

B. Statistical CSI Error Model: For the statistical CSI error model, the channel error is modeled as a random variable. When the linear minimum mean square error (LMMSE) is used for channel estimation, the channel estimation error generally follows a complex Gaussian distribution. Hence, the channel error can be modeled as

$$\mathrm{vec}\left(\tilde{\mathbf{G}}_k\right) \sim \mathcal{CN}\left(0, \boldsymbol{\Sigma}_{c,k}\right), \boldsymbol{\Sigma}_{c,k} \succ 0, \forall k, \tag{96}$$

$$\tilde{\mathbf{h}}_{d,k} \sim \mathcal{CN}\left(0, \boldsymbol{\Sigma}_{d,k}\right), \boldsymbol{\Sigma}_{d,k} \succ 0, \forall k, \tag{97}$$

where $\boldsymbol{\Sigma}_{c,k} \in \mathbb{C}^{MN\times MN}$ and $\boldsymbol{\Sigma}_{h,k} \in \mathbb{C}^{N\times N}$ are positive definite error covariance matrices. Due to the randomness of the channel error, the design objective is mainly to optimize the beamforming vectors at the BS and the phase shifts at the RIS while ensuring a minimum non-outage probability. Specifically, the robust design problem can be formulated as

$$\begin{aligned} \min_{\mathbf{W},\boldsymbol{\theta}} \ & \|\mathbf{W}\|_2^2 \\ \text{s.t.} \ & \theta_m \in \mathcal{S}_1 \text{ or } \mathcal{S}_2, \forall m = 1,2,\ldots,M, \\ & \Pr\left\{R_k(\mathbf{W},\boldsymbol{\theta}) \geq D_k\right\} \geq 1-\rho_k, \forall k, \end{aligned} \tag{98}$$

where ρ_k is the maximum outage probability.

The main challenge in solving the optimization problem in (98) lies in the fact that the rate outage probability constraints do not admit a closed-form expression. One method to circumvent this issue is the Bernstein-Type inequality [180], which transforms an intractable outage probability constraint into a tractable linear matrix inequality [114]. This method has been

applied to RIS-aided physical layer security systems [177], [181] and RIS-aided vehicular communications [116]. Another method that can be applied is the constrained stochastic SCA algorithm proposed in [182], which is shown to guarantee the desired outage probability performance of the users. In addition, the robust design of RIS-aided cognitive radio networks was considered in [115], where it was ensured that the probability that the interference power perceived at the primary users (PUs) is below an interference temperature (IT) limit was larger than a threshold. Both the triangle inequality and the inverse Chi-square distribution techniques were invoked to derive a tractable approximate constraint.

2) Two-Timescale CSI

As illustrated in Fig. 5, transmission designs based on the knowledge of the instantaneous CSI require that the BS needs to estimate the cascaded channel $\mathbf{G}_k$ and direct channel $\mathbf{h}_{d,k}$ in each channel coherence block, in which the number of time slots required for channel training is often proportional to the number of reflecting elements, e.g., for the unstructured channel models discussed in Section II. Since the RIS is not endowed with power amplification and signal processing capabilities, it is expected to be equipped with hundreds or even thousands of reflecting elements for ensuring the desired coverage. As a result, the channel training overhead may be excessive, and there may be only a few or even no time slots left for data transmission. Most recently, in instantaneous CSI cases, it was revealed in [130], [183], [184] that the net data rate that accounts for the penalty due to channel estimation overhead first increases and then decreases with the number of reflecting elements. It is well known that time division duplexing (TDD) is the preferred option in massive MIMO systems since the channel training overhead depends on the number of users, and is not related to the number of BS antennas. In RIS-aided wireless systems, no matter whether TDD or frequency division duplexing (FDD) is used, the required channel training overhead is always proportional to the number of reflecting elements under the assumption of unstructured channel models. Furthermore, in each coherence block with a duration of at most several hundred milliseconds, the BS needs to optimize the phase shifts and active beamforming vectors, which requires the BS to have sufficient computational capabilities. In addition, the phase shifts computed at the BS need to be sent to the RIS controller for updating the phase shifts in each coherence block. This would incur high feedback overhead.

To address these issues, a novel two-timescale beamforming design was first proposed in [185] and its transmission protocol is shown in Fig. 5. The main idea is that the active beamforming vectors at the BS are designed based on the instantaneous effective/aggragated BS-user channels that are the superposition of the direct and RIS-reflected channels, while the phase shifts at the RIS are designed based on long-term CSI, such as their distribution parameters including channel mean and channel covariance matrices. In each coherence block, only the instantaneous effective channel of each user needs to be estimated and the channel training overhead is equal to the number of users, which is the same as for legacy massive MIMO systems without RISs. Furthermore, since the long-term CSI remains usually invariant for a large number of channel coherence blocks, the phase shifts of the RIS can be updated at a much lower rate than the fast fading fluctuations, which significantly reduces the computational burden and feedback overhead.

To better understand, let us consider the system model in Fig. 1. All the channels are assumed to be subject to correlated Rician fading

$$\mathbf{H} = \sqrt{\beta}\left(\sqrt{\frac{\delta}{1+\delta}}\bar{\mathbf{H}} + \sqrt{\frac{1}{1+\delta}}\tilde{\mathbf{H}}\right),$$
$$\mathbf{h}_{r,k} = \sqrt{\alpha_k}\left(\sqrt{\frac{\varepsilon_k}{1+\varepsilon_k}}\bar{\mathbf{h}}_{r,k} + \sqrt{\frac{1}{1+\varepsilon_k}}\tilde{\mathbf{h}}_{r,k}\right), \forall k,$$
$$\mathbf{h}_{d,k} = \sqrt{\gamma_k}\left(\sqrt{\frac{\varpi_k}{1+\varpi_k}}\bar{\mathbf{h}}_{d,k} + \sqrt{\frac{1}{1+\varpi_k}}\tilde{\mathbf{h}}_{d,k}\right), \forall k,$$

where β, α_k and γ_k are the large-scale path loss coefficients, δ, ε_k and ϖ_k are the Rician factors, $\bar{\mathbf{H}}$, $\bar{\mathbf{h}}_{r,k}$, and $\bar{\mathbf{h}}_{d,k}$ are LoS components, $\tilde{\mathbf{H}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\mathrm{HR}} \otimes \mathbf{R}_{\mathrm{HB}})$, $\tilde{\mathbf{h}}_{r,k} \sim \mathcal{CN}(0, \mathbf{R}_{\mathrm{h_rR},k})$ and $\tilde{\mathbf{h}}_{d,k} \sim \mathcal{CN}(0, \mathbf{R}_{\mathrm{h_dB},k})$ are NLoS components with $\mathbf{R}_{\mathrm{HR}}$, $\mathbf{R}_{\mathrm{HB}}$, $\mathbf{R}_{\mathrm{h_rR},k}$ and $\mathbf{R}_{\mathrm{h_dB},k}$ being the corresponding spatial covariance matrices. When the Rician factors are equal to zero, the channels become correlated Rayleigh channels. When the spatial covariance matrices are identity matrices, the channels are independent and identically distributed Rician channels. When both conditions hold, the channels reduce to independent and identically distributed Rayleigh channels. The spatial covariance matrices can be estimated by using the method in [186] or the model proposed in [187]. Also, the LoS components mainly depend on location/angle information, which can be estimated using the method in [67]. In general, the NLoS components vary rapidly and need to be estimated at each coherence block, while the LoS components change slowly and usually remain constant over a number of channel coherence blocks.

Let $\mathbf{q}_k \triangleq \mathbf{h}_{d,k} + \mathbf{H}\boldsymbol{\Theta}\mathbf{h}_{r,k}$ be the effective channel of user k and let $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_K]$ be the collection of all effective channels. In the downlink, the beamforming vector $\mathbf{w}_k$ is a function of the effective channels $\mathbf{Q}$, i.e., $\mathbf{w}_k = z_k(\mathbf{Q})$. By substituting $\mathbf{w}_k$ into the optimization problem in (56), the original problem can be converted into an optimization problem that depends only on the optimization variable $\boldsymbol{\theta}$

$$\begin{aligned} \min_{\boldsymbol{\theta}} \ & \mathbb{E}\{f(\boldsymbol{\theta})\} \\ \text{s.t.} \ & \theta_m \in \mathcal{S}_1 \text{ or } \mathcal{S}_2, \forall m = 1, 2, \ldots, M, \\ & \mathbb{E}\{g_i(\boldsymbol{\theta})\} \geq D_i, i = 1, \cdots, I, \end{aligned} \tag{99}$$

where the expectation is taken over the NLoS components of the channels. The optimization problem in the uplink can be formulated *mutatis mutandis*.

In the following, we discuss the existing works from two aspects: (1) Downlink transmission; (2) Uplink transmission.

(1) *Downlink transmission*: In the two-timescale scheme, to reduce the complexity of solving the optimization problem in (99), most existing works have adopted maximum-ratio-transmission (MRT) at the BS:

$$\mathbf{w}_k = \sqrt{P_k}\frac{\mathbf{q}_k}{\|\mathbf{q}_k\|}, \tag{100}$$

where P_k is the transmission power of user k.

For the single-user case, the MRT precoder is optimal. This setup has been studied in [185], [188]–[190]. Specifically, by assuming Rician fading for the RIS-related channels and Rayleigh fading for the direct channel, an upper bound for the ergodic data rate was derived in [185] using Jensen's inequality, based on which the optimal phase shift was obtained in closed form as a function of the LoS components of the channels. The work in [185] was further extended in [188] to the case where there is a co-channel BS transmitting interference signals. The authors of [189] first derived an approximated closed-form expression of the coverage probability of a multi-RIS-aided system in the presence of correlated Rayleigh fading, by using the deterministic equivalent analysis, and then optimized the phase shifts relying on the projected gradient method. Instead of using Jensen's inequality, the stochastic gradient descent method was used in [190] for solving the phase shift optimization problem, which is applicable to any channel distribution.

In the multiuser case, MRT at the BS was adopted in [131], [191]. In particular, by taking into account the use of low-resolution digital-analog converters (DACs) at the BS, the authors of [131] derived an approximate expression for the ergodic data rate of RIS-aided massive MIMO systems. Because of the complicated expression of the obtained data rate, the heuristic PSO method was invoked to solve the rate maximization problem under the assumptions of either continuous or discrete phase shifts. Simulation results showed that three quantization bits for the phase shifts are sufficient to obtain similar performance as for the continuous phase shifts. By considering the channel estimation error of the effective channel, the authors of [191] derived deterministic equivalents for the sum rate, where a LoS deterministic RIS-BS channel and Rician fading RIS-user/BS-user channels were assumed. The projected gradient ascent-based algorithm was used for solving the phase shift optimization problem.

In addition to the MRT, another beamforming vector that can be utilized at the BS is the optimal linear precoder studied in [192], which is the optimal solution to the problem of maximizing the minimum data rate. Specifically, the authors of [192] developed deterministic approximations of the minimum data rate using random matrix theory (RMT) tools by assuming a deterministic BS-RIS channel and correlated Rayleigh RIS-user channels.

In [137] and [193], more sophisticated algorithms, such as deep unfolding and the stochastic successive convex approximation, were utilized to jointly optimize the beamforming vectors and the phase shifts. Although the algorithms in [137] and [193] are applicable to any channel distributions, the computational complexity would be excessive when the number of antennas at the BS is large, and thus not suitable for massive MIMO systems.

(2) *Uplink transmission*: Similar to the downlink, maximum-ratio-combining (MRC) at the BS is widely used in the uplink. In this case, the decoding vector is

$$\mathbf{v}_k = \mathbf{q}_k. \tag{101}$$

The MRC scheme has been studied in [126], [128]–[130], [194], [195]. The achievable data rate was derived in [128] for RIS-aided massive MIMO systems assuming Rician fading for the BS-RIS and RIS-user channels. Based on the obtained result, the scaling laws with respect to the number of RIS elements and the number of BS antennas were analyzed. The work was extended in [129] to the case study where the direct link between the BS and the users are present, and the authors provided the conditions under which RIS-aided massive MIMO systems outperform conventional massive MIMO systems. In [194], the achievable data rate was analyzed in the presence of transceiver hardware impairments (HWIs) and RIS phase noise over Rician fading RIS-related channels, and it was demonstrated that the hardware imperfections do not affect the power scaling law. Most recently, the authors of [130] adopted the LMMSE estimator to estimate the effective channel, and derived a closed-form expression for the approximate achievable data rate by taking into account the channel estimation errors. A framework for the power scaling law analysis was introduced in [130] under various channel fading distributions. Due to the complicated data rate expression, the phase shift optimization problem was solved by using the heuristic GA method in [128]–[130], [194]. In addition to the effective channel estimation error, the impact of both the transceiver HWIs and the RIS phase noise on the achievable data rate was analyzed in [126] for transmission over a deterministic BS-RIS channel and other correlated Rayleigh fading channels. In [195], the achievable data rate for imperfect effective channels was derived where all channels were assumed to undergo correlated Rayleigh fading. The authors of [195] demonstrated that the channel hardening and favorable propagation conditions still hold for the effective channels.

Besides the MRC, another low-complexity decoding vector is the zero-forcing (ZF) detector. Let $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_K]$ denote the collection of all users' decoding vectors. Then, the ZF detector at the BS is

$$\mathbf{V} = \mathbf{Q}\big(\mathbf{Q}^{\mathrm{H}}\mathbf{Q}\big)^{-1}. \tag{102}$$

The two-timescale design for RIS-aided massive MIMO with ZF was studied in [125], [196]. By approximating the non-central Wishart distribution with a central Wishart distribution, a closed-form expression of the ergodic data rate was derived in [125] under the assumption of a BS-RIS channel that undergoes Rician fading, and LoS RIS-user channels and BS-user channels that undergo Rayleigh fading. Based on the derived expression of the ergodic data rate, the gradient ascent algorithm was proposed to optimize the phase shifts of the RIS. The simulation results in [125] demonstrated that the ZF detectors significantly outperform the MRC detectors in RIS-aided massive MIMO systems. The reason is that the users share the same RIS-BS channel and then their cascaded channels are highly correlated over Rician fading channels. Therefore, by effectively eliminating the interference, the ZF detector provides a higher ergodic rate as compared with the MRC. The analysis in [125] was extended to the case with imperfect effective CSI [196], where closed-form expressions for the uplink data rate were derived. The analytical results

showed that the rate of all the users scales at least on the order of $O\left(\log_2\left(MN\right)\right)$. The low-complexity MM algorithm was proposed to optimize the RIS phase shifts.

In the uplink, the optimal decoding vector is the LMMSE receiver given by

$$\mathbf{v}_k = \left(\sum\nolimits_{i=1}^{K} P_i \mathbf{q}_i \mathbf{q}_i^{\mathrm{H}} + \sigma^2 \mathbf{I}_N\right)^{-1} \mathbf{q}_k. \tag{103}$$

Using the LMMSE receiver, the authors of [197] analyzed the approximate SINR of each user by leveraging the deterministic equivalent method, where the BS-RIS channel is assumed to be deterministic and the other channels follow a correlated Rayleigh distribution. In [197], both RIS phase noise and transceiver hardware impairments were taken into account, and the phase shifts of the RIS were optimized using the gradient ascent algorithm to maximize the minimum SINR.

3) Fully long-term CSI

In the two-timescale design, the instantaneous effective CSI needs to be estimated in each coherence block. To further reduce the channel estimation overhead, an appealing approach is based on transmission designs that require only long-term CSI. The operation procedure is shown in Fig. 5. Specifically, at the beginning of the transmission, the BS estimates or measures the long-term CSI, based on which the BS computes the beamforming vector and the phase shifts that are used in all the subsequent coherence blocks until the long-term CSI changes. By adopting the same channel fading distribution as in the two-timescale CSI subsection, the long-term CSI problem is formulated as

$$\begin{aligned} \min_{\boldsymbol{\theta},\mathbf{W}} \ & \mathbb{E}\left\{f\left(\mathbf{W},\boldsymbol{\theta}\right)\right\} \\ \text{s.t.} \ & \theta_m \in \mathcal{S}_1 \text{ or } \mathcal{S}_2, \forall m = 1,2,\ldots,M, \\ & \mathbb{E}\left\{g_i\left(\mathbf{W},\boldsymbol{\theta}\right)\right\} \geq D_i, i = 1,\cdots,I, \end{aligned} \tag{104}$$

where the expectation is taken over the NLoS components of the channels. In contrast to the two-timescale optimization problem in (99), the beamforming vectors at the BS are designed based on long-term CSI. The main difficulty in solving Problem (104) lies in the lack of explicit expressions for the objective function and constraints, i.e., the expectation in (104) cannot in general be formulated in a closed-form expression. In general, three optimization techniques can be utilized to circumvent this issue: (1) Applying Jensen's inequality; (2) Utilizing large system analysis; (3) Leveraging deep reinforcement learning methods.

(1) *Jensen's inequality*: Jensen's inequality is an effective and simple method to tackle the randomness in the channels, and has been widely used in the existing literature [198]–[203]. The main idea of this method is to derive an upper bound for the function $f(\cdot)$ in (104) that is a function of the optimization variables and the long-term CSI. In these works, usually, the function $f(\cdot)$ in (104) is the data rate. Once the upper-bound is obtained, the many optimization methods overviewed in Subsection III-A can be invoked for solving the approximated phase shift optimization problem. Based on this approach, various research works can be found in the literature. The single-user case was studied in [198] and [199] for Rician fading channels and correlated Rician fading channels, respectively. Given the beamforming vectors, the optimal phase shifts were derived in closed form in [198], and the SDR method was used in [199] to solve the phase shift optimization problem. The authors of [200] derived an upper bound for the data rate in both downlink and uplink multiuser scenarios, based on which the ADMM method was used for solving the phase shift optimization problem. The more complex multicell networks scenario in the presence of interference channels was studied in [201] and [202] assuming Rician fading channels, and the phase shifts were optimized using the GA and CCM methods, respectively. Jensen's inequality was used in [203] to derive an upper bound of the ergodic data rate in RIS-aided UAV communication systems, based on which the trajectory of the UAV and the phase shifts of the RIS were jointly optimized.

(2) *Large system analysis*: Another widely used method to tackle the computation of the expectation in (104) relies on large system analysis [204]–[207]. The main idea is to apply the replica method in large dimension random matrix theory to derive a deterministic approximation of the ergodic data rate, based on which various optimization techniques can be used to optimize the phase shifts. Specifically, the authors of [204] and [205] considered the RIS-aided single-user case under correlated Rician fading channels and the double-scattering channel model, respectively. The gradient decent method and element-wise BCD method were used for determining the RIS phase shifts in [204] and [205], respectively. Recently, the authors of [206] and [207] studied RIS-aided MIMO multiple-access channels. Specifically, in [206], the BS-RIS channel is assumed to be deterministic and the RIS-user channels are distributed according to a spatially correlated Rayleigh fading distribution. The MM method is used to solve the phase shift optimization problem. In [207], all the channels are assumed to be subject to correlated Rician fading and the gradient decent method is used to solve the phase shift optimization problem.

(3) *Deep reinforcement learning*: For example, a novel deep deterministic policy gradient (DDPG) based algorithm was proposed in [208] for solving the optimization problem in (104). The idea is that the BS first estimates the long-term CSI, based on which the BS randomly generates a set of instantaneous CSI samples based on the channel distribution in an offline manner. Then, the generated data set is employed for DDPG training. The final trained solutions can be used in subsequent coherence blocks. Compared with the above two methods, the appealing feature of the DDPG approach is the low training overhead, while providing almost the same performance as the existing methods.

In [118] and [209], the authors considered the case study in the presence of imperfect long-term CSI. Specifically, in [118], the effective angles from the BS to the user are first derived, and are then used for optimizing the BS beamforming and the RIS phase shifts by accounting for the errors in angle estimation. In [209], the authors exploited user location information to obtain the angle information and derived the corresponding angle error distribution based on the distribution of estimated location information. The achievable data rate was then derived based on the statistical information of the angle error distribution.

Simulation results: Fig. 6 illustrates the performance

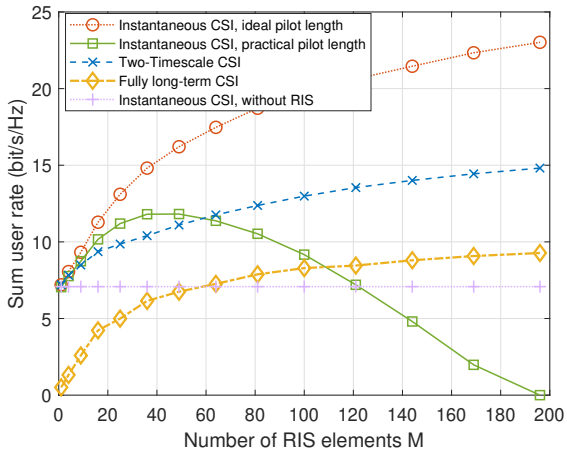

Fig. 6. Comparison of achievable rate under three different levels of CSI availability in an RIS-aided massive MIMO system, where $K = 4$, $U = 50$, $\delta = 1$, $\varepsilon_k = 20, \forall k$ and $N = 100$. The number of time slots in each coherence block is $T_c = 196$ [210], [211]. For brevity, we set $\varpi_k = 0, \forall k$ and consider the rank-1 LoS BS-RIS channel $\overline{\mathbf{H}} = \mathbf{a}_N \mathbf{a}_M^H$.

of transmission schemes based on instantaneous CSI, two-timescale CSI, and fully long-term CSI. We consider the uplink transmission in massive MIMO systems where low-complexity MRC is employed at the BS. The channel estimation overhead for long-term CSI estimation is ignored. The pilot lengths required for acquiring the necessary instantaneous CSI and two-timescale CSI are equal to $T_{\text{ins}} = K + M + \max\left(K-1, (K-1)\left\lceil M/N \right\rceil\right)$ [52] and $T_{\text{two}} = K$ [130], respectively. The fully long-term CSI scheme does not require estimation of the instantaneous (effective) CSI, which leads to a pilot length equal to $T_{\text{ful}} = 0$.

As far as the instantaneous CSI case is concerned, in the u-th coherence block, the cascaded and direct channels of user k and the phase shifts of the RIS are denoted by $\mathbf{G}_k^u$, $\mathbf{h}_{d,k}^u$ and $\boldsymbol{\theta}^u$, respectively. Then, in the u-th coherence block, the instantaneous CSI-based MRC decoding vector is set as $(\mathbf{w}_k^u(\boldsymbol{\theta}^u))^H = (\mathbf{G}_k^u \boldsymbol{\theta}^u + \mathbf{h}_{d,k}^u)^H$ and the corresponding SINR of user k is

$$\text{SINR}_{\text{ins},k}^u\left(\boldsymbol{\theta}^u\right) = \frac{P_k \left| \left(\mathbf{w}_k^u(\boldsymbol{\theta}^u)\right)^H \left(\mathbf{G}_k^u \boldsymbol{\theta}^u + \mathbf{h}_{d,k}^u\right) \right|^2}{\sum\limits_{i=1, i\neq k}^{K} P_i \left| \left(\mathbf{w}_k^u(\boldsymbol{\theta}^u)\right)^H \left(\mathbf{G}_i^u \boldsymbol{\theta}^u + \mathbf{h}_{d,i}^u\right) \right|^2 + \sigma^2 \left\| \mathbf{w}_k^u(\boldsymbol{\theta}^u) \right\|^2}, \tag{105}$$

and the corresponding average sum user rate is given by

$$R_{\text{ins}} = \left(1 - \frac{T_{\text{ins}}}{T_c}\right) \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{K} \log_2\left(1 + \text{SINR}_{\text{ins},k}^u(\boldsymbol{\theta}^u)\right), \tag{106}$$

where the factor $1 - \frac{T_{\text{ins}}}{T_c}$ accounts for the rate loss due to the pilot overhead.

As far as the two-timescale CSI scheme is concerned, the effective channel of user k in the u-th coherence block is denoted by $\mathbf{q}_k^u = \mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,k}^u + \mathbf{h}_{d,k}^u$, where the phase shift matrix $\boldsymbol{\Theta}$ remains constant for all coherence blocks $1 \leq u \leq U$. In the u-th coherence block, the instantaneous CSI-based MRC decoding vector is set equal to $(\mathbf{w}_k^u(\boldsymbol{\Theta}))^H = (\mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,k}^u + \mathbf{h}_{d,k}^u)^H$ and then the SINR of user k is given by

$$\text{SINR}_{\text{two},k}^u(\boldsymbol{\Theta}) = \frac{P_k \left| \left(\mathbf{w}_k^u(\boldsymbol{\Theta})\right)^H \left(\mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,k}^u + \mathbf{h}_{d,k}^u\right) \right|^2}{\sum\limits_{i=1, i\neq k}^{K} P_i \left| \left(\mathbf{w}_k^u(\boldsymbol{\Theta})\right)^H \left(\mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,i}^u + \mathbf{h}_{d,i}^u\right) \right|^2 + \sigma^2 \left\| \mathbf{w}_k^u(\boldsymbol{\Theta}) \right\|^2}. \tag{107}$$

It is worth noting that the phase shift $\boldsymbol{\Theta}$ needs to be optimized only once in the considered series of coherence blocks. However, in the instantaneous CSI case, the phase shift needs to be optimized U times. The average sum user rate of the two-timescale scheme is given by

$$R_{\text{two}} = \left(1 - \frac{K}{T_c}\right) \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{K} \log_2\left(1 + \text{SINR}_{\text{two},k}^u(\boldsymbol{\Theta})\right). \tag{108}$$

As for the fully long-term CSI case is concerned, the long-term CSI-based MRC beamformer is set equal to $(\mathbf{w}_k(\boldsymbol{\Theta}))^H = \left(\mathbb{E}\left\{\mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,k}^u + \mathbf{h}_{d,k}^u\right\}\right)^H$ which only contains large-scale LoS channel components and remains constant for all coherence blocks. Therefore, in the u-th coherence block, the SINR of user k is given by

$$\text{SINR}_{\text{ful},k}^u(\boldsymbol{\Theta}) = \frac{P_k \left| \left(\mathbf{w}_k(\boldsymbol{\Theta})\right)^H \left(\mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,k}^u + \mathbf{h}_{d,k}^u\right) \right|^2}{\sum\limits_{i=1, i\neq k}^{K} P_i \left| \left(\mathbf{w}_k(\boldsymbol{\Theta})\right)^H \left(\mathbf{H}^u \boldsymbol{\Theta} \mathbf{h}_{r,i}^u + \mathbf{h}_{d,i}^u\right) \right|^2 + \sigma^2 \left\| \mathbf{w}_k(\boldsymbol{\Theta}) \right\|^2}, \tag{109}$$

and the corresponding average sum rate is

$$R_{\text{ful}} = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{K} \log_2\left(1 + \text{SINR}_{\text{ful},k}^u(\boldsymbol{\Theta})\right). \tag{110}$$

The phase shifts of the above three schemes are obtained by using the GA method with the objective of optimizing the sum data rate. However, in the instantaneous CSI scheme, the phase shift of the RIS, $\boldsymbol{\theta}^u$, $1 \leq u \leq U$, needs to be designed in each coherence block. In the two-timescale and the fully long-term CSI schemes, we only need to design the variable $\boldsymbol{\Theta}$ once.

Fig. 6 clearly unveils the performance tradeoffs of the considered schemes as a function of the level of CSI available. The instantaneous CSI scheme with an ideal pilot length equal to $T_{\text{ins}}^{(\text{ideal})} = K = T_{\text{two}}$ offers, as expected, the best performance. If, the actual pilot training overhead is taken into account, however, we note that the average rate first increases and then decreases as a function of the number M of RIS elements. On the other hand, the fully long-term CSI scheme has the lowest pilot overhead, but it offers a relatively low achievable rate since the decoding vector at the BS cannot be adjusted according to the instantaneous effective CSI. By contrast, the two-timescale scheme offers good performance while maintaining a low pilot overhead.

IV. RIS-aided Radio Localization

In wireless communication networks, radio localization [212] offers a viable alternative for obtaining user location information in GPS-denied environments [213]. Radio localization techniques are based on the general idea that the radio signals provide information on the position of network nodes. More precisely, the location of agent nodes (e.g., mobile devices or vehicles) are estimated with the aid of known-position anchor nodes (e.g., BSs or APs) and the exchange of radio signal between the anchor and agent nodes. In general, the position of the agent node is estimated by using a two-step approach [214]. First, the distance/angle-related measurements are extracted from the received signal, then the location is estimated from the measurements, such as the time of arrival (ToA), the time difference of arrival (TDoA), the received signal strength (RSS), the AoA, and the AoD.

Third generation (3G) communication systems can provide a positioning accuracy of the order of tens of meters by using TDoA measurements [215]. This accuracy is improved to some extent in fourth generation (4G) systems. Existing studies have shown that the position errors in 5G mmWave communication systems are on the order of centimeters [213], [216], [217]. Due to thriving new applications such as smart factories, automated/assisted driving, and augmented reality, the requirements in terms of positioning accuracy for 5G/6G communication networks are becoming increasingly stringent. In addition, the reliability of the localization provided by 5G/6G communications is of particular importance. Since 5G/6G systems can be deployed in high-frequency mmWave and THz bands, the links are vulnerable to obstacles. Since LoS propagation is usually required for accurately estimating the location, existing localization methods result in prohibitively large estimation errors if the LoS link is blocked.

The use of RISs can yield reliable and high-precision position estimates at a low cost and high energy efficiency. The RISs can be integrated into existing radio localization systems to co-work with other anchor nodes, and thus provide better positioning estimates of the agent nodes. The benefits of RISs for positioning are as follows. First, an RIS can establish a virtual LoS link when the LoS link is blocked. Thus, the RIS can restore the positioning capability of the network when the GPS or the BS signals are weak. Second, RISs can be regarded, unlike active anchor nodes, as quasi-passive anchor nodes that need no power amplifiers and radio frequency chains. Therefore, they can provide high-precision estimate of the locations with low hardware cost and low energy consumption. Third, RISs can be constructed with a large physical aperture. Hence, they can offer a higher angular resolution, which is appealing for radio localization. Fourth, unlike a non-reconfigurable scatterer in the environment, RISs can provide a high beamforming gain by tuning the phase shifts of the reflecting elements. As a result, the use of RISs provides several promising opportunities for assisting localization systems in next generation wireless networks.

Current state-of-the-art research on RIS-aided radio localization has mainly focused on three aspects: performance analysis, development of algorithms, and the interplay between communication and localization. As far as the performance analysis is concerned, the authors of [218] studied the theoretical performance of a 2D RIS-aided mmWave positioning system by computing the CRB in the far field region. The authors of [219] derived the Fisher information matrix (FIM) and the CRB for evaluating the performance of a three-dimensional (3D) RIS-assisted positioning system, where a near-field channel model was utilized. The authors of [220] derived the CRB for evaluating the localization and orientation performance of synchronous and asynchronous signalling schemes in an RIS-assisted localization system in the near-field region. As far as the localization algorithms, the authors of [221] considered an RIS-assisted 3D localization system, where the AoAs and AoDs were estimated by using the MLE algorithm, and user localization was obtained via a Taylor series algorithm. The localization accuracy of an indoor localization system was improved in [222] and [223] by choosing proper RIS reflection coefficients that maximize the differences of RSS values among adjacent locations. As far as the interplay between communication and localization is concerned, the localization accuracy and the data rate were balanced by optimizing the time allocation of an RIS-aided wireless communication system in [224]. In particular, the CRB and the effective achievable data rate were used as the performance metrics. The authors of [225] proposed an adaptive RIS phase shifter design based on hierarchical codebooks and feedback from the mobile user for enabling accurate localization and high data rate transmission.

Under the spherical wave-front channel model, there exist only a few contributions that studied the RIS-aided radio localization problem. A near-field codebook was developed in [226] for extremely large-scale RIS (XL-RIS) beam training by dividing the two-dimensional (2D) plane into several sampled points in the x-y coordinate system. Also, the authors of [227] showed that the characteristics of the transmitted signal (transmit antenna type, size, orientation etc.) may profoundly affect the received signals in the near field, which needs to be taken into consideration for high-precision localization. In [228], the FIM was analysed for an uplink localization system using an RIS-based lens, and the PEB and OEB were also evaluated by exploiting the wavefront curvature. The theoretical localization performance of multipath-aided localization in both LoS and NLoS conditions was characterized in [229], where it was shown that using the RIS's reflected signal wavefront curvature in the near field is enough to deduce the user's position even in the absence of a direct path.

In the following, we discuss RIS-aided localization with focus on two channel conditions: 1) Far-field channel model; 2) Near-field channel model.

A. Far-Field Localization Techniques

1) System Model: In Fig. 7, as an illustrative example we consider a mmWave MIMO system for user localization in a 3D space. The considered system model consists of a BS equipped with a ULA with N_t antennas, a mobile user (MU) with a ULA with N_r antennas, and $K-1$ RISs. Each RIS is equipped with a square UPA with L^2 reflecting elements. The

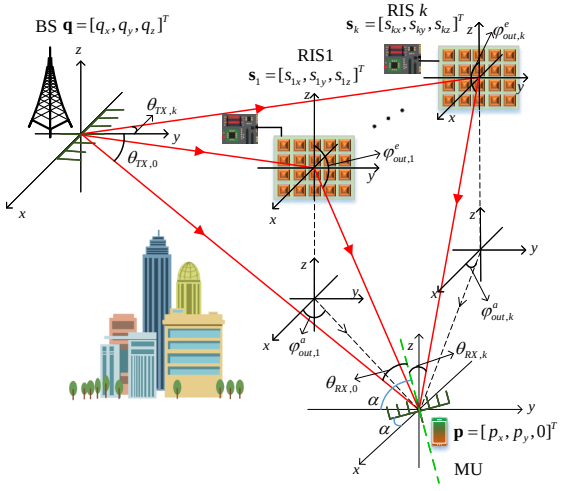

Fig. 7. Illustration of an RIS-aided localization system.

carrier frequency is f_c with a corresponding wavelength equal to λ and bandwidth equal to B. A 3D Cartesian coordinate system is used to describe the locations of the anchor nodes. The ULA of the BS is placed parallel to the x axis without rotation, and its center is located at $\mathbf{q}=[q_x,q_y,q_z]^{\mathrm{T}}\in\mathbb{R}^{3\times1}$. The UPA of the k-th RIS is placed parallel to the yz-plane, and its center is located at $\mathbf{s}_k=[s_{kx},s_{ky},s_{kz}]^{\mathrm{T}}\in\mathbb{R}^{3\times1}$. The ULA of the MU is placed on the xy-plane, and is rotated by $\alpha\in[0,\pi)$ radians relative to the x-axis. The center of the ULA of the MU is located at $\mathbf{p}=[p_x,p_y,0]^{\mathrm{T}}\in\mathbb{R}^{3\times1}$. The symbols q_z and s_{kz} denote the heights of the BS and the k-th RIS with respect to the MU on the ground. The locations of the BS and the RIS are assumed to be known. Then, the objective of the positioning system is to estimate the location (p_x,p_y) of the MU and its rotation angle α. To simplify the notation and to facilitate the analysis, we collect the parameters to be estimated in a vector $\tilde{\boldsymbol{\eta}}=[p_x,p_y,\alpha]^{\mathrm{T}}$.

A narrow-band channel model is considered. The overall channel from the MU to the BS is expressed as the combination of the direct BS-MU channel and the channels reflected by all the RISs. In mathematical terms, we have

$$\mathbf{H}=\mathbf{H}_{\mathrm{BM}}+\sum_{k=1}^{K-1}\mathbf{H}_{\mathrm{RIS},k}, \tag{111}$$

where $\mathbf{H}_{\mathrm{BM}}\in\mathbb{C}^{N_r\times N_t}$ is the BS-MU channel and $\mathbf{H}_{\mathrm{RIS},k}\in\mathbb{C}^{N_r\times N_t}$ is the k-th RIS's reflected channel. For the sake of analysis, we only consider the LoS paths in the channels of $\mathbf{H}_{\mathrm{BM}}$ and $\mathbf{H}_{\mathrm{RIS},k}$, which can be further expressed as

$$\mathbf{H}_{\mathrm{BM}}=h_0\mathbf{a}_{\mathrm{RX}}(\theta_{\mathrm{RX},0})\mathbf{a}^{\mathrm{H}}_{\mathrm{TX}}(\theta_{\mathrm{TX},0}), \tag{112a}$$

$$\begin{aligned}\mathbf{H}_{\mathrm{RIS},k}=&h_k\mathbf{a}_{\mathrm{RX}}(\theta_{\mathrm{RX},k})\mathbf{a}^{\mathrm{H}}_{\mathrm{RIS,OUT}}(\varphi^a_{\mathrm{out},k},\varphi^e_{\mathrm{out},k})\boldsymbol{\Theta}_k\\&\cdot\mathbf{a}_{\mathrm{RIS,IN}}(\varphi^a_{\mathrm{in},k},\varphi^e_{\mathrm{in},k})\mathbf{a}^{\mathrm{H}}_{\mathrm{TX}}(\theta_{\mathrm{TX},k}),\end{aligned} \tag{112b}$$

where h_0 and $\{h_1,\cdots,h_{K-1}\}$ are the complex channel gains of the BS-MU direct link and the BS-RIS-MU reflected links, respectively. $\boldsymbol{\Theta}_k$ is the reflection coefficient matrix of the k-th RIS. $\mathbf{a}_{\mathrm{TX}}(\theta_{\mathrm{TX},0})\in\mathbb{C}^{N_t\times1}$ and $\mathbf{a}_{\mathrm{RX}}(\theta_{\mathrm{RX},0})\in\mathbb{C}^{N_r\times1}$ are the antenna response vectors of the transmitter and the receiver, in which $\theta_{\mathrm{TX},0}$ denotes the AoD and $\theta_{\mathrm{RX},0}$ denotes the AoA of the BS-MU direct link. Since the transmitter and the receiver use ULAs, $\mathbf{a}_{\mathrm{TX}}(\theta)$ and $\mathbf{a}_{\mathrm{RX}}(\theta)$ can be expressed as

$$\mathbf{a}_{\mathrm{TX}}(\theta)=[1,e^{j\frac{2\pi}{\lambda}d\sin(\theta)},\cdots,e^{j(N_t-1)\frac{2\pi}{\lambda}d\sin(\theta)}]^{\mathrm{T}}, \tag{113a}$$

$$\mathbf{a}_{\mathrm{RX}}(\theta)=[1,e^{j\frac{2\pi}{\lambda}d\sin(\theta)},\cdots,e^{j(N_r-1)\frac{2\pi}{\lambda}d\sin(\theta)}]^{\mathrm{T}}, \tag{113b}$$

where d denotes the distance between adjacent antennas and θ is the AoA or AoD. The $\mathbf{a}_X(\varphi^a_{\mathrm{x},k},\varphi^e_{\mathrm{x},k})$ in (112b) denotes the array response vector of the RIS array, which is equal to $\mathbf{a}_{\mathrm{RIS,IN}}(\varphi^a_{\mathrm{in},k},\varphi^e_{\mathrm{in},k})$ and $\mathbf{a}_{\mathrm{RIS,OUT}}(\varphi^a_{\mathrm{out},k},\varphi^e_{\mathrm{out},k})$ when x is equal to IN and OUT, respectively. Since the RISs use UPAs, $\mathbf{a}_X(\varphi^a_{\mathrm{x},k},\varphi^e_{\mathrm{x},k})$ can be expressed as

$$\begin{aligned}\mathbf{a}_X(\varphi^a_{\mathrm{x},k},\varphi^e_{\mathrm{x},k})=&[1,e^{j\frac{2\pi}{\lambda}d\cos(\varphi^e_{\mathrm{x},k})},\cdots,e^{j\frac{2\pi}{\lambda}(L-1)d\cos(\varphi^e_{\mathrm{x},k})}]^{\mathrm{T}}\\&\otimes[1,e^{j\frac{2\pi}{\lambda}d\sin(\varphi^e_{\mathrm{x},k})\sin(\varphi^a_{\mathrm{x},k})}\cdots,e^{j\frac{2\pi}{\lambda}(L-1)d\sin(\varphi^e_{\mathrm{x},k})\sin(\varphi^a_{\mathrm{x},k})}]^{\mathrm{T}},\end{aligned}$$

where $\varphi^a_{\mathrm{in},k}$ and $\varphi^e_{\mathrm{in},k}$ are the azimuth AoA and elevation AoA at the k-th RIS from the BS, while $\varphi^a_{\mathrm{out},k}$ and $\varphi^e_{\mathrm{out},k}$ are the azimuth AoD and elevation AoD at the k-th RIS to the MU.

The complex channel gains $h_k,k=0,1,\cdots,K-1$ in (112) follow the distribution of $\mathcal{CN}(0,10^{\frac{-\mathrm{PL}_k}{10}})$, where $\mathrm{PL}_0,\mathrm{PL}_1,\cdots,\mathrm{PL}_{K-1}$ are the corresponding path losses. Based on [21] and [230], the path loss of the BS-MU channel in (112a) can be expressed in dB as

$$\begin{aligned}\mathrm{PL}_0=&10\log_{10}(64\pi^3)+10\alpha_0\log_{10}d_{BM}\\&+20\log_{10}f_c+\xi_0,\end{aligned} \tag{114}$$

where d_{BM} is the distance (in meters) between the BS and the MU, α_0 is the path loss exponent. $\xi_0\sim\mathcal{N}(0,\sigma^2_{SF_0})$ is the log-normal term accounting for the shadow fading, where $\sigma^2_{SF_0}$ denotes the log-normal shadowing variance.

The path loss of the RIS-reflected channel via the k-th RIS can be expressed in dB as

$$\begin{aligned}\mathrm{PL}_k=&10\log_{10}(64\pi^3)+10\alpha_k\log_{10}(d_{\mathrm{BR}}\cdot d_{\mathrm{RM}})\\&+40\log_{10}f_c+\xi_k,\end{aligned} \tag{115}$$

where d_{BR} and d_{RM} are the distances (in meters) of the BS-RIS k link and RIS k-MU link, respectively. α_k is the path loss exponent of the reflected channel, and ξ_k is the log-normal term accounting for the shadow fading following distribution $\mathcal{N}(0,\sigma^2_{SF_k})$ with variance $\sigma^2_{SF_k}$.

For estimating the locations of users, we assume a pilot-based transmission scheme. The transmitted pilot signal is a continuous time-domain waveform $x(t)$ with bandwidth B and duration of T_o seconds. Assuming M_t pilot signals, i.e., $\mathbf{x}(t)=[x_1(t),x_2(t),\cdots,x_{M_t}(t)]^{\mathrm{T}}\in\mathbb{C}^{M_t\times1}$, is transmitted at time t with corresponding beamforming matrix denoted by $\mathbf{F}=[\mathbf{f}_1,\mathbf{f}_2,\cdots,\mathbf{f}_{M_t}]\in\mathbb{C}^{N_t\times M_t}$ where $M_t\ll N_t$, then the signal $\mathbf{y}(t)\in\mathbb{C}^{N_r\times1}$ received at the MU is given by

$$\mathbf{y}(t)=\mathbf{H}_{\mathrm{BM}}\mathbf{F}\mathbf{x}(t-\tau_0)+\sum_{k=1}^{K-1}\mathbf{H}_{\mathrm{RIS},k}\mathbf{F}\mathbf{x}(t-\tau_k)+\mathbf{n}(t),$$

where τ_0 is the propagation delay of the direct BS-MU path, and $\tau_k,k=1,2,\cdots,K-1$ is the propagation delay from the BS to the MU via the k-th RIS. The delays $\tau_k,k=0,1,2,\cdots,K-1$ is also called TOAs. The received noise $\mathbf{n}(t)$ is Gaussian with zero mean and two-sided power spectral density equal to $N_0/2$.

2) Two-Step Localization Scheme: The position and orientation of the MU can be estimated based on the observed signal $\mathbf{y}(t)$ by using a two-step localization scheme. In the first step, the angles, channel gains, and time delays are estimated using some existing methods. For example, the channel gains $\{\hat{h}_k, \hat{h}_0\}$ can be estimated with the LS approach, while the angles $\{\hat{\theta}_{TX,0}, \hat{\varphi}^a_{out,k}, \hat{\varphi}^e_{out,k}, \hat{\theta}_{RX,0}, \hat{\theta}_{RX,k}\}$ can be estimated by using array signal processing algorithms in the spatial domain, such as MUSIC and ESPRIT. The estimated time delays $\{\hat{\tau}_k, \hat{\tau}_0\}$ can be extracted from the pilot signal. In the second step, the position of the MU can be obtained by using multi-angulation (or triangulation) from the angle-related measurements, or by using multi-lateration (or trilateration) methods from distance-related measurements, or by a combination of both.

As shown in Fig. 7, the localization measurements are closely related with the coordinates and rotation angle of MU, which means that there exists a mapping from the MU's location to the measurements obtained in the first step, which can be described as follows.

ToA: The propagation delay of the BS-MU path, which is related with the MU coordinate $\mathbf{p}$, is given by

$$\hat{\tau}_0 = \|\mathbf{q} - \mathbf{p}\| / c + n_{t_0}, \tag{116}$$

where c is the speed of light and n_{t_0} denotes the ToA error. The ToA $\hat{\tau}_{k0}$ from the MU to the k-th RIS can be obtained by subtracting the propagation delay of the k-th RIS-BS path from the propagation delay of the path from BS to MU via the k-th RIS, which are expressed as

$$\hat{\tau}_{k0} = \|\mathbf{p} - \mathbf{s}_k\| / c + n_{t_k} = \hat{\tau}_k - \|\mathbf{q} - \mathbf{s}_k\| / c + n_{t_k}, \tag{117}$$

where n_{t_k} denotes the ToA error.

The ToA measurements are related with the path lengths, which are further related with the MU position. Specifically, the $\hat{\tau}_0$ and $\hat{\tau}_{k0}$ can describe the estimated distance of the direct path as well as the estimated distance from the MU to the k-th RIS, which are expressed as follows

$$\hat{d}_0 = c\hat{\tau}_0 = \|\mathbf{q} - \mathbf{p}\| + n_{d_0}, \tag{118a}$$

$$\hat{d}_k = c\hat{\tau}_{k0} = \|\mathbf{p} - \mathbf{s}_k\| + n_{d_k}, \tag{118b}$$

where n_{d_0} and n_{d_k} are the estimation errors.

TDoA: The location of the MU can be derived from the TDoA, which describes the distance differences between different pairs of the direct MU-BS path and the k MU-RIS paths.

Specifically, by taking the direct path from the BS to MU as the reference path, the TDoA can be expressed as

$$\hat{\tau}_{td,k} = \hat{\tau}_{k0} - \hat{\tau}_0. \tag{119}$$

Then, the distance difference corresponding to the TDoA is represented as

$$\begin{aligned} c\hat{\tau}_{td,k} &= c\hat{\tau}_{k0} - c\hat{\tau}_0 = \hat{d}_{d,k} = (\hat{d}_k - \hat{d}_0) \\ &= \|\mathbf{p} - \mathbf{s}_k\| - \|\mathbf{q} - \mathbf{p}\| + n_{td_k}, \end{aligned} \tag{120}$$

where n_{td_k} is the estimation error of $d_{d,k}$.

AoA and AoD: The angle-related measurements (AoA and AoD) are estimated at the arrays of BS, MU, and RIS using angle estimation algorithms. They are also closely related with the MU position, which can be described as

$$\hat{\theta}_{TX,0} = \arcsin\left(\frac{p_x - q_x}{\|\mathbf{p} - \mathbf{q}\|_2}\right) + n_{\theta_{TX,0}}, \tag{121}$$

$$\hat{\varphi}^a_{out,k} = \arcsin\left(\frac{p_y - s_{ky}}{\sqrt{(p_x - s_{kx})^2 + (p_y - s_{ky})^2}}\right) + n_{a_k}, \tag{122}$$

$$\hat{\varphi}^e_{out,k} = \arccos\left(\frac{-s_{kz}}{\|\mathbf{p} - \mathbf{s}_k\|_2}\right) + n_{e_k}, \tag{123}$$

$$\begin{aligned} \hat{\theta}_{RX,0} = \arcsin&\left(\frac{(p_x - q_x)\cos\alpha - (p_y - q_y)\sin\alpha}{\|\mathbf{p} - \mathbf{q}\|_2}\right) \\ &+ n_{\theta_{RX,0}}, \end{aligned} \tag{124}$$

$$\begin{aligned} \hat{\theta}_{RX,k} = \arcsin&\left(\frac{(p_x - s_{kx})\cos\alpha - (p_y - s_{ky})\sin\alpha}{\|\mathbf{p} - \mathbf{s}_k\|_2}\right) \\ &+ n_{\theta_{RX,k}}. \end{aligned} \tag{125}$$

where $n_{\theta_{TX,0}}$,n_{a_k},n_{e_k}, $n_{\theta_{RX,0}}$, and $n_{\theta_{RX,k}}$ denote error associated to the AoA and AoD estimations. Besides, $\theta_{TX,k}$ are known values, which can be calculated in advance from the known coordinates of the BS and RISs.

Channel Gains: Some recent works, e.g., [218], [220], extract the distance information from the channel gains h_k according to the relations $h_0 = \frac{\lambda}{4\pi}\frac{1}{d_{BM}}$ and $h_k = \frac{\lambda}{4\pi}\frac{1}{d_{BR}d_{RM}}$. However, the channel gains depend on the distance-dependent large-scale path loss as well as the distance-independent shadowing fading. Thus, it may not be accurate to extract the distance information from the channel gains. Therefore, this measurement is not considered in the following numerical results.

In the second step, by collecting the equations mapping the MU position into the measurements illustrated above, a set of interlinked equations can be formulated. The position and rotation angle of the MU, i.e., the vector $\tilde{\boldsymbol{\eta}} = [p_x, p_y, \alpha]^{\mathrm{T}}$ can be retrieved by jointly solving the set of equations. The larger the number of measurements (equations) is available, the better the estimate of the position is. The obtained equations are, however, nonlinear, and the measurements are affected by errors. Consequently, they are difficult to solve. To circumvent this difficulty, wireless localization algorithms, such as the Chan algorithm [231] and the Taylor series algorithm [232] can be utilized.

3) Position Error Bound (PEB) and Rotation Error Bound (REB) Analysis: To evaluate the performance of the position and orientation estimation, we analyze the FIM and CRB. To provide a benchmark for the accuracy of the position estimation, the CRB and the PEB/REB are investigated in [218], [225], by assuming a 2D localization problem formulation. In the following, we derive the CRB and PEB/REB in a general 3D space. Further details can be found in [233].

First, we construct the FIM based on the channel parameters, including the ToA, AoA, AoD, and the complex channel

gains. The unknown parameters are collected in the following vector

$$\boldsymbol{\eta} = [\boldsymbol{\tau}, \theta_{\mathrm{TX},0}, \boldsymbol{\theta}_{\mathrm{RX}}, \boldsymbol{\varphi}^a_{\mathrm{out}}, \boldsymbol{\varphi}^e_{\mathrm{out}}, \mathbf{h}_R, \mathbf{h}_I]^{\mathrm{T}}, \tag{126}$$

where

$$\boldsymbol{\tau} = [\tau_0, \tau_1, \cdots, \tau_{K-1}]^{\mathrm{T}}, \tag{127a}$$
$$\boldsymbol{\theta}_{\mathrm{RX}} = [\theta_{\mathrm{RX},0}, \theta_{\mathrm{RX},1}, \cdots, \theta_{\mathrm{RX},K-1}]^{\mathrm{T}}, \tag{127b}$$
$$\boldsymbol{\varphi}^a_{\mathrm{out}} = [\varphi^a_{\mathrm{out},1}, \varphi^a_{\mathrm{out},2}, \cdots, \varphi^a_{\mathrm{out},K-1}]^{\mathrm{T}}, \tag{127c}$$
$$\boldsymbol{\varphi}^e_{\mathrm{out}} = [\varphi^e_{\mathrm{out},1}, \varphi^e_{\mathrm{out},2}, \cdots, \varphi^e_{\mathrm{out},K-1}]^{\mathrm{T}}, \tag{127d}$$
$$\mathbf{h}_R = [h_{R,0}, h_{R,1}, \cdots, h_{R,K-1}]^{\mathrm{T}}, \tag{127e}$$
$$\mathbf{h}_I = [h_{I,0}, h_{I,1}, \cdots, h_{I,K-1}]^{\mathrm{T}}, \tag{127f}$$

and $\theta_{\mathrm{TX},0}$ denotes the AOD from the BS to the MU. The real parts of the channel gains $h_{R,k} = \mathrm{Re}\{h_k\}$ are collected in the vector $\mathbf{h}_R$ and the imaginary parts $h_{I,k} = \mathrm{Im}(h_k)$ are collected in the vector $\mathbf{h}_I$. We denote the unbiased estimator of the unknown channel parameters as $\hat{\boldsymbol{\eta}}$. Then, the mean-square error bound is

$$\mathbb{E}[(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{H}}] \geq \mathbf{J}^{-1}_{\boldsymbol{\eta}}, \tag{128}$$

where $\mathbf{J}_{\boldsymbol{\eta}}$ is the $(6K-1) \times (6K-1)$ dimensional FIM. Assuming that the noise is a wide-sense stationary (WSS) Gaussian random vector and is written as $\boldsymbol{n}(t)$, the element in the m-th row and n-th column of matrix $\mathbf{J}_{\boldsymbol{\eta}}$ can be written as [234]

$$[\mathbf{J}_{\boldsymbol{\eta}}]_{mn} = \mathbb{E}_{\mathbf{y}(t)|\boldsymbol{\eta}} \left[-\frac{\partial^2 \ln p(\mathbf{y}(t)\,|\boldsymbol{\eta})}{\partial \eta_m \partial \eta_n} \right] \approx \frac{2}{N_0} \int_0^{T_0} \mathrm{Re}\left\{ \frac{\partial \boldsymbol{\mu}^{\mathrm{H}}(t)}{\partial \eta_m} \frac{\partial \boldsymbol{\mu}(t)}{\partial \eta_n} \right\}. \tag{129}$$

where η_m denotes the m-th element of $\boldsymbol{\eta}$, T_0 denotes the observation time, $\boldsymbol{\mu}(t) = \mathbf{y}(t) - \boldsymbol{n}(t)$, and $p(\mathbf{y}(t)\,|\boldsymbol{\eta})$ denotes the probability density function (PDF) (i.e, the likelihood function) of the random vector $\mathbf{y}(t)$ conditioned on the parameter vector $\boldsymbol{\eta}$.

The FIM of the position and orientation of the MU, i.e., $\tilde{\eta}$, can be obtained by applying a transformation of variables from the channel parameters $\boldsymbol{\eta}$ to the location parameters $\tilde{\boldsymbol{\eta}}$. Specifically, the FIM of $\tilde{\boldsymbol{\eta}}$ is obtained by applying the transformation matrix $\mathbf{T}$ defined as

$$[\mathbf{T}]_{1:3,m} = [\frac{\partial \eta_m}{\partial p_x}, \frac{\partial \eta_m}{\partial p_y}, \frac{\partial \eta_m}{\partial \alpha}]^{\mathrm{T}}. \tag{130}$$

The equation for the FIM of the position and orientation parameters in $\tilde{\boldsymbol{\eta}}$ is given by [234]

$$\mathbf{J} = \mathbf{T}\mathbf{J}_{\boldsymbol{\eta}}\mathbf{T}^{\mathrm{H}}. \tag{131}$$

Once the FIM is obtained, the CRB is obtained by inverting the matrix. Thus, the PEB is equal to the square root of the trace of the first 2×2 sub-matrix and can be formulated as

$$\mathrm{PEB} = \sqrt{\mathrm{tr}\left(\mathbf{J}^{-1}_{1:2,1:2}\right)}, \tag{132}$$

and the REB is equal to the square root of the third entry of diagonal

$$\mathrm{REB} = \sqrt{\mathrm{tr}\left(\mathbf{J}^{-1}_{3,3}\right)}. \tag{133}$$

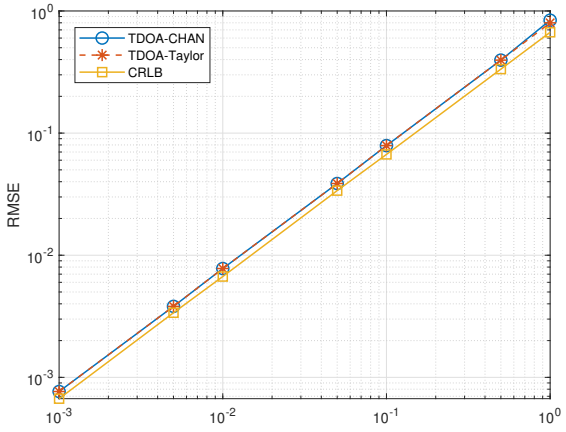

Fig. 8. RMSE versus the standard variance σ_{td}, when $N_t = 32$, $N_r = 8$, $L^2 = 16 \times 16$, $f_c = 9$ GHz, and $K = 4$.

Simulation results: To illustrate the position accuracy that can be obtained in the presence of RISs, we report some numerical results based on the measurements of the TDoA in (120). Specifically, we compare the root mean square error (RMSE) of the position estimation obtained by using Taylor and Chan's algorithms against the the CRB. The locations in metre of the BS and MU are $\mathbf{q} = [0, 0, 40]^T$ and $\mathbf{p} = [90, 30, 0]^T$, respectively, and the locations of three RISs are $\mathbf{s1} = [60, 45, 15]^T$, $\mathbf{s2} = [50, 50, 5]^T$ and $\mathbf{s3} = [40, 20, 10]^T$. The setup with RISs is considered as a case study. The phase shift matrix of the RIS is set to a unit identity matrix for the sake of analysis, if the matrix is optimized, the estimation accuracy will be improved. The other simulation parameters are given in the caption of the figure, where σ_{td} denotes the variance of the Gaussian estimation error n_{td_k}. It can be seen that Taylor's and Chan's algorithms achieve almost the same performance and have negligible performance loss compared with the CRB.

B. Near-Field Localization Techniques

Thanks to the absence of power amplifiers, digital processing units, and RF chains for each scattering element, a large number of quasi passive reflecting elements can be integrated on the RIS panel at an affordable cost and power consumption. Consequently, an RIS may evolve into a panel with a very large size, which is sometimes called XL-RIS.

In this case, as illustrated in Fig. 9, the RIS may be large enough and the observation point may be sufficiently close to it that the far-field planar wave assumption may not hold anymore. Specifically, the spherical curvature of the wavefront may not be ignored. In these cases, specifically, a signal formulation that accounts for the near-field is needed. According to [235], the far-field of the RIS can be defined as the set of observation distances R that are greater than the Fraunhofer distance R_f, i.e.,

$$R \geq R_f = \frac{2L^2}{\lambda}, \tag{134}$$

where L is the maximum aperture of the RIS UPA and λ is the carrier wavelength. This conventional definition of far-field

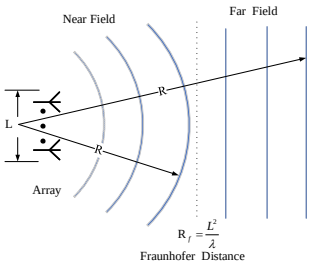

Fig. 9. Illustration of the Fraunhofer distance and the far-field planar wave approximation.

region corresponds to a maximum phase error equal to $\pi/8$ over the entire length L.

According to (134), the Fraunhofer distance can be quite large if an XL-RIS is utilized and its size is sufficiently large with respect to the wavelength. If the RIS is deployed at high-frequency bands, such as mmwave/sub-THz, the Fraunhofer distance also increases due to the smaller wavelength [236]–[238]. For instance, let us consider a 100×100-UPA RIS panel that operates at the carrier frequency of 200 GHz. The RIS elements are assumed to be spaced by one half-wavelength, i.e. $\frac{\lambda}{2} = 0.75$ mm, and, thus, the maximum aperture (the diagonal) of the RIS panel is $\sqrt{2}\times100\times\frac{\lambda}{2}\approx10.6$ cm. Then, the Fraunhofer distance for this scenario is $R_f = 7.5$ meters. Therefore, it is likely that some users are in the near-field when the RIS is deployed for indoor scenarios. In this case, the spherical wavefront of the electromagnetic waves cannot be ignored. Recent system-level simulations have shown that the near-field region may not be ignored in outdoor scenarios either [239].

1) Near-Field Channel Model: In the near-field region, spherical wave-front provide an underlying generic parametric model for estimating the positions of MUs and scatterers [240]. Under the assumption of a spherical wave-front, the steering responses of the RIS and the BS arrays need to be parameterized by taking into account the 3D location of the signal source rather than by simply utilizing the AoDs and the AoAs.

As an example, let us consider the uplink transmission where MU is equipped with one antenna, the RIS is an $M_x \times M_z$ UPA that lies on the XOZ plane, and the BS is equipped with an N-element ULA. In the near-field, the channel from the MU to the RIS is

$$\boldsymbol{h}_r = \sum\nolimits_{l=0}^{L_{\mathrm{RU}}} \tilde{\beta}_l \boldsymbol{a}_R(\boldsymbol{r}_l) = \boldsymbol{A}_{\mathrm{R}}(\boldsymbol{r}_l)\tilde{\boldsymbol{\beta}}, \tag{135}$$

where L_{RU} is the number of scatterers or spatial paths, the vector $\tilde{\boldsymbol{\beta}} = [\tilde{\beta}_0, \ldots, \tilde{\beta}_{L_{\mathrm{RU}}}]^{\mathrm{T}}$ contains the complex path gains, and the matrix $\boldsymbol{A}_{\mathrm{R}}(\boldsymbol{r}_l) = [\boldsymbol{a}_{\mathrm{R}}(\boldsymbol{r}_0), \ldots, \boldsymbol{a}_R(\boldsymbol{r}_{L_{\mathrm{RU}}})]$ collects the steering vectors associated with the scatterers that produce the multiple propagation paths. The LoS path from the MU to the RIS is denoted as path 0, and the location of the MU is denoted as $\boldsymbol{r}_0 = [x_0, y_0, z_0]^T$. The steering vector associated with the l-th scatterer located at $\boldsymbol{p}_l = [x_l, y_l, z_l]^T$ is denoted as $\boldsymbol{a}(\boldsymbol{p}_l)$, and is given by

$$\begin{aligned}\boldsymbol{a}(\boldsymbol{p}_l) = [&e^{-j\frac{2\pi}{\lambda}d_{1,1}(\boldsymbol{p}_l)}, \cdots, e^{-j\frac{2\pi}{\lambda}d_{1,M_z}(\boldsymbol{p}_l)}, \cdots, \\ &e^{-j\frac{2\pi}{\lambda}d_{M_x,1}(\boldsymbol{p}_l)}, \cdots, e^{-j\frac{2\pi}{\lambda}d_{M_x,M_z}(\boldsymbol{p}_l)}]^{\mathrm{T}},\end{aligned} \tag{136}$$

where $d_{m,n}(\boldsymbol{p}_l)$ represents the distance from the scatterer located at $\boldsymbol{p}_l$ to the (m,n)-th RIS element.

We concentrate on the LoS channel between the RIS and the AP and assume that the communication channel is slowly varying and frequency-flat. By taking the spherical nature of the wave propagation, the path length from each antenna elements of the BS to each elements of the RIS determines the phase shifts of the received signals. Similar to [241] and [242], we assume that the path loss of all the received paths is the same and is denoted by α. Under this assumption, the elements of the BS-RIS LoS channel are collected in the matrix $\boldsymbol{H}$ whose elements are

$$\boldsymbol{H}(m,n) = \alpha \exp\left(j\frac{2\pi}{\lambda}r_{m,n}\right), \tag{137}$$

where $r_{m,n}$ denotes the path length between the m-th antenna element of the BS and the n-th RIS element.

The signal transmitted by the MU is denoted by s and the received data at the BS is

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{\Theta}\boldsymbol{h}_r s + \boldsymbol{n}, \tag{138}$$

where $\boldsymbol{n} \sim \mathcal{CN}\left(\boldsymbol{0}, \sigma^2\mathbf{I}\right)$ and $\boldsymbol{\Theta}$ denote the AWGN and the reflection coefficient matrix defined in Section II, respectively.

The set of unknown parameters are

$$\boldsymbol{\eta} = [\boldsymbol{p}_0, \boldsymbol{p}_1 \cdots, \boldsymbol{p}_{L_{\mathrm{RU}}}, \boldsymbol{\beta}^T]^T, \tag{139}$$

where $\boldsymbol{\beta} = [\alpha\tilde{\beta}_0, \ldots, \alpha\tilde{\beta}_{L_{\mathrm{RU}}}]^{\mathrm{T}}$ denotes the vector of cascaded channel fading coefficients.

A commonly adopted localization method is the maximum-likelihood estimation. The log-likelihood function of the received signal $\mathbf{y}$ is given by $f(\boldsymbol{y}|\boldsymbol{\Theta},\boldsymbol{\eta}) \propto -||\boldsymbol{y} - \mathbf{H}\boldsymbol{\Theta}\mathbf{h}_r^{\mathrm{H}}(\boldsymbol{\eta})s||^2$, where irrelevant constant terms are ignored. Then, the maximum likehood estimation problem for the locations of the MU and the scatterers is given by

$$\begin{aligned}&\max_{\boldsymbol{\eta},\boldsymbol{\Theta}} \ f(\boldsymbol{y}|\boldsymbol{\Theta},\boldsymbol{\eta}) = \min_{\boldsymbol{\eta},\boldsymbol{\Theta}} \ ||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\Theta}\boldsymbol{h}_r(\boldsymbol{\eta})s||^2 \\ &\text{s.t.} \quad \mathrm{C1}: \boldsymbol{\eta} \in \mathcal{S}_\eta, \\ &\qquad \mathrm{C2}: \theta_m \in \mathcal{S}_1 \text{ or } \mathcal{S}_2, \forall m = 1,2,\ldots,M,\end{aligned} \tag{140}$$

where $\mathcal{S}_\eta$ denotes the set of feasible locations.

2) Near-Field Localization Scheme: The location of the MU can be estimated via a two-stage near-field localization scheme. As the phase shift of the RIS can be obtained as $\hat{\boldsymbol{\Theta}}$ by leveraging the methods discussed in Section III, here we only focus on the localization problem. In the first stage, the coarse estimation of the positions of MU and scatterers is derived [226], [243]. In the second stage, the finer estimation of the positions of MU and scatterers can be estimated with the iterative methods, e.g., two-dimensional search [244]. The details of the two-stage near-field localization scheme are outlined below.

Stage 1: Suppose that the transmit signal is $s = 1$. As the received power from the scatterer-reflected paths is much weaker than the LoS path, we first try to find the location of MU based on the LoS path with the largest received power. In the first stage, the location of the MU is estimated as follows:

$$\hat{\boldsymbol{p}}_0 = \arg\min_{\boldsymbol{p}_0} \ ||\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{\Theta}}\boldsymbol{a}_R(\boldsymbol{p}_0)||^2. \tag{141}$$

Without any prior information for the source location, the objective function in (141) needs to be evaluated over a set of grid locations. For the sake of reducing complexity, a non-uniform grid may be used. Specifically, the angle and distance are sampled in the polar domain using a uniform grid in [226]. In general, the grids should be dense near the array and sparse away from the array [243].

With the obtained location, the corresponding channel gain can be obtained by using the projection method as

$$\hat{\beta}_0 = \frac{(\boldsymbol{H}\hat{\boldsymbol{\Theta}}\boldsymbol{a}_R(\hat{\boldsymbol{p}}_0))^H \boldsymbol{y}}{||\boldsymbol{H}\hat{\boldsymbol{\Theta}}\boldsymbol{a}_R(\hat{\boldsymbol{p}}_0)||^2}. \tag{142}$$

Then, the locations of the scatterers are estimated as

$$\hat{\boldsymbol{p}}_l = \arg\min_{\boldsymbol{p}_l} \left\| \boldsymbol{y}_{-l} - \boldsymbol{H}\hat{\boldsymbol{\Theta}}\boldsymbol{a}_R(\boldsymbol{p}_l) \right\|^2, \tag{143}$$

where $\boldsymbol{y}_{-l} = \boldsymbol{y} - \sum_{s=1}^{l-1} \boldsymbol{H}\hat{\boldsymbol{\Theta}}\hat{\beta}_s \boldsymbol{a}_R(\hat{\boldsymbol{p}}_s)$, and the corresponding channel gain is obtained as

$$\hat{\beta}_l = \frac{(\boldsymbol{H}\hat{\boldsymbol{\Theta}}\boldsymbol{a}_R(\hat{\boldsymbol{p}}_l))^H \boldsymbol{y}_{-l}}{||\boldsymbol{H}\hat{\boldsymbol{\Theta}}\boldsymbol{a}_R(\hat{\boldsymbol{p}}_l)||^2}. \tag{144}$$

Stage 2: With the coarse stimation, a refined estimation can be further conducted in the second stage. The position of the MU and the scatterers can be refined using a finer grid while keeping the estimation of other locations fixed. Specifically,

$$\min_{\boldsymbol{p}_l} \; ||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\Theta}\boldsymbol{h}_r(\boldsymbol{p}_l, \boldsymbol{\eta}_{-l})||^2, \tag{145}$$

where $\boldsymbol{\eta}_{-l} = [\hat{\boldsymbol{r}}_0, \cdots, \hat{\boldsymbol{r}}_{l-1}, \hat{\boldsymbol{r}}_{l+1}, \cdots, \hat{\boldsymbol{r}}_{L_{\mathrm{RU}}}, \hat{\boldsymbol{\beta}}^T]^T$. Several iterations are needed for high-resolution location estimates.

Simulation results: For the near field localization case, we evaluate the localization performance in terms of the MSE, which is obtained by the proposed two-stage near-field localization scheme. The locations of BS and RIS are $\mathbf{q} = [3, 0, 3]^T$ and $\mathbf{p} = [0, 0.5, 2]^T$, while the locations of the MU and scatterers are randomly generated in the 5 m $\times$ 5 m cell range. The number of RIS elements is 21×21, and the phase shift matrix is obtained by maximizing the received signal power. It can be seen that the MSE performance degrades with the number of scatterers due to the more scattered energy and the feedback estimation error propagation, while increasing the SNR can help improve the localization accuracy.

V. Future Directions

In this section, we summarize some open research issues that are focused on the modeling, the analysis, and the optimization of RIS-aided wireless systems.

A. Mobility

Most existing works focus on quasi-static scenarios where the users and the RISs are assumed to be nearly stationary. However, it is also imperative to investigate scenarios with mobility [70], [245]–[248], such as wireless systems in the presence of vehicles, trains, and UAVs. In these scenarios, in particular, the RISs may even be mobile as they may be carried by UAVs or they may be deployed in trains. In these scenarios, several challenging research problems emerge. For example, the estimation of rapidly time-varying channels becomes more challenging. Therefore, new channel estimation methods are required to quickly and accurately track the channels, in order to enable the design of accurate phase shifts at the RIS. Also, since the channel is rapidly changing, the channel estimation and the RIS phase shift configuration need to be executed more frequently. As a result, it is crucial to design effective signal processing schemes to avoid a prohibitively high overhead in high-mobility scenarios. Furthermore, the RIS has the potential to reduce the Doppler and delay spread [247], which deserves further research in more practical scenarios.

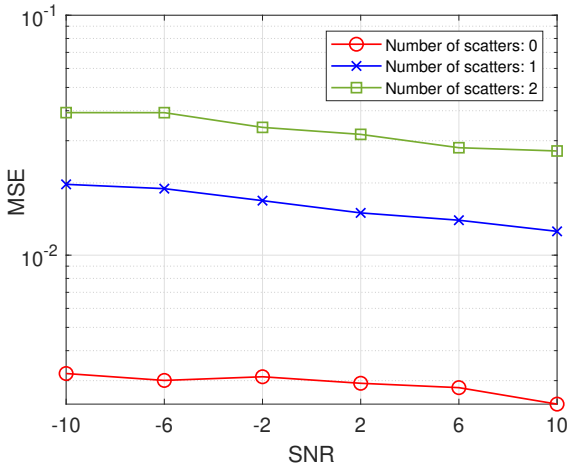

Fig. 10. MSE versus the SNR, when $N_t = 32$ and $f_c = 240$ GHz.

B. Near-Field Channel

To ensure a sufficiently large coverage, the aperture of the RIS planar array is typically large in order to compensate for the absence of power amplifiers and digital signal processing units for regenerating the signals. Depending on the size of the RIS, the operating frequency, and its distance from the transmitter and the receiver, it may be necessary to utilize a near-field communication model [22], [249]–[251]. Near-field communication models are, however, less understood than their far-field counterparts. Therefore, it is essential to consider accurate channel models to unveil the fundamental performance limits and the scaling laws of RISs in this context. It is well known, for example, that the scaling laws of the electromagnetic field radiated by the RIS is different in the near-field and far-field regions [252]. Also, new low-overhead channel estimation strategies are needed since the channel sparsity in the angle domain, which is based on the far-field planar wavefront assumption, may no longer hold when using the near-field spherical wavefront assumption [226].

C. Active RIS

Quasi Passive RIS architectures have attracted significant attention due to their low hardware cost and energy consumption. Nevertheless, quasi passive architectures have their limitations. Since the signals reflected by the RIS is determined by the product of the distances from the transmitter to the RIS and from the RIS to the receiver in the far-field region, the

received signal strength is relatively small compared with that from direct links, especially when the direct links are strong. The recently proposed active RIS architecture is a promising solution to overcome this issue [253]–[256]. After integrating active reflection-type amplifiers [253], in addition to adjusting the phase shifts, active RISs can simultaneously amplify the magnitude of the reflected signals. However, the amplification at the active RIS requires an additional power supply. Therefore, more dedicated beamforming at the BS and phase shifts at the RIS should be designed to balance the performance and energy consumption. Meanwhile, channel estimation with active RISs needs to incorporate the statistical properties of amplification-induced noise, which is more challenging and worth further investigation. Besides, how to decide the optimal deployment of the active RISs is still an open question.

D. Double/multi-RIS

Most of the existing works have considered quasi passive beamforming designs and channel estimation schemes in systems with a single RIS. In some scenarios, however, it may be convenient to enable the transmission of signals through reflections from multiple RISs in order to route the signals and bypass the blocking objects in a smart manner, directly at the electromagnetic level (electromagnetic routing) [5]. Thus, double/multiple RISs may be utilized to realize a blockage-free communication network via multiple signal reflections. Moreover, the proper design of the cooperative quasi passive beamforming can eliminate the inter-RIS interference, and achieve multiplicative beamforming gain from the inter-RIS reflection channel [257]–[260]. However, these methods require highly accurate CSI, which is challenging to obtain due to the coupling of the different reflected signal links and more channel coefficients to be estimated [261]–[263]. In this context, channel estimation methods in double/multi-RIS aided communication scenarios in the presence of single/double/multi-reflection links need to be investigated. In addition, existing research works on double/multi-RIS aided systems often ignore the impact of the secondary reflections among the RISs, which may be a reasonable approximation if the RISs are in the far-field of each other but it may not hold anymore if the RISs are closely located. Finally, the analysis and design of multi-RIS communications at high frequency bands is an open research issue as well.

E. Multifunction RISs

In the existing literature, most of the research works concentrate on RISs that operate as anomalous reflectors or as reflecting lenses. An RIS, however, can realize multiple signal transformations depending on how the scattering matrix (or equivalently the surface impedance) is designed [5]. Recently, notably, a few research attempts have been made to design RISs that operate as anomalous refracting mirrors or as anomalous refracting lenses [264] as well RISs that can simultaneously realize reflections and refractions in order to guarantee omni-coverage performance [265]–[268]. Multifunction RISs are an emerging research topic, and the corresponding modeling, performance evaluation, and optimization are still at its infancy.

VI. Conclusions

In this paper, we provided a comprehensive overview of state-of-the-art research for new and revolutionary RIS/IRS-aided wireless systems, with an emphasis on signal processing techniques for solving various channel estimation, transmission design and radio localization problems. Specifically, we first reviewed existing results on channel estimation under unstructured and structured channel models. Next, we provided a detailed overview of the research results from the perspective of different optimization techniques and availability of CSI. In particular, several optimization techniques were described for optimizing RIS with discrete and continuous time shifts. As far the availability of CSI is concerned, three main cases were considered, namely, fully instantaneous CSI, two-timescale CSI, and fully long-term CSI. Simulation results demonstrated that the two-timescale CSI scheme constitutes a promising approach when the pilot overhead is taken into account. Furthermore, radio localization is an important application of RISs, and it has been thoroughly reviewed by considering deployment scenarios and channel models that account for far-field and near-field propagation. Finally, several open research problems have been discussed.

References

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[2] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, Nov. 2020.

[3] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sept. 2019.

[4] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[5] M. Di Renzo, M. Debbah, D.-T. Phan-Huy, A. Zappone, M.-S. Alouini, C. Yuen, V. Sciancalepore, G. C. Alexandropoulos, J. Hoydis, H. Gacanin *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wirel. Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, May 2019.

[6] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[7] C. Pan, H. Ren, K. Wang, J. F. Kolb, M. Elkashlan, M. Chen, M. Di Renzo, Y. Hao, J. Wang, A. L. Swindlehurst, X. You, and L. Hanzo, "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, Jun. 2021.

[8] X. Yuan, Y.-J. A. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 136–143, Apr. 2021.

[9] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. Di Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[10] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light Sci. Appl.*, vol. 3, no. 10, pp. e218–e218, Oct. 2014.

[11] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[12] Q. Tao, J. Wang, and C. Zhong, "Performance analysis of intelligent reflecting surface aided communication systems," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2464–2468, Nov. 2020.

[13] L. Yang, J. Yang, W. Xie, M. O. Hasna, T. Tsiftsis, and M. Di Renzo, "Secrecy performance analysis of RIS-aided wireless communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12 296–12 300, Oct. 2020.

[14] Z. Ding and H. Vincent Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, 2020.

[15] H. Ren, K. Wang, and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *early access in IEEE Trans. Commun.*, 2021.

[16] L. Yang, F. Meng, J. Zhang, M. O. Hasna, and M. Di Renzo, "On the performance of RIS-assisted dual-hop UAV communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10 385–10 390, Sept. 2020.

[17] J. Zhang, H. Du, Q. Sun, B. Ai, and D. W. K. Ng, "Physical layer security enhancement with reconfigurable intelligent surface-aided networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 3480–3495, May 2021.

[18] P. Xu, G. Chen, G. Pan, and M. Di Renzo, "Ergodic secrecy rate of ris-assisted communication systems in the presence of discrete phase shifts and multiple eavesdroppers," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 629–633, Mar. 2021.

[19] D. Li, "Ergodic capacity of intelligent reflecting surface-assisted communication systems with phase errors," *IEEE Commun. Lett.*, vol. 24, no. 8, pp. 1646–1650, Aug. 2020.

[20] S. Zhou, W. Xu, K. Wang, M. Di Renzo, and M.-S. Alouini, "Spectral and energy efficiency of IRS-assisted MISO communication with hardware impairments," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1366–1369, 2020.

[21] W. Tang, M. Z. Chen, X. Chen, J. Y. Dai, Y. Han, M. Di Renzo, Y. Zeng, S. Jin, Q. Cheng, and T. J. Cui, "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 421–439, Jan. 2021.

[22] E. Björnson and L. Sanguinetti, "Power scaling laws and near-field behaviors of massive MIMO and intelligent reflecting surfaces," *IEEE Open Journal of the Commun. Society*, vol. 1, pp. 1306–1324, Sept. 2020.

[23] M. Najafi, V. Jamali, R. Schober, and H. V. Poor, "Physics-based modeling and scalable optimization of large intelligent reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2673–2691, Apr. 2021.

[24] I. Yildirim, A. Uyrus, and E. Basar, "Modeling and analysis of reconfigurable intelligent surfaces for indoor and outdoor applications in future wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1290–1301, Feb. 2021.

[25] Y. Sun, C.-X. Wang, J. Huang, and J. Wang, "A 3D non-stationary channel model for 6G wireless systems employing intelligent reflecting surfaces with practical phase shifts," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 2, pp. 496–510, Jun. 2021.

[26] L. Dai, B. Wang, M. Wang, X. Yang, J. Tan, S. Bi, S. Xu, F. Yang, Z. Chen, M. Di Renzo, C.-B. Chae, and L. Hanzo, "Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45 913–45 923, Mar. 2020.

[27] W. Tang, J. Y. Dai, M. Z. Chen, K.-K. Wong, X. Li, X. Zhao, S. Jin, Q. Cheng, and T. J. Cui, "MIMO transmission through reconfigurable intelligent surface: System design, analysis, and implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2683–2699, Nov. 2020.

[28] X. Pei, H. Yin, L. Tan, L. Cao, Z. Li, K. Wang, K. Zhang, and E. Björnson, "RIS-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," *early access in IEEE Trans. Commun.*, pp. 1–1, 2021.

[29] L. Zhang, M. Z. Chen, W. Tang, J. Y. Dai, L. Miao, X. Y. Zhou, S. Jin, Q. Cheng, and T. J. Cui, "A wireless communication scheme based on space-and frequency-division multiplexing using digital metasurfaces," *Nature Electronics*, vol. 4, no. 3, pp. 218–227, Mar. 2021.

[30] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[31] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, thirdquarter 2021.

[32] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, Fourthquarter 2020.

[33] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[34] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," 2021. [Online]. Available: https://arxiv.org/abs/2110.01292

[35] A. Lee Swindlehurst, G. Zhou, R. Liu, C. Pan, and M. Li, "Channel estimation with reconfigurable intelligent surfaces – a general framework," 2021. [Online]. Available: https://arxiv.org/abs/2110.00553

[36] M.-M. Zhao, Q. Wu, M.-J. Zhao, and R. Zhang, "Exploiting amplitude control in intelligent reflecting surface aided wireless communication with imperfect CSI," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4216–4231, Jun. 2021.

[37] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[38] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2621–2636, Nov. 2020.

[39] G. de Araújo, A. de Almeida, and R. Boyer, "Channel estimation for intelligent reflecting surface assisted MIMO systems: A tensor modeling approach," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 3, pp. 789–802, Apr. 2021.

[40] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[41] C. Hu, L. Dai, S. Han, and X. Wang, "Two-timescale channel estimation for reconfigurable intelligent surface aided wireless communications," *early access in IEEE Trans. Commun.*, pp. 1–1, 2021.

[42] X. Hu, R. Zhang, and C. Zhong, "Semi-passive elements assisted channel estimation for intelligent reflecting surface-aided communications," *IEEE Trans. Wireless Commun.*, early access, 2021.

[43] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 4659–4663.

[44] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4522–4535, Jul. 2020.

[45] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Jul. 2020.

[46] T. L. Jensen and E. De Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 5000–5004.

[47] Z. Zhou, N. Ge, Z. Wang, and L. Hanzo, "Joint transmit precoding and reconfigurable intelligent surface phase adjustment: A decomposition-aided channel estimation approach," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1228–1243, Feb. 2021.

[48] B. Zheng, C. You, and R. Zhang, "Fast channel estimation for IRS-assisted OFDM," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 580–584, Mar. 2021.

[49] C. You, B. Zheng, and R. Zhang, "Channel estimation and passive beamforming for intelligent reflecting surface: Discrete phase shift and progressive refinement," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2604–2620, Nov. 2020.

[50] N. K. Kundu and M. R. McKay, "Channel estimation for reconfigurable intelligent surface aided MISO communications: From LMMSE to deep learning solutions," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 471–487, Mar. 2021.

[51] Q.-U.-A. Nadeem, H. Alwazani, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface-assisted multi-user MISO communication: Channel estimation and beamforming design," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 661–680, May 2020.

[52] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.

[53] B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface assisted multi-user OFDMA: Channel estimation and training design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8315–8329, Dec. 2020.

[54] Y. Wei, M.-M. Zhao, M.-J. Zhao, and Y. Cai, "Channel estimation for IRS-aided multiuser communications with reduced error propagation," *early access in IEEE Trans. Wireless Commun.*, 2021.

[55] H. Guo and V. K. Lau, "Cascaded channel estimation for intelligent reflecting surface assisted multiuser MISO systems," 2021. [Online]. Available: https://arxiv.org/abs/2108.09002

[56] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.

[57] P. Wang, J. Fang, H. Duan, and H. Li, "Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems," *IEEE Signal Process. Lett.*, vol. 27, no. May, pp. 905–909, 2020.

[58] H. Liu, J. Zhang, Q. Wu, H. Xiao, and B. Ai, "ADMM based channel estimation for RISs aided millimeter wave communications," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 2894–2898, Sept. 2021.

[59] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Trans. Antennas Propagat.*, vol. 34,, no. 3, pp. 276–280, Mar. 1986.

[60] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT - A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 34, no. 5, pp. 1340–1342, Oct. 1986.

[61] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, "Channel estimation for reconfigurable intelligent surface aided multi-user MIMO systems," 2019. [Online]. Available: https://arxiv.org/abs/1912.03619

[62] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[63] K. Ardah, S. Gherekhloo, A. de Almeida, and M. Haardt, "TRICE: A channel estimation framework for RIS-aided millimeter-wave MIMO systems," *IEEE Signal Process. Lett.*, vol. 28, pp. 513–517, Feb. 2021.

[64] T. Lin, X. Yu, Y. Zhu, and R. Schober, "Channel estimation for IRS-assisted millimeter-wave MIMO systems: Sparsity-inspired approaches," 2021. [Online]. Available: https://arxiv.org/abs/2107.11605

[65] J. He, H. Wymeersch, and M. Juntti, "Channel estimation for RIS-aided mmWave MIMO systems via atomic norm minimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5786–5797, Sept. 2021.

[66] X. Wei, D. Shen, and L. Dai, "Channel estimation for RIS assisted wireless communications: Part II - an improved solution based on double-structured sparsity," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1403–1407, May 2021.

[67] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. Lee Swindlehurst, "Channel estimation for RIS-aided multiuser millimeter-wave systems," 2021. [Online]. Available: https://arxiv.org/abs/2106.14792

[68] C. You, B. Zheng, and R. Zhang, "Fast beam training for IRS-assisted multiuser communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1845–1849, Nov. 2020.

[69] B. Ning, Z. Chen, W. Chen, Y. Du, and J. Fang, "Terahertz multi-user massive MIMO with intelligent reflecting surface: Beam training and hybrid beamforming," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1376–1393, Feb. 2021.

[70] S. E. Zegrar, L. Afeef, and H. Arslan, "A general framework for RIS-aided mmWave communication networks: Channel estimation and mobile user tracking," 2020. [Online]. Available: https://arxiv.org/abs/2009.01180

[71] J. Fang, F. Wang, Y. Shen, H. Li, and R. S. Blum, "Super-resolution compressed sensing for line spectral estimation: An iterative reweighted approach," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4649–4662, Sept. 2016.

[72] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[73] Z. Zhang and L. Dai, "A joint precoding framework for wideband reconfigurable intelligent surface-aided cell-free network," *IEEE Trans. Signal Process.*, vol. 69, pp. 4085–4101, Jun. 2021.

[74] T. Bai, C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for intelligent reflecting surface aided wireless powered mobile edge computing in OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5389–5407, Aug. 2021.

[75] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sept. 2020.

[76] W. Cai, H. Li, M. Li, and Q. Liu, "Practical modeling and beamforming for intelligent reflecting surface aided wideband systems," *IEEE Commun. Lett.*, vol. 24, no. 7, pp. 1568–1571, Jul. 2020.

[77] H. Li, W. Cai, Y. Liu, M. Li, Q. Liu, and Q. Wu, "Intelligent reflecting surface enhanced wideband MIMO-OFDM communications: From practical model to reflection optimization," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4807–4820, Jul. 2021.

[78] G. Gradoni and M. Di Renzo, "End-to-end mutual coupling aware communication model for reconfigurable intelligent surfaces: An electromagnetic-compliant approach based on mutual impedances," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 938–942, May 2021.

[79] X. Qian and M. Di Renzo, "Mutual coupling and unit cell aware optimization for reconfigurable intelligent surfaces," *IEEE Wireless Commun. Lett.*, vol. 10, no. 6, pp. 1183–1187, Jun. 2021.

[80] A. Abrardo, D. Dardari, M. Di Renzo, and X. Qian, "MIMO interference channels assisted by reconfigurable intelligent surfaces: Mutual coupling aware sum-rate optimization based on a mutual impedance channel model," *early access in IEEE Wireless Commun. Lett.*, pp. 1–1, 2021.

[81] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[82] K. Shen and W. Yu, "Fractional programming for communication systems - part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[83] J. Chen, Y.-C. Liang, Y. Pei, and H. Guo, "Intelligent reflecting surface: A programmable wireless environment for physical layer security," *IEEE Access*, vol. 7, pp. 82 599–82 612, Jun. 2019.

[84] G. Zhou, C. Pan, H. Ren, K. Wang, M. Elkashlan, and M. Di Renzo, "Stochastic learning-based robust beamforming design for RIS-aided millimeter-wave systems in the presence of random blockages," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 1057–1061, Jan. 2021.

[85] Y. Zhang, C. Zhong, Z. Zhang, and W. Lu, "Sum rate optimization for two way communications with intelligent reflecting surface," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1090–1094, May 2020.

[86] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[87] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[88] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 108–112, Jan. 2020.

[89] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14 960–14 973, Dec. 2020.

[90] G. Yang, X. Xu, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface-assisted non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3137–3151, May 2021.

[91] G. Zhou, C. Pan, H. Ren, K. Wang, and Z. Peng, "Secure wireless communication in RIS-aided MISO system with hardware impairments," *IEEE Wireless Commun. Lett.*, vol. 10, no. 6, pp. 1309–1313, Jun. 2021.

[92] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2021.

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

[94] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jan. 2006.

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

[96] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Intelligent reflecting surface aided multigroup multicast MISO communication systems," *IEEE Trans. Signal Process.*, vol. 68, pp. 3236–3251, Apr. 2020.

[97] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1488–1492, Sept. 2019.

[98] L. Dong and H.-M. Wang, "Secure MIMO transmission via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 787–790, Jun. 2020.

[99] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.

[100] Z. Peng, Z. Zhang, C. Pan, L. Li, and A. L. Swindlehurst, "Multiuser full-duplex two-way communications via intelligent reflecting surface," *IEEE Trans. Signal Process.*, vol. 69, pp. 837–851, Jan. 2021.

[101] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

[102] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces: (invited paper)," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2019, pp. 735–740.

[103] P. Wang, J. Fang, L. Dai, and H. Li, "Joint transceiver and large intelligent surface design for massive MIMO mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1052–1064, Feb. 2021.

[104] H.-M. Wang, J. Bai, and L. Dong, "Intelligent reflecting surfaces assisted secure transmission without eavesdropper's CSI," *IEEE Signal Process. Lett.*, vol. 27, pp. 1300–1304, Jul. 2020.

[105] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[106] X. Yu, D. Xu, and R. Schober, "Enabling secure wireless communications via intelligent reflecting surfaces," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Dec. 2019, pp. 1–6.

[107] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided MIMO communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1823–1838, Aug. 2020.

[108] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted SWIPT under QoS constraints," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1735–1748, Aug. 2020.

[109] Y. Omid, S. M. Shahabi, C. Pan, Y. Deng, and A. Nallanathan, "Low-complexity robust beamforming design for IRS-aided MISO systems with imperfect channels," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1697–1701, May 2021.

[110] M. Fu, Y. Zhou, and Y. Shi, "Intelligent reflecting surface for downlink non-orthogonal multiple access networks," in *Proc. IEEE Global Commun. Conf. Workshops (GC Wkshps)*, Dec. 2019, pp. 1–6.

[111] X. Hu, C. Masouros, and K.-K. Wong, "Reconfigurable intelligent surface aided mobile edge computing: From optimization-based to location-only learning-based solutions," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3709–3725, Jun. 2021.

[112] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 664–674, Jan. 2021.

[113] H. Yang, X. Yuan, J. Fang, and Y.-C. Liang, "Reconfigurable intelligent surface aided constant-envelope wireless power transfer," *IEEE Trans. Signal Process.*, vol. 69, pp. 1347–1361, Feb. 2021.

[114] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A framework of robust transmission design for IRS-aided MISO communications with imperfect cascaded channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 5092–5106, Aug. 2020.

[115] L. Zhang, C. Pan, Y. Wang, H. Ren, and K. Wang, "Robust beamforming design for intelligent reflecting surface aided cognitive radio systems with imperfect cascaded CSI," *early access in IEEE Trans. Cogn. Commun. Netw.*, pp. 1–1, 2021.

[116] Y. Chen, Y. Wang, and L. Jiao, "Robust transmission for reconfigurable intelligent surface aided millimeter wave vehicular communications with statistical CSI," *early access in IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[117] J. Ye, S. Guo, and M.-S. Alouini, "Joint reflecting and precoding designs for SER minimization in reconfigurable intelligent surfaces assisted MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5561–5574, Aug. 2020.

[118] X. Hu, C. Zhong, and Z. Zhang, "Angle-domain intelligent reflecting surface systems: Design and analysis," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4202–4215, Jun. 2021.

[119] M. Shao, Q. Li, and W.-K. Ma, "Minimum symbol-error probability symbol-level precoding with intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1601–1605, Oct. 2020.

[120] S. Wang, Q. Li, and M. Shao, "One-bit symbol-level precoding for MU-MISO downlink with intelligent reflecting surface," *IEEE Signal Process. Lett.*, vol. 27, pp. 1784–1788, Sept. 2020.

[121] N. S. Perović, M. Di Renzo, and M. F. Flanagan, "Channel capacity optimization using reconfigurable intelligent surfaces in indoor mmwave environments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–7.

[122] N. S. Perović, L.-N. Tran, M. Di Renzo, and M. F. Flanagan, "Optimization of RIS-aided MIMO systems via the cutoff rate," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1692–1696, Aug. 2021.

[123] ——, "On the maximum achievable sum-rate of the RIS-aided MIMO broadcast channel," 2021. [Online]. Available: https://arxiv.org/abs/2110.01700

[124] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[125] K. Zhi, C. Pan, H. Ren, and K. Wang, "Ergodic rate analysis of reconfigurable intelligent surface-aided massive MIMO systems with ZF detectors," *early access in IEEE Commun. Lett.*, 2021.

[126] A. Papazafeiropoulos, C. Pan, P. Kourtessis, S. Chatzinotas, and J. M. Senior, "Intelligent reflecting surface-assisted MU-MISO systems with imperfect hardware: Channel estimation and beamforming design," *early access in IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[127] N. S. Perovic, L.-N. Tran, M. Di Renzo, and M. F. Flanagan, "Achievable rate optimization for MIMO systems with reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3865–3882, 2021.

[128] K. Zhi, C. Pan, H. Ren, and K. Wang, "Power scaling law analysis and phase shift optimization of RIS-aided massive MIMO systems with statistical CSI," 2020. [Online]. Available: https://arxiv.org/abs/2010.13525

[129] ——, "Statistical CSI-based design for reconfigurable intelligent surface-aided massive MIMO systems with direct links," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 1128–1132, May 2021.

[130] K. Zhi, C. Pan, H. Ren, K. Wang, M. Elkashlan, M. Di Renzo, R. Schober, H. V. Poor, J. Wang, and L. Han, "Two-timescale design for reconfigurable intelligent surface-aided massive MIMO systems with imperfect CSI," 2021. [Online]. Available: https://arxiv.org/abs/2108.07622

[131] J. Dai, Y. Wang, C. Pan, K. Zhi, H. Ren, and K. Wang, "Reconfigurable intelligent surface aided massive MIMO systems with low-resolution DACs," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 3124–3128, Sept. 2021.

[132] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[133] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.

[134] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[135] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, Mar. 2020.

[136] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[137] M.-M. Zhao, Q. Wu, M.-J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 2–17, Jan. 2021.

[138] J. Yuan, Y.-C. Liang, J. Joung, G. Feng, and E. G. Larsson, "Intelligent reflecting surface-assisted cognitive radio system," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 675–687, Jan. 2021.

[139] S. Hu, Z. Wei, Y. Cai, C. Liu, D. W. K. Ng, and J. Yuan, "Robust and secure sum-rate maximization for multiuser MISO downlink systems with self-sustainable IRS," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 7032–7049, Oct. 2021.

[140] M. Shao, Q. Li, W.-K. Ma, and A. M.-C. So, "A framework for one-bit and constant-envelope precoding over multiuser massive MISO channels," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5309–5324, Oct. 2019.

[141] L. You, J. Xiong, D. W. K. Ng, C. Yuen, W. Wang, and X. Gao, "Energy efficiency and spectral efficiency tradeoff in RIS-aided multiuser MIMO uplink transmission," *IEEE Trans. Signal Process.*, vol. 69, pp. 1407–1421, Jan. 2021.

[142] C. Pradhan, A. Li, L. Song, B. Vucetic, and Y. Li, "Hybrid precoding design for reconfigurable intelligent surface aided mmWave communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 1041–1045, Jul. 2020.

[143] Y. Pan, K. Wang, C. Pan, H. Zhu, and J. Wang, "Sum rate maximization for intelligent reflecting surface assisted terahertz communications," 2020. [Online]. Available: https://arxiv.org/abs/2008.12246

[144] Z. Wan, Z. Gao, F. Gao, M. Di Renzo, and M.-S. Alouini, "Terahertz massive MIMO with holographic reconfigurable intelligent surfaces," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4732–4750, Jul. 2021.

[145] Y. Pan, K. Wang, C. Pan, H. Zhu, and J. Wang, "UAV-assisted and intelligent reflecting surfaces-supported terahertz communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 6, pp. 1256–1260, Jun. 2021.

[146] H. Xie, J. Xu, and Y.-F. Liu, "Max-min fairness in IRS-aided multi-cell MISO systems with joint transmit and reflective beamforming," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1379–1393, Feb. 2021.

[147] S. Buzzi, C. D'Andrea, A. Zappone, M. Fresia, Y.-P. Zhang, and S. Feng, "RIS configuration, beamformer design, and power control in single-cell and multi-cell wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 2, pp. 398–411, Jun. 2021.

[148] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7851–7866, Dec. 2020.

[149] X. Guan, Q. Wu, and R. Zhang, "Intelligent reflecting surface assisted secrecy communication: Is artificial noise helpful or not?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 778–782, Jun. 2020.

[150] Q. Wu and R. Zhang, "Weighted sum power maximization for intelligent reflecting surface aided SWIPT," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 586–590, May 2020.

[151] Y. Pan, K. Wang, C. Pan, H. Zhu, and J. Wang, "Self-sustainable reconfigurable intelligent surface aided simultaneous terahertz information and power transfer (STIPT)," 2021. [Online]. Available: https://arxiv.org/abs/2102.04053

[152] B. Lyu, P. Ramezani, D. T. Hoang, S. Gong, Z. Yang, and A. Jamalipour, "Optimized energy and information relaying in self-sustainable IRS-empowered WPCN," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 619–633, Jan. 2021.

[153] Y. Zheng, S. Bi, Y.-J. A. Zhang, X. Lin, and H. Wang, "Joint beamforming and power control for throughput maximization in IRS-assisted MISO WPCNs," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8399–8410, May 2021.

[154] Z. Feng, B. Clerckx, and Y. Zhao, "Waveform and beamforming design for intelligent reflecting surface aided wireless power transfer: Single-user and multi-user solutions," 2021. [Online]. Available: https://arxiv.org/abs/2101.02674

[155] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2666–2682, Nov. 2020.

[156] T. Bai, C. Pan, C. Han, and L. Hanzo, "Reconfigurable intelligent surface aided mobile edge computing," *Accepted by IEEE Wireless Commun.*, 2021.

[157] F. Zhou, C. You, and R. Zhang, "Delay-optimal scheduling for IRS-aided mobile edge computing," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 740–744, Apr. 2021.

[158] Z. Chu, P. Xiao, M. Shojafar, D. Mi, J. Mao, and W. Hao, "Intelligent reflecting surface assisted mobile edge computing for internet of things," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 619–623, Mar. 2021.

[159] Z. Li, M. Chen, Z. Yang, J. Zhao, Y. Wang, J. Shi, and C. Huang, "Energy efficient reconfigurable intelligent surface enabled mobile edge computing networks with NOMA," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 2, pp. 427–440, Jun. 2021.

[160] Y. Cao, T. Lv, Z. Lin, and W. Ni, "Delay-constrained joint power control, user detection and passive beamforming in intelligent reflecting surface-assisted uplink mmWave system," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 2, pp. 482–495, Jun. 2021.

[161] L. Du, S. Shao, G. Yang, J. Ma, Q. Liang, and Y. Tang, "Capacity characterization for reconfigurable intelligent surfaces assisted multiple-antenna multicast," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6940–6953, Oct. 2021.

[162] L. Zhang, Y. Wang, W. Tao, Z. Jia, T. Song, and C. Pan, "Intelligent reflecting surface aided MIMO cognitive radio systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11 445–11 457, Oct. 2020.

[163] X. Guan, Q. Wu, and R. Zhang, "Joint power control and passive beamforming in IRS-assisted spectrum sharing," *IEEE Commun. Lett.*, vol. 24, no. 7, pp. 1553–1557, Jul. 2020.

[164] B. Zheng, Q. Wu, and R. Zhang, "Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA?" *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 753–757, Apr. 2020.

[165] H. Wang, C. Liu, Z. Shi, Y. Fu, and R. Song, "On power minimization for IRS-aided downlink NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1808–1811, Nov. 2020.

[166] J. Zhu, Y. Huang, J. Wang, K. Navaie, and Z. Ding, "Power efficient IRS-assisted NOMA," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 900–913, Feb. 2021.

[167] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Capacity and optimal resource allocation for IRS-assisted multi-user communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3771–3786, Jun. 2021.

[168] J. Zuo, Y. Liu, Z. Qin, and N. Al-Dhahir, "Resource allocation in intelligent reflecting surface assisted NOMA systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7170–7183, Nov. 2020.

[169] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, Oct. 2020.

[170] Y. Cai, M.-M. Zhao, K. Xu, and R. Zhang, "Intelligent reflecting surface aided full-duplex communication: Passive beamforming and deployment design," *IEEE Trans. Wireless Commun.*, pp. early access,, 2021.

[171] G. Zhou, C. Pan, H. Ren, K. Wang, M. Di Renzo, and A. Nallanathan, "Robust beamforming design for intelligent reflecting surface aided MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1658–1662, Oct. 2020.

[172] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[173] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for IRS-assisted full-duplex cognitive radio systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7376–7394, Dec. 2020.

[174] Q. Wang, F. Zhou, R. Q. Hu, and Y. Qian, "Energy efficient robust beamforming and cooperative jamming design for IRS-assisted MISO networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2592–2607, Apr. 2021.

[175] J. Zhang, Y. Zhang, C. Zhong, and Z. Zhang, "Robust design for intelligent reflecting surfaces assisted MISO systems," *IEEE Commun. Lett.*, vol. 24, no. 10, pp. 2353–2357, Oct. 2020.

[176] Z. Zhang, L. Lv, Q. Wu, H. Deng, and J. Chen, "Robust and secure communications in intelligent reflecting surface assisted NOMA networks," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 739–743, Mar. 2021.

[177] S. Hong, C. Pan, H. Ren, K. Wang, K. K. Chai, and A. Nallanathan, "Robust transmission design for intelligent reflecting surface-aided secure communication systems with imperfect cascaded CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2487–2501, Apr. 2021.

[178] L. Dong, H.-M. Wang, and H. Xiao, "Secure cognitive radio communication via intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4678–4690, Jul. 2021.

[179] X. Yu, D. Xu, D. W. K. Ng, and R. Schober, "IRS-assisted green communication systems: Provable convergence and robust optimization," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6313–6329, Sept. 2021.

[180] K.-Y. Wang, A. M.-C. So, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5690–5705, Nov. 2014.

[181] S. Hong, C. Pan, G. Zhou, H. Ren, and K. Wang, "Outage constrained robust transmission design for IRS-aided secure communications with direct communication links," https://arxiv.org/abs/2011.09822 2020.

[182] M.-M. Zhao, A. Liu, and R. Zhang, "Outage-constrained robust beamforming for intelligent reflecting surface aided wireless communication," *IEEE Trans. Signal Process.*, vol. 69, pp. 1301–1316, Feb. 2021.

[183] N. K. Kundu and M. R. Mckay, "Large intelligent surfaces with channel estimation overhead: Achievable rate and optimal configuration," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 986–990, May 2021.

[184] A. Zappone, M. Di Renzo, F. Shams, X. Qian, and M. Debbah, "Overhead-aware design of reconfigurable intelligent surfaces in smart radio environments," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 126–141, Jan. 2021.

[185] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, Aug. 2019.

[186] Y.-C. Liang and F. Chin, "Downlink channel covariance matrix (DCCM) estimation and its applications in wireless DS-CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 2, pp. 222–232, Feb. 2001.

[187] E. Björnson and L. Sanguinetti, "Rayleigh fading modeling and channel hardening for reconfigurable intelligent surfaces," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 830–834, Apr. 2021.

[188] Y. Jia, C. Ye, and Y. Cui, "Analysis and optimization of an intelligent reflecting surface-assisted system with interference," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8068–8082, Dec. 2020.

[189] A. Papazafeiropoulos, C. Pan, A. Elbir, P. Kourtessis, S. Chatzinotas, and J. M. Senior, "Coverage probability of distributed IRS systems under spatially correlated channels," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1722–1726, Aug. 2021.

[190] H. Guo, Y.-C. Liang, and S. Xiao, "Intelligent reflecting surface configuration with historical channel observations," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1821–1824, Nov. 2020.

[191] B. Al-Nahhas, Q.-U.-A. Nadeem, and A. Chaaban, "Reconfigurable intelligent surface aided communications: Asymptotic analysis under imperfect CSI," 2021. [Online]. Available: https://arxiv.org/abs/2108.09388

[192] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.

[193] M.-M. Zhao, A. Liu, Y. Wan, and R. Zhang, "Two-timescale beamforming optimization for intelligent reflecting surface aided multiuser communication with QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6179–6194, Sept. 2021.

[194] J. Dai, F. Zhu, C. Pan, H. Ren, and K. Wang, "Statistical CSI-based transmission design for reconfigurable intelligent surface-aided massive MIMO systems with hardware impairments," *IEEE Wireless Commun. Lett.*, pp. 1–1, Oct. 2021.

[195] Z. Wang, L. Liu, S. Zhang, and S. Cui, "Massive MIMO communication with intelligent reflecting surface," 2021. [Online]. Available: https://arxiv.org/abs/2107.04255

[196] K. Zhi, C. Pan, G. Zhou, H. Ren, M. Elkashlan, and R. Schober, "Is RIS-aided massive MIMO promising with ZF detectors and imperfect CSI?" 2021. [Online]. Available: https://arxiv.org/abs/2111.01585

[197] A. Papazafeiropoulos, C. Pan, A. Elbir, V. Nguyen, P. Kourtessis, and S. Chatzinotas, "Asymptotic analysis of max-min weighted SINR for IRS-assisted MISO systems with hardware impairments," *IEEE Wireless Commun. Lett.*, pp. 1–1, Jul. 2021.

[198] X. Hu, J. Wang, and C. Zhong, "Statistical CSI based design for intelligent reflecting surface assisted MISO systems," *Sci. China Inf. Sci.*, vol. 63, no. 12, pp. 1–10, Oct. 2020.

[199] J. Wang, H. Wang, Y. Han, S. Jin, and X. Li, "Joint transmit beamforming and phase shift design for reconfigurable intelligent surface assisted MIMO systems," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 2, pp. 354–368, Jun. 2021.

[200] X. Gan, C. Zhong, C. Huang, and Z. Zhang, "RIS-assisted multiuser MISO communications exploiting statistical CSI," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6781–6792, Oct. 2021.

[201] Z. Peng, T. Li, C. Pan, H. Ren, W. Xu, and M. Di Renzo, "Analysis and optimization for RIS-aided multi-pair communications relying on statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3897–3901, Apr. 2021.

[202] C. Luo, X. Li, S. Jin, and Y. Chen, "Reconfigurable intelligent surface-assisted multi-cell MISO communication systems exploiting statistical CSI," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2313–2317, Oct. 2021.

[203] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. L. Swindlehurst, "UAV-assisted intelligent reflecting surface symbiotic radio system," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5769–5785, Sept. 2021.

[204] J. Zhang, J. Liu, S. Ma, C.-K. Wen, and S. Jin, "Large system achievable rate analysis of RIS-assisted MIMO wireless communication with statistical CSIT," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5572–5585, Sept. 2021.

[205] X. Zhang, X. Yu, S. Song, and K. B. Letaief, "Large-scale IRS-aided MIMO over double-scattering channel: An asymptotic approach," 2021. [Online]. Available: https://arxiv.org/abs/2104.06125

[206] L. You, J. Xiong, Y. Huang, D. W. K. Ng, C. Pan, W. Wang, and X. Gao, "Reconfigurable intelligent surfaces-assisted multiuser MIMO uplink transmission with partial CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5613–5627, Sept. 2021.

[207] K. Xu, J. Zhang, X. Yang, S. Ma, and G. Yang, "On the sum-rate of RIS-assisted MIMO multiple-access channels over spatially correlated rician fading," *early access in IEEE Trans. Commun.*, pp. 1–1, 2021.

[208] H. Ren, C. Pan, L. Wang, Z. Kou, and K. Wang, "Long-term CSI-based design for RIS-aided multiuser MISO systems exploiting deep reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/2111.12212v1

[209] X. Hu, C. Zhong, Y. Zhang, X. Chen, and Z. Zhang, "Location information aided multiple intelligent reflecting surface systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7948–7962, Dec. 2020.

[210] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[211] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.

[212] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.

[213] Z. Abu-Shaban, X. Zhou, T. Abhayapala, G. Seco-Granados, and H. Wymeersch, "Error bounds for uplink and downlink 3D localization in 5G millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 4939–4954, Aug. 2018.

[214] J. J. Caffery, *Wireless location in CDMA cellular radio systems*. Springer Science & Business Media, 2000.

[215] H. Wymeersch, G. Seco-Granados, G. Destino, D. Dardari, and F. Tufvesson, "5G mmWave positioning for vehicular networks," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 80–86, Dec. 2017.

[216] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "Position and orientation estimation through millimeter-wave MIMO in 5G systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1822–1835, Mar. 2017.

[217] A. Guerra, F. Guidi, and D. Dardari, "Single-anchor localization and orientation performance limits using massive arrays: MIMO vs. beamforming," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5241–5255, Aug. 2018.

[218] J. He, H. Wymeersch, L. Kong, O. Silvén, and M. Juntti, "Large intelligent surface for positioning in millimeter wave MIMO systems," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, Jun. 2020, pp. 1–5.

[219] Y. Liu, E. Liu, R. Wang, and Y. Geng, "Reconfigurable intelligent surface aided wireless localization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Aug. 2021, pp. 1–6.

[220] A. Elzanaty, A. Guerra, F. Guidi, and M.-S. Alouini, "Reconfigurable intelligent surfaces for localization: Position and orientation error bounds," *IEEE Trans. Signal Process.*, vol. 69, pp. 5386–5402, Aug. 2021.

[221] W. Wang and W. Zhang, "Joint beam training and positioning for intelligent reflecting surfaces assisted millimeter wave communications," *IEEE Trans. Wireless Commun.*, Oct. 2021.

[222] H. Zhang, H. Zhang, B. Di, K. Bian, Z. Han, and L. Song, "Towards ubiquitous positioning by leveraging reconfigurable intelligent surface," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 284–288, Jan. 2020.

[223] ——, "Metalocalization: Reconfigurable intelligent surface aided multi-user wireless indoor localization," *early access in IEEE Trans. Wireless Commun.*, 2021.

[224] R. Wang, Z. Xing, and E. Liu, "Joint location and communication study for intelligent reflecting surface aided wireless communication system," 2021. [Online]. Available: https://arxiv.org/abs/2103.01063

[225] J. He, H. Wymeersch, T. Sanguanpuak, O. Silvén, and M. Juntti, "Adaptive beamforming design for mmWave RIS-aided joint localization and communication," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Jun. 2020, pp. 1–6.

[226] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?" 2021. [Online]. Available: https://arxiv.org/abs/2108.07581

[227] B. Friedlander, "Localization of signals in the near-field of an antenna array," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 3885–3893, Aug. 2019.

[228] Z. Abu-Shaban, K. Keykhosravi, M. F. Keskin, G. C. Alexandropoulos, G. Seco-Granados, and H. Wymeersch, "Near-field localization with a reconfigurable intelligent surface acting as lens," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Aug. 2021, pp. 1–6.

[229] M. Rahal, B. Denis, K. Keykhosravi, B. Uguen, and H. Wymeersch, "RIS-enabled localization continuity under near-field conditions," 2021. [Online]. Available: http://arxiv.org/pdf/2109.11965v1

[230] *Study on channel model for frequencies from 0.5 to 100 GHz (V15.0.0; 3GPP TR 38.901 version 15.0.0 Release 15)*, Std.

[231] Y.-T. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994.

[232] W. H. Foy, "Position-location solutions by Taylor-series estimation," *IEEE Trans. Aerosp. Electron. Syst.*, no. 2, pp. 187–194, Mar. 1976.

[233] Y. Liu, S. Hong, C. Pan, Y. Wang, Y. Pan, and M. Chen, "Optimization of RIS configurations for multiple-RIS-aided mmwave positioning systems based on CRLB analysis," 2021. [Online]. Available: https://arxiv.org/abs/2111.14023

[234] S. M. Kay, "Fundamentals of statistical signal processing," *Technometrics*, vol. 37, no. 4, pp. 465–466, 1993.

[235] K. T. Selvan and R. Janaswamy, "Fraunhofer and fresnel distances: Unified derivation for aperture antennas," *IEEE Antennas Propag. Mag.*, vol. 59, no. 4, pp. 12–15, Aug. 2017.

[236] K. Dovelos, S. D. Assimonis, H. Quoc Ngo, B. Bellalta, and M. Matthaiou, "Intelligent reflecting surfaces at Terahertz bands: Channel modeling and analysis," pp. 1–6, Jul. 2021.

[237] F. H. Danufane, M. Di Renzo, J. de Rosny, and S. Tretyakov, "On the path-loss of reconfigurable intelligent surfaces: An approach based on green's theorem applied to vector fields," 2020.

[238] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, 2020.

[239] B. Sihlbom, M. I. Poulakis, and M. Di Renzo, "Reconfigurable intelligent surfaces: Performance assessment through a system-level simulator," 2021.

[240] X. Yin, S. Wang, N. Zhang, and B. Ai, "Scatterer localization using large-scale antenna arrays based on a spherical wave-front parametric model," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6543–6556, Oct. 2017.

[241] F. Bohagen, P. Orten, and G. E. Oien, "Design of optimal high-rank line-of-sight MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1420–1425, Apr. 2007.

[242] F. Bohagen, P. Orten, and G. Oien, "On spherical vs. plane wave modeling of line-of-sight mimo channels," *IEEE Trans. Commun.*, vol. 57, no. 3, pp. 841–849, 2009.

[243] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1843–1854, Aug. 2002.

[244] X. Wei and L. Dai, "Channel estimation for extremely large-scale massive mimo: Far-field, near-field, or hybrid-field?" *IEEE Communications Letters*, pp. 1–1, 2021.

[245] Z. Mao, M. Peng, and X. Liu, "Channel estimation for reconfigurable intelligent surface assisted wireless communication systems in mobility scenarios," *China Commun.*, vol. 18, no. 3, pp. 29–38, Mar. 2021.

[246] S. Sun and H. Yan, "Channel estimation for reconfigurable intelligent surface-assisted wireless communications considering doppler effect," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 790–794, Apr. 2021.

[247] B. Matthiesen, E. Björnson, E. De Carvalho, and P. Popovski, "Intelligent reflecting surface operation under predictable receiver mobility: A continuous time propagation model," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 216–220, Feb. 2021.

[248] Z. Huang, B. Zheng, and R. Zhang, "Transforming fading channel from fast to slow: Intelligent refracting surface aided high-mobility communication," 2021. [Online]. Available: https://arxiv.org/abs/2106.02274

[249] C. Feng, H. Lu, Y. Zeng, S. Jin, and R. Zhang, "Wireless communication with extremely large-scale intelligent reflecting surface," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Sept. 2021, pp. 165–170.

[250] E. Björnson, Ö. T. Demir, and L. Sanguinetti, "A primer on near-field beamforming for arrays and reconfigurable intelligent surfaces," 2021. [Online]. Available: https://arxiv.org/abs/2110.06661

[251] X. Wei, L. Dai, Y. Zhao, G. Yu, and X. Duan, "Codebook design and beam training for extremely large-scale RIS: Far-field or near-field?" 2021. [Online]. Available: https://arxiv.org/abs/2109.10143

[252] F. H. Danufane, M. Di Renzo, J. de Rosny, and S. Tretyakov, "On the path-loss of reconfigurable intelligent surfaces: An approach based on green?s theorem applied to vector fields," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5573–5592, Aug. 2021.

[253] Z. Zhang, L. Dai, X. Chen, C. Liu, F. Yang, R. Schober, and H. V. Poor, "Active RIS vs. passive RIS: Which will prevail in 6G?" 2021. [Online]. Available: https://arxiv.org/abs/2103.15154

[254] D. Xu, X. Yu, D. W. K. Ng, and R. Schober, "Resource allocation for active IRS-assisted multiuser communication systems," 2021. [Online]. Available: https://arxiv.org/abs/2108.13033

[255] C. You and R. Zhang, "Wireless communication aided by intelligent reflecting surface: Active or passive?" *early access in IEEE Wireless Commun. Lett.*, 2021.

[256] M. H. Khoshafa, T. M. N. Ngatched, M. H. Ahmed, and A. R. Ndjiongue, "Active reconfigurable intelligent surfaces-aided wireless communication system," *early access in IEEE Commun.Lett.*, 2021.

[257] Y. Han, S. Zhang, L. Duan, and R. Zhang, "Cooperative double-IRS aided communication: Beamforming design and power scaling," *IEEE Wireless Commun. Letters*, vol. 9, no. 8, pp. 1206–1210, Aug. 2020.

[258] B. Zheng, C. You, and R. Zhang, "Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4513–4526, Jul. 2021.

[259] W. Mei and R. Zhang, "Multi-beam multi-hop routing for intelligent reflecting surfaces aided massive MIMO," *IEEE Trans. Wireless Commun., early access*, pp. 1–1, 2021.

[260] H. Niu, Z. Chu, F. Zhou, C. Pan, D. W. K. Ng, and H. X. Nguyen, "Double intelligent reflecting surface-assisted multi-user MIMO mmwave systems with hybrid precoding," *IEEE Trans. Veh. Tech., early access*, pp. 1–1, 2021.

[261] C. You, B. Zheng, and R. Zhang, "Wireless communication via double IRS: Channel estimation and passive beamforming designs," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 431–435, Feb. 2021.

[262] B. Zheng, C. You, and R. Zhang, "Uplink channel estimation for double-IRS assisted multi-user MIMO," in *IEEE Int. Conf. Commun. (ICC), Montreal, Canada*, Jun. 2021, pp. 1–6.

[263] ——, "Efficient channel estimation for double-IRS aided multi-user MIMO system," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3818–3832, Jun. 2021.

[264] K. W. Cho, M. H. Mazaheri, J. Gummeson, O. Abari, and K. Jamieson, "mmWall: A reconfigurable metamaterial surface for mmWave networks," in *Proc. 22nd Int. Workshop Mobile Comput. Syst. Appl.*, 2021, pp. 119–125.

[265] S. Zhang, H. Zhang, B. Di, Y. Tan, M. Di Renzo, Z. Han, H. Vincent Poor, and L. Song, "Intelligent omni-surfaces: Ubiquitous wireless transmission by reflective-refractive metasurfaces," *IEEE Trans. Wireless Commun.*, early access, 2021.

[266] H. Zhang, S. Zeng, B. Di, Y. Tan, M. Di Renzo, M. Debbah, L. Song, Z. Han, and H. V. Poor, "Intelligent omni-surfaces for full-dimensional wireless communications: Principle, technology, and implementation," 2021. [Online]. Available: http://arxiv-export-lb.library.cornell.edu/abs/2104.12313

[267] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, H. V. Poor, and L. Hanzo, "STAR: Simultaneous transmission and reflection for 360° coverage by intelligent surfaces," 2021. [Online]. Available: https://arxiv.org/abs/2103.09104

[268] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," 2021. [Online]. Available: https://arxiv.org/abs/2104.01421