Spatial Channel Covariance Estimation and Two-Timescale Beamforming for IRS-Assisted Millimeter Wave Systems

Hongwei Wang, Jun Fang, Huiping Duan, and Hongbin Li, *Fellow, IEEE*

***Abstract*— We consider the problem of spatial channel covariance matrix (CCM) estimation for intelligent reflecting surface (IRS)-assisted millimeter wave (mmWave) communication systems. Spatial CCM is essential for two-timescale beamforming in IRS-assisted systems; however, estimating the spatial CCM is challenging due to the passive nature of reflecting elements and the large size of the CCM resulting from massive reflecting elements of the IRS. In this paper, we propose a CCM estimation method by exploiting the low-rankness as well as the positive semi-definite (PSD) 3-level Toeplitz structure of the CCM. Estimation of the CCM is formulated as a semidefinite programming (SDP) problem and an alternating direction method of multipliers (ADMM) algorithm is developed. Our analysis shows that the proposed method is theoretically guaranteed to attain a reliable CCM estimate with a sample complexity much smaller than the dimension of the CCM. Thus the proposed method can help achieve a significant training overhead reduction. Simulation results are presented to illustrate the effectiveness of our proposed method and the performance of two-timescale beamforming scheme based on the estimated CCM.**

***Index Terms*— Intelligent reflecting surface, millimeter wave communications, spatial channel covariance estimation.**

I. Introduction

Millimeter Wave (mmWave) communication is considered as a promising technology for future cellular networks due to its potential to offer gigabits-per-second communication data rates [1]. Nevertheless, due to the small wavelength, mmWave signals have limited diffraction and scattering abilities. As a result, mmWave communications are vulnerable to blockage events, which can be frequent in indoor and dense urban environments. Intelligent reflecting surface (IRS) has been recently introduced as a cost-effective and energy-efficient solution to address the blockage issue for mmWave communications [2]. The IRS, also referred to as reconfigurable intelligent surface (RIS), is a planar array made of a newly developed metamaterial. It comprises a large number of reconfigurable passive elements, each of which can independently reflect the incident signal with a reconfigurable phase shift. By properly adjusting the phase shifts of the passive elements, IRS can help realize a programmable and desirable wireless propagation environment [3]–[5].

Channel state information (CSI) acquisition is a prerequisite to achieve the full potential of IRS-assisted mmWave systems. There have been a plethora of studies on how to acquire the instantaneous CSI (I-CSI) for IRS-assisted mmWave systems. Specifically, to reduce the training overhead, some works exploited the inherent sparsity of mmWave channels and developed compressed sensing-based methods to estimate the cascade channel [6]–[8]. Other works, e.g., [9]–[11], developed tensor decomposition-based channel estimation methods by utilizing some intrinsic multi-dimensional structure of cascade channels. Despite these efforts, system optimization based on I-CSI is still considered as a formidable task due to the following difficulties. First, the coherence time of mmWave channels is drastically shorter than that of sub-6GHz channels. This implies that channel estimation and system optimization (i.e. joint active/passive beamforming) should be performed more frequently, which entails a significant amount of training overhead and tremendous computational resources. Second, system optimization based on I-CSI requires frequent transmissions of control signals from the base station (BS) to the IRS, which involves a considerable amount of signalling overhead.

To address the above difficulties, some attempts have been made by exploiting channel statistics for joint active and passive beamforming, e.g., [12]–[19]. Specifically, in [15], a two-timescale beamforming protocol was proposed for IRS-assisted systems, where the reflecting coefficients at the IRS are designed according to the long-term (i.e., statistical) CSI, and the transmit beamforming matrix is devised based on the instantaneous equivalent channel in a short-term scale. In addition, a model-driven particle swarm optimization scheme was proposed in [19] for two timescale beamforming based on the statistical CSI. Statistical CSI, usually characterized by the spatial CCM, is essential for two-timescale beamforming in IRS-assisted systems. Obtaining the spatial CCM, however, is challenging due to the passive nature of reflecting elements and the large size of the CCM resulting from massive reflecting elements of the IRS. To our best knowledge, how to estimate the spatial CCM for IRS-assisted mmWave systems has not been reported before. Although there are some works on CCM estimation for conventional mmWave systems, e.g., [20]–[22],

Hongwei Wang, and Jun Fang are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China, Email: JunFang@uestc.edu.cn

Huiping Duan is with the School of Information and Communications Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, Email: huipingduan@uestc.edu.cn

Hongbin Li is with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA, E-mail: Hongbin.Li@stevens.edu

This work was supported in part by the National Science Foundation of China under Grant 61829103. The work of H. Li was supported in part by the National Science Foundation under Grants ECCS-1923739 and ECCS-2212940.

these methods cannot be straightforwardly extended to the IRS-aided systems.

In this paper, we propose a CCM estimation method for IRS-assisted mmWave systems. The proposed method exploits the low-rankness as well as the positive semi-definite (PSD) 3-level Toeplitz structure of the CCM, and is theoretically guaranteed to attain a reliable CCM estimate with a sample complexity much smaller than the dimension of the CCM. Particularly, our work has the following contributions:

1) Due to the passive nature of the IRS, it is almost impossible to obtain the CCM for the BS-IRS link and the IRS-user link. The first contribution of this work is to show that the knowledge of the cascade channel's CCM can be utilized to perform long-term beamforming (i.e., optimization of reflection coefficients) for both single-user and multi-user scenarios.
2) The second contribution of this paper is to formulate the CCM estimation problem as a semidefinite programming (SDP) problem by exploiting the low-rank and Toeplitz structures of the CCM. We also develop an alternating direction method of multipliers (ADMM) algorithm for CCM estimation. Although ADMM has a standard procedure, how to efficiently solve some of its subproblems is non-trivial.
3) We provide a theoretical analysis of the proposed CCM estimation method, which shows that a significant overhead reduction can be achieved by exploiting the low-rank structure of the CCM. Our analysis generalizes the partial observation model [23] to an arbitrarily linear compression model. The generalized result is more useful for CCM estimation in practical mmWave systems, where, due to the hybrid analog-and-digital structure, direct access to the entries of the sample CCM is usually unavailable.

The rest of the paper is organized as follows. Section II discusses the system model, channel model, and the motivation of the work. Details about the downlink training and the received signal model are presented in Section III. Section IV proposes an ADMM-based algorithm for CCM estimation. Performance of the proposed CCM estimation method is analyzed in Section V. Section VI discusses how to perform two-timescale beamforming based on the estimated CCM. Simulation results are provided in Section VII, followed by concluding remarks in Section VIII.

Notations: Italic letters denote scalars. Boldface lowercase and uppercase denote the vectors and matrices, respectively. Superscripts $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ denote conjugate, transpose, and conjugate transpose, respectively. $\mathbb{E}(\cdot)$ is the expectation operator and $j=\sqrt{-1}$. vec$(\cdot)$ is the vectorization operation, which stacks the columns of a matrix on top of each other. $\boldsymbol{A} \succcurlyeq \boldsymbol{B}$ means $\boldsymbol{A}-\boldsymbol{B}$ is a positive semidefinite matrix. The transposed Khatri-Rho, Hadamard and Kronecker product are denoted by $\bullet$, $\circ$, and $\otimes$ respectively. $\mathbb{CN}(\mu,\sigma^2)$ means a circularly symmetric complex Gaussian distribution with mean μ and variance σ^2. $\mathbb{N}(\mu,\sigma^2)$ denotes a real Gaussian distribution with mean μ and variance σ^2. $\mathbb{C}^{N\times M}$ represents the complex space with $N\times M$ dimension.

II. Models and Motivations

A. System Model

We consider a point-to-point IRS-aided mmWave communication system, where an IRS is deployed to assist data transmission from the base station (BS) to an omnidirectional-antenna user. The BS is equipped with a uniform linear array with N antennas. The IRS is a uniform planar array consisting of $M=M_v\times M_h$ passive reflecting elements. Each element can independently reflect the incident signal with a reconfigurable phase shift. Let $\boldsymbol{\Psi}=\text{diag}\left(e^{j\psi_1},\cdots,e^{j\psi_M}\right)$ denote the reflecting coefficient matrix of the IRS, where ψ_m is the phase shift associated with the mth passive element.

For simplicity, we assume that the direct link between the BS and the user is blocked due to poor propagation conditions, and the transmitted signal arrives at the user via the BS-IRS-user channel. Let $\boldsymbol{G}\in\mathbb{C}^{M\times N}$ and $\boldsymbol{h}\in\mathbb{C}^{M\times 1}$ denote the BS-IRS channel and the IRS-user channel, respectively. The effective channel between the BS and the user can thus be expressed as

$$\tilde{\boldsymbol{h}}^H=\boldsymbol{h}^H\boldsymbol{\Psi}\boldsymbol{G}=\boldsymbol{\psi}^T\text{diag}(\boldsymbol{h}^H)\boldsymbol{G}\triangleq\boldsymbol{\psi}^T\boldsymbol{H} \tag{1}$$

where $\boldsymbol{\psi}=\text{diag}(\boldsymbol{\Psi})$ and $\boldsymbol{H}\triangleq\text{diag}(\boldsymbol{h}^H)\boldsymbol{G}$ is referred to as the cascade channel.

We adopt a geometric mmWave channel model [24] to characterize the channel. For notational convenience, we first define

$$\boldsymbol{a}(\nu,D)\triangleq[1\ \ \cdots\ \ e^{j(D-1)\nu}]^T \tag{2}$$

The BS-IRS channel $\boldsymbol{G}$ can be expressed as

$$\boldsymbol{G}=\sum_{l=1}^{L}\alpha_l\boldsymbol{a}_r(\theta_{v,l},\theta_{h,l})\boldsymbol{a}_t^H(\gamma_l) \tag{3}$$

where L is the number of paths between the BS and the IRS, α_l is the complex gain which is assumed to follow a complex Gaussian distribution $\mathbb{CN}(0,\varpi_l^2)$, $\{\gamma_l,\theta_{h,l},\theta_{v,l}\}$ are, respectively, the angle of departure (AoD), the elevation and azimuth angle of arrival (AoA) associated with the lth path; and $\boldsymbol{a}_t(\gamma_l)$ and $\boldsymbol{a}_r(\theta_{v,l},\theta_{h,l})$ are the transmit and receive array response vectors. Specifically, we have

$$\boldsymbol{a}_t(\gamma_l)=\boldsymbol{a}(\nu_{1,l},N) \tag{4}$$

where $\nu_{1,l}\triangleq\frac{2\pi d}{\lambda}\sin(\gamma_l)$ with λ and d representing the signal wavelength and the antenna spacing, respectively. $\boldsymbol{a}_r(\theta_{v,l},\theta_{h,l})$ has a form of a Kronecker product as

$$\boldsymbol{a}_r(\theta_{v,l},\theta_{h,l})=\boldsymbol{a}(\nu_{2,l},M_v)\otimes\boldsymbol{a}(\nu_{3,l},M_h) \tag{5}$$

where $\nu_{2,l}\triangleq\frac{2\pi d}{\lambda}\cos(\theta_{v,l})$ and $\nu_{3,l}\triangleq\frac{2\pi d}{\lambda}\sin(\theta_{v,l})\cos(\theta_{h,l})$. Define $\boldsymbol{\alpha}\triangleq[\alpha_1\ \ \cdots\ \ \alpha_L]^T$, $\boldsymbol{\Sigma}\triangleq\text{diag}(\boldsymbol{\alpha})$, $\boldsymbol{F}_1\triangleq[\boldsymbol{a}_r(\theta_{v,1},\theta_{h,1}))\ \cdots\ \boldsymbol{a}_r(\theta_{v,L},\theta_{h,L})]$, and $\boldsymbol{F}_2\triangleq[\boldsymbol{a}(\nu_{1,1},N)\ \cdots\ \boldsymbol{a}(\nu_{1,L},N)]$, we can express $\boldsymbol{G}$ as

$$\boldsymbol{G}=\boldsymbol{F}_1\boldsymbol{\Sigma}\boldsymbol{F}_2^H \tag{6}$$

Similarly, the IRS-user channel is characterized as

$$\boldsymbol{h}=\sum_{p=1}^{P}\beta_p\boldsymbol{a}_r(\psi_{v,p},\psi_{h,p}) \tag{7}$$

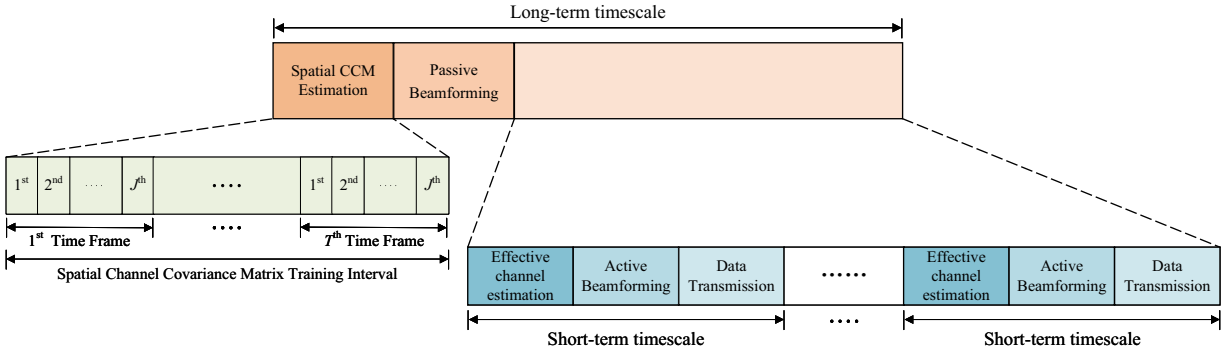


Fig. 1: Two-timescale beamforming and spatial CCM training protocol.

where P is the number of paths between the IRS and the user, β_p is the complex channel gain which follows $\mathbb{CN}(0, \chi_p^2)$, $\{\psi_{v,p}, \psi_{h,p}\}$ denotes the elevation and azimuth AoD associated with the pth path, and $\boldsymbol{a}_r(\psi_{v,p}, \psi_{h,p})$ can be expressed as

$$\boldsymbol{a}_r(\psi_{v,p}, \psi_{h,p}) = \boldsymbol{a}(\nu_{4,p}, M_v) \otimes \boldsymbol{a}(\nu_{5,p}, M_h) \tag{8}$$

where $\nu_{4,p} \triangleq \frac{2\pi d}{\lambda} \cos(\psi_{v,p})$ and $\nu_{5,p} \triangleq \frac{2\pi d}{\lambda} \sin(\psi_{v,p}) \cos(\psi_{h,p})$. The IRS-user channel can be written as

$$\boldsymbol{h} = \boldsymbol{F}_3 \boldsymbol{\beta} \tag{9}$$

where $\boldsymbol{F}_3 \triangleq [\boldsymbol{a}_r(\psi_{v,1}, \psi_{h,1}) \ \cdots \ \boldsymbol{a}_r(\psi_{v,P}, \psi_{h,P})]$ and $\boldsymbol{\beta} \triangleq [\beta_1 \ \cdots \ \beta_P]^T$.

B. Motivations

Although I-CSI helps achieve optimal beamforming performance, system optimization based on I-CSI is a challenging task due to the following reasons. First, to conduct optimal beamforming, the knowledge of a large-size $N \times M$ cascade channel matrix is required [6], whose estimation involves a large amount of training overhead. Second, due to the short coherence time of mmWave channels, estimation of I-CSI needs to be frequently performed. In addition, optimal beamforming requires the reflection coefficients to be updated and configured to the IRS for each channel realization, which entails a significant amount of signalling overhead from the BS to the IRS.

A two-timescale beamforming scheme [15] is a promising solution to address the above challenge. The key idea of two-timescale beamforming is to perform beamforming on two time scales, namely, a long-term timescale and a short-term timescale (see Fig. 1). Specifically, the passive beamforming is conducted in a long-term timescale, where the reflecting coefficients at the IRS are devised according to the statistical CSI which varies much more slowly than the I-CSI, whereas the active beamforming is performed in a short-term timescale, where the BS's precoder is determined based on the instantaneous effective (or equivalent) channel that remains unaltered over an interval of channel coherent time.

For the two-timescale beamforming scheme, at each small-scale interval, we only need to estimate the I-CSI of the effective channel between the BS and the user, which has a much smaller size than the cascade channel. Also, the reflection coefficients are devised based on statistical CSI, which can remain fixed over a large-scale interval that is much larger than the channel coherence time. Therefore, compared with beamforming based on I-CSI, two-timescale beamforming has the potential to achieve a significant reduction in training and signalling overhead.

A prerequisite for the two-timescale beamforming scheme is to obtain the statistical CSI that is necessary for performing the long-term passive beamforming. Due to the passive nature of the IRS, it is almost impossible to obtain the statistical CSI for each individual link. To address this challenge, in this work the statistical CSI acquisition problem is formulated into a cascade channel CCM estimation problem and we show that long-term passive beamforming can be accomplished with the knowledge of the cascade channel's CCM.

III. Downlink Training

We consider a downlink training procedure, where the downlink training process consists of T time frames, and each time frame is further divided into J time slots (see Fig. 1). We assume that each time frame has a short period of time so that the channel remains unaltered. During the training process, the BS sends a same pilot signal, i.e., $s_{t,j} = 1, \ \forall j \in \{1, \cdots, J\}$, to the receiver. The pilot signal is precoded via a transmit precoding vector $\boldsymbol{f}_{t,j}$ that changes over different time frames and different time slots. At the jth time slot of the tth time frame, the signal received at the user can be expressed as

$$\begin{aligned} y_{t,j} &= \boldsymbol{h}_t^H \boldsymbol{\Psi}_{t,j} \boldsymbol{G}_t \boldsymbol{f}_{t,j} + n_{t,j} \\ &= \boldsymbol{w}_{t,j} \mathrm{vec}\left(\boldsymbol{H}_t\right) + n_{t,j} \end{aligned} \tag{10}$$

where $\boldsymbol{h}_t$ and $\boldsymbol{G}_t$ respectively represent the IRS-user channel and the BS-IRS channel at the tth time frame, $\boldsymbol{\Psi}_{t,j}$ is the phase shift matrix that is employed at the jth time slot of the tth time frame, $n_{t,j}$ is the additive white Gaussian noise, $\boldsymbol{H}_t \triangleq \mathrm{diag}(\boldsymbol{h}_t^H)\boldsymbol{G}_t$ is referred to as the cascade channel at the tth frame, and $\boldsymbol{w}_{t,j} \triangleq \boldsymbol{f}_{t,j}^T \otimes \boldsymbol{\psi}_{t,j}^T$, in which $\boldsymbol{\psi}_{t,j} = \mathrm{diag}(\boldsymbol{\Psi}_{t,j})$.

Define $\boldsymbol{y}_t \triangleq [y_{t,1} \cdots y_{t,J}]^T$, $\boldsymbol{n}_t \triangleq [n_{t,1} \cdots n_{t,J}]^T$ and $\boldsymbol{W}_t \triangleq [\boldsymbol{w}_{t,1}^T \ \cdots \ \boldsymbol{w}_{t,J}^T]^T$. To facilitate CCM estimation, we employ the same $\boldsymbol{W} = \boldsymbol{W}_t, \forall t$ for each time frame. Thus

the received signal at the tth frame can be written in a matrix form as

$$\boldsymbol{y}_t = \boldsymbol{W}\mathrm{vec}\left(\boldsymbol{H}_t\right) + \boldsymbol{n}_t = \boldsymbol{W}\bar{\boldsymbol{h}}_t + \boldsymbol{n}_t \tag{11}$$

where $\bar{\boldsymbol{h}}_t \triangleq \mathrm{vec}\left(\boldsymbol{H}_t\right)$.

As reported by previous studies [25], [26], an important fact about mmWave channels is that the angle parameters such as the AoA and AoD depend only on the relative positions of the BS, the user, and the scatterers, which vary much more slowly than I-CSI. Also, the statistics of the complex path gain keep invariant over an interval that is much longer than the channel coherent time. Hence it is reasonable to make the following basic assumption:

A1 The angle parameters and the variances of the path gains, remains unchanged over a long-term period.

Our objective is to obtain an estimate of $\boldsymbol{R}_h \triangleq \mathbb{E}[\bar{\boldsymbol{h}}_t\bar{\boldsymbol{h}}_t^H]$ from the received signal $\{\boldsymbol{y}_t\}_{t=1}^T$. Note that the CCM $\boldsymbol{R}_h$ has a dimension of $NM \times NM$.

A. Discussions

Although this paper considers CCM estimation of the downlink channel, the proposed method can be readily applied to the uplink's CCM estimation. Similar to the downlink training, the uplink training contains T time frames and each time frame consists of J time slots. During the training phase, the user sends a same pilot signal $s_{t,j} = 1$ to the BS, where the received signal is combined via a combing vector $\boldsymbol{g}_{t,j}$. The channel in the same time frame is assumed to be time-invariant. Therefore the signal received at the jth time slot of the tth time frame is given as

$$\begin{aligned} y_{t,j}^u &= \boldsymbol{g}_{t,j}^H\boldsymbol{G}_t^H\boldsymbol{\Psi}_{t,j}\boldsymbol{h}_t s_{t,j} + n_{t,j}^u \\ &= \boldsymbol{w}_{t,j}^u\mathrm{vec}(\boldsymbol{H}_t^u) + n_{t,j}^u \end{aligned} \tag{12}$$

where $\boldsymbol{G}_t$ and $\boldsymbol{h}_t$ are, respectively, the IRS-BS channel and the user-IRS channel, $\boldsymbol{w}_{t,j}^u \triangleq \boldsymbol{\psi}_{t,j}^T \otimes \boldsymbol{g}_{t,j}^H$ with $\boldsymbol{\psi}_{t,j} = \mathrm{diag}(\boldsymbol{\Psi}_{t,j})$ and $\boldsymbol{H}_t^u = \boldsymbol{G}_t^H\mathrm{diag}(\boldsymbol{h}_t)$. Define $\boldsymbol{y}_t^u \triangleq [y_{t,1}^u \ \cdots \ y_{t,J}^u]^T$, $\boldsymbol{n}_t^u \triangleq [n_{t,1}^u \ \cdots \ n_{t,J}^u]^T$, and $\boldsymbol{W}_t^u \triangleq [(\boldsymbol{w}_{t,1}^u)^T \ \cdots \ (\boldsymbol{w}_{t,J}^u)^T]$. The received signal at the BS can be expressed as

$$\boldsymbol{y}_t^u = \boldsymbol{W}_t^u\mathrm{vec}(\boldsymbol{H}_t^u) + \boldsymbol{n}_t^u \tag{13}$$

If we set $\boldsymbol{W}_t^u$ the same for different time frames, the signal model for the uplink training has a same form as that for the downlink scenario. For time division duplex (TDD) systems, due to the channel reciprocity between opposite links (downlink and uplink), the estimated uplink CCM can be used for downlink precoding/beamforming.

IV. Spatial Channel Covariance Matrix Estimation

According to (11), we have

$$\boldsymbol{R}_y = \mathbb{E}\{\boldsymbol{y}_t\boldsymbol{y}_t^H\} = \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H + \sigma^2\boldsymbol{I} \tag{14}$$

Generally the true covariance matrix $\boldsymbol{R}_y$ is unavailable. But it can be estimated via the following sample covariance matrix

$$\hat{\boldsymbol{R}}_y = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t\boldsymbol{y}_t^H \tag{15}$$

Intuitively, one can directly estimate $\boldsymbol{R}_h$ from $\hat{\boldsymbol{R}}_{\boldsymbol{y}}$ via solving the following least squares problem

$$\mathrm{vec}(\hat{\boldsymbol{R}}_{\boldsymbol{y}}) = (\boldsymbol{W}^* \otimes \boldsymbol{W})\,\mathrm{vec}(\boldsymbol{R}_h) + \mathrm{vec}(\sigma^2\boldsymbol{I}) \tag{16}$$

provided that the dimension of $\hat{\boldsymbol{R}}_{\boldsymbol{y}}$ (i.e. $J \times J$) is larger than the dimension of $\boldsymbol{R}_h$ (i.e. $MN \times MN$). Nevertheless, the condition $J \geq NM$ is unlikely to be satisfied in practice since the coherence time in mmWave systems is relatively small. Therefore, estimating $\boldsymbol{R}_h$ from $\hat{\boldsymbol{R}}_{\boldsymbol{y}}$ is in fact an underdetermined problem, and in order to handle such an issue we have to exploit the structure of $\boldsymbol{R}_h$.

A. Exploiting The Structure of $\boldsymbol{R}_h$

To exploit the structure of $\boldsymbol{R}_h$, we first obtain a sparse representation of the cascade channel $\boldsymbol{H}_t$. For notational convenience, in the following we will omit the subscript t in $\boldsymbol{H}_t$, $\boldsymbol{h}_t$ and $\boldsymbol{G}_t$. Utilizing the matrix properties, the cascade channel $\boldsymbol{H}$ can be expressed as

$$\begin{aligned} \boldsymbol{H} &= \mathrm{diag}(\boldsymbol{h}^H)\boldsymbol{G} = \boldsymbol{h}^* \bullet \boldsymbol{G} \\ &= (\boldsymbol{F}_3^*\boldsymbol{\beta}^*) \bullet (\boldsymbol{F}_1\boldsymbol{\Sigma}\boldsymbol{F}_2^H) \\ &= (\boldsymbol{F}_3^* \bullet \boldsymbol{F}_1)\left(\boldsymbol{\beta}^* \otimes \left(\boldsymbol{\Sigma}\boldsymbol{F}_2^H\right)\right) \\ &= (\boldsymbol{F}_3^* \bullet \boldsymbol{F}_1)\left(\boldsymbol{\beta}^* \otimes \boldsymbol{\Sigma}\right)\boldsymbol{F}_2^H \\ &\triangleq \boldsymbol{F}_4\boldsymbol{\Pi}\boldsymbol{F}_2^H \end{aligned} \tag{17}$$

where

$$\boldsymbol{F}_4 \triangleq \boldsymbol{F}_3^* \bullet \boldsymbol{F}_1 \tag{18}$$

$$\boldsymbol{\Pi} \triangleq \boldsymbol{\beta}^* \otimes \boldsymbol{\Sigma} \tag{19}$$

Note that the ςth ($\varsigma = (p-1)L + l$) column of $\boldsymbol{F}_4 \in \mathbb{C}^{M \times PL}$ is given by

$$\begin{aligned} &\boldsymbol{a}_r^*(\psi_{v,p}, \psi_{h,p}) \circ \boldsymbol{a}_r(\theta_{v,l}, \theta_{h,l}) \\ &\overset{(a)}{=} (\boldsymbol{a}^*(\nu_{4,p}, M_v) \circ \boldsymbol{a}(\nu_{2,l}, M_v)) \otimes (\boldsymbol{a}^*(\nu_{5,p}, M_h) \circ \boldsymbol{a}(\nu_{3,l}, M_h)) \\ &\triangleq \boldsymbol{a}(\nu_{6,\varsigma}, M_v) \otimes \boldsymbol{a}(\nu_{7,\varsigma}, M_h) \end{aligned} \tag{20}$$

where $\nu_{6,\varsigma} \triangleq \nu_{2,l} - \nu_{4,p}$, $\nu_{7,\varsigma} \triangleq \nu_{3,l} - \nu_{5,p}$, and (a) follows from the property: $(\boldsymbol{A} \otimes \boldsymbol{B}) \circ (\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A} \circ \boldsymbol{C}) \otimes (\boldsymbol{B} \circ \boldsymbol{D})$.

Vectorizing the cascade channel in (17) leads to

$$\mathrm{vec}(\boldsymbol{H}) = \mathrm{vec}(\boldsymbol{F}_4\boldsymbol{\Pi}\boldsymbol{F}_2^H) \triangleq \boldsymbol{F}\boldsymbol{x} \tag{21}$$

where $\boldsymbol{F} \triangleq \boldsymbol{F}_2^* \otimes \boldsymbol{F}_4 \in \mathbb{C}^{MN \times L^2P}$ and $\boldsymbol{x} \triangleq \mathrm{vec}(\boldsymbol{\Pi}) \in \mathbb{C}^{L^2P}$. $\boldsymbol{F}$ can be further expressed as

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{F}_2^* \otimes \boldsymbol{F}_4 \\ &= [\boldsymbol{a}^*(\nu_{1,1}, N) \ \cdots \ \boldsymbol{a}^*(\nu_{1,L}, N)] \\ &\quad \otimes \Big[\boldsymbol{a}(\nu_{6,1}, M_v) \otimes \boldsymbol{a}(\nu_{7,1}, M_h) \ \cdots \\ &\qquad \boldsymbol{a}(\nu_{6,LP}, M_v) \otimes \boldsymbol{a}(\nu_{7,LP}, M_h)\Big] \end{aligned} \tag{22}$$

Substituting (22) into (21), we have

$$\begin{aligned} \bar{\boldsymbol{h}} &= \mathrm{vec}(\boldsymbol{H}) \\ &= \sum_{l=1}^{L}\sum_{\zeta=1}^{LP} x_{\varrho}\boldsymbol{a}^*(\nu_{1,l}, N) \otimes \boldsymbol{a}(\nu_{6,\zeta}, M_v) \otimes \boldsymbol{a}(\nu_{7,\zeta}, M_h) \end{aligned} \tag{23}$$

where x_ϱ with $\varrho = \zeta + (l-1)LP$ is the ϱth element of $\boldsymbol{x}$. Due to the fact that $\boldsymbol{\Pi} = \boldsymbol{\beta}^* \otimes \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix, there are at most LP non-zeros elements in $\boldsymbol{x}$, and each element in $\boldsymbol{x}$ is given by

$$x_\varrho = \begin{cases} \alpha_l \beta_p^*, & \varrho \in \boldsymbol{O} \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

where the set $\boldsymbol{O}$ is defined as

$$\Big\{(p-1)L + l + (l-1)LP | l \in \{1, \cdots, L\}; p \in \{1, \cdots, P\}\Big\}$$

Recall that $\alpha_l \sim \mathbb{CN}(0, \varpi_l^2)$ and $\beta_p \sim \mathbb{CN}(0, \chi_p^2)$, and α_l and β_p are mutually uncorrelated. Therefore for $x_\varrho = \alpha_l \beta_p^*$, its mean and variance are given as

$$\begin{aligned} \mathbb{E}[x_\varrho] &= 0 \\ \mathbb{E}[x_\varrho x_\varrho^*] &= \eta_\varrho^2 \triangleq \varpi_l^2 \chi_p^2 \end{aligned} \tag{25}$$

We see that $\bar{\boldsymbol{h}}$ can be characterized by a geometric channel model. It has LP uncorrelated composite paths in total. The complex gain of each composite path is a random variable with zero mean and finite variance. The angular parameters $(\{\nu_{1,l}, \nu_{6,\zeta}, \nu_{7,\zeta}\})$ associated with each path are treated as deterministic parameters as angle parameters vary slowly relative to the complex path gains. Hence the channel covariance matrix $\boldsymbol{R}_h$ can be expressed as

$$\begin{aligned} \boldsymbol{R}_h &= \mathbb{E}(\bar{\boldsymbol{h}}\bar{\boldsymbol{h}}^H) = \boldsymbol{F}\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^H)\boldsymbol{F}^H \\ &\overset{(a)}{=} \sum_{\varrho=1}^{L^2P} \mathbb{E}(x_\varrho x_\varrho^*) \boldsymbol{R}_\varrho \\ &= \sum_{\varrho \in \boldsymbol{O}} \eta_\varrho^2 \boldsymbol{R}_\varrho \end{aligned} \tag{26}$$

where (a) follows from the fact that $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^H)$ is a diagonal matrix, and $\boldsymbol{R}_\varrho \in \mathbb{C}^{NM \times NM}$ is defined as

$$\boldsymbol{R}_\varrho = \big(\boldsymbol{a}^*(\nu_{1,l}, N)\boldsymbol{a}^T(\nu_{1,l}, N)\big) \otimes \big(\boldsymbol{a}(\nu_{6,\zeta}, M_v)\boldsymbol{a}^H(\nu_{6,\zeta}, M_v)\big) \otimes \big(\boldsymbol{a}(\nu_{7,\zeta}, M_h)\boldsymbol{a}^H(\nu_{7,\zeta}, M_h)\big) \tag{27}$$

It can be easily verified that $\boldsymbol{R}_\varrho$ is a PSD 3-level Toeplitz matrix. As a result, $\boldsymbol{R}_h \in \mathbb{C}^{NM \times NM}$ is also a PSD 3-level Toeplitz matrix. Although there are N^2M^2 elements in $\boldsymbol{R}_h$, owing to the specific structure of PSD 3-level Toeplitz matrix, $\boldsymbol{R}_h$ can be characterized by $(2N-1)(2M_v-1)(2M_h-1)$ parameters which can be represented by a third-order tensor $\boldsymbol{V} \in \mathbb{C}^{(2N-1)\times(2M_v-1)\times(2M_h-1)}$, i.e. we can write $\boldsymbol{R}_h = \mathbb{T}_3(\boldsymbol{V})$. How to map a third-order tensor to a 3-level Toeplitz matrix can be found in [27]. Furthermore, from (26), we know that $\boldsymbol{R}_h$ can be represented by a summation of LP rank-one matrices. Due to the sparse scattering characteristics of mmWave channels, LP is usually much smaller than the dimension of $\boldsymbol{R}_h$ (i.e., NM), meaning that $\boldsymbol{R}_h$ has a low-rank structure.

Utilizing the PSD 3-level Toeplitz structure and the low-rank property of $\mathbb{T}_3(\boldsymbol{V})$, the estimation of $\boldsymbol{R}_h$ can be cast into the following low-rank structured covariance reconstruction problem:

$$\begin{aligned} \hat{\boldsymbol{R}}_h = \arg\min_{\mathbb{T}_3(\boldsymbol{V})} \quad & \frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\mathbb{T}_3(\boldsymbol{V})\boldsymbol{W}^H\right\|_F^2 + \lambda \text{rank}(\mathbb{T}_3(\boldsymbol{V})) \\ \text{s.t.} \quad & \mathbb{T}_3(\boldsymbol{V}) \succcurlyeq 0 \end{aligned} \tag{28}$$

where λ is a regularization parameter to balance the tradeoff between data fitting and low-rankness. Nonetheless, such a problem is generally NP-hard due to the rank function. To make it tractable, we resort to convex relaxation to replace $\text{rank}(\mathbb{T}_3(\boldsymbol{V}))$ with the nuclear-norm of $\mathbb{T}_3(\boldsymbol{V})$. Since $\mathbb{T}_3(\boldsymbol{V})$ is confined to be a PSD matrix, its nuclear norm is equivalent to its trace. Consequently, the resulting optimization can be given by

$$\begin{aligned} \hat{\boldsymbol{R}}_h = \arg\min_{\mathbb{T}_3(\boldsymbol{V})} \quad & \frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\mathbb{T}_3(\boldsymbol{V})\boldsymbol{W}^H\right\|_F^2 + \lambda \text{tr}(\mathbb{T}_3(\boldsymbol{V})) \\ \text{s.t.} \quad & \mathbb{T}_3(\boldsymbol{V}) \succcurlyeq 0 \end{aligned} \tag{29}$$

The above optimization is a convex semidefinite programming (SDP) problem which can be solved by many standard off-the-shelf solvers, e.g., CVX. Unfortunately, these solvers are usually computationally expensive. To reduce the computational complexity, we develop an alternating direction method of multipliers (ADMM) algorithm for solving (29) in the next section.

B. ADMM-Based Algorithm

To solve (29), we first introduce two auxiliary variables, $\boldsymbol{A}$ and $\boldsymbol{B}$, and reformulate (29) into the following optimization

$$\begin{aligned} \{\hat{\boldsymbol{V}}, \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\} = \arg\min_{\boldsymbol{V},\boldsymbol{A},\boldsymbol{B}} \quad & \Big(\frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{A}\boldsymbol{W}^H\right\|_F^2 + \lambda \text{tr}(\boldsymbol{A}) + \mathbb{I}_\infty(\boldsymbol{B} \succcurlyeq 0 \\ \text{s.t.} \quad & \boldsymbol{A} = \mathbb{T}_3(\boldsymbol{V}),\ \boldsymbol{B} = \boldsymbol{A}, \end{aligned} \tag{30}$$

where $\mathbb{I}_\infty(a)$ is an indicator function defined as

$$\mathbb{I}_\infty(a) = \begin{cases} 0, & \text{if } a \text{ is true} \\ \infty, & \text{otherwise} \end{cases} \tag{31}$$

The augmented Lagrangian of the above optimization is read as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Upsilon}, \boldsymbol{\Lambda}) \\ &= \frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{A}\boldsymbol{W}^H\right\|_F^2 + \lambda \text{tr}(\boldsymbol{A}) \\ &\quad + \langle \boldsymbol{\Upsilon}, \boldsymbol{A} - \mathbb{T}_3(\boldsymbol{V})\rangle + \frac{\eta}{2}\|\boldsymbol{A} - \mathbb{T}_3(\boldsymbol{V})\|_F^2 \\ &\quad + \langle \boldsymbol{\Lambda}, \boldsymbol{B} - \boldsymbol{A}\rangle + \frac{\rho}{2}\|\boldsymbol{B} - \boldsymbol{A}\|_F^2 + \mathbb{I}_\infty(\boldsymbol{B} \succcurlyeq 0) \end{aligned} \tag{32}$$

where $\langle \boldsymbol{A}, \boldsymbol{B}\rangle$ is defined as $\text{Re}\big(\text{Tr}(\boldsymbol{B}^H\boldsymbol{A})\big)$, $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Lambda}$ are the dual parameters, and $\eta, \rho > 0$ are the penalty parameters. According to the updating rule of the ADMM algorithm, it consists of solving the following sub-problems:

$$\hat{\boldsymbol{A}}^{k+1} = \arg\min_{\boldsymbol{A}} \ \mathcal{L}(\boldsymbol{V}^k, \boldsymbol{A}, \boldsymbol{B}^k, \boldsymbol{\Upsilon}^k, \boldsymbol{\Lambda}^k) \tag{33}$$

$$\hat{\boldsymbol{V}}^{k+1} = \arg\min_{\boldsymbol{V}} \ \mathcal{L}(\boldsymbol{V}, \boldsymbol{A}^{k+1}, \boldsymbol{B}^k, \boldsymbol{\Upsilon}^k, \boldsymbol{\Lambda}^k) \tag{34}$$

$$\hat{\boldsymbol{B}}^{k+1} = \arg\min_{\boldsymbol{B}} \ \mathcal{L}(\boldsymbol{V}^{k+1}, \boldsymbol{A}^{k+1}, \boldsymbol{B}, \boldsymbol{\Upsilon}^k, \boldsymbol{\Lambda}^k) \tag{35}$$

$$\boldsymbol{\Upsilon}^{k+1} = \boldsymbol{\Upsilon}^k + \eta\big(\boldsymbol{A}^{k+1} - \mathbb{T}_3(\boldsymbol{V}^{k+1})\big) \tag{36}$$

$$\boldsymbol{\Lambda}^{k+1} = \boldsymbol{\Lambda}^k + \rho\big(\boldsymbol{B}^{k+1} - \boldsymbol{A}^{k+1}\big) \tag{37}$$

Update of $\boldsymbol{A}$: We first solve the sub-problem (33). Calculating the derivative of $\mathcal{L}(\boldsymbol{V}^k, \boldsymbol{A}, \boldsymbol{B}^k, \boldsymbol{\Upsilon}^k, \boldsymbol{\Lambda}^k)$ with respect to

$\boldsymbol{A}$, we have

$$\nabla_{\boldsymbol{A}}\mathcal{L} = (\boldsymbol{W}^H\boldsymbol{W})\boldsymbol{A}(\boldsymbol{W}^H\boldsymbol{W})^H + (\eta-\rho)\boldsymbol{A} - \eta\mathbb{T}_3(\boldsymbol{V}^k) + \rho\boldsymbol{B}^k - \boldsymbol{W}^H\hat{\boldsymbol{R}}_y\boldsymbol{W} + \boldsymbol{\Upsilon}^k - \boldsymbol{\Lambda}^k + \lambda\boldsymbol{I} \quad (38)$$

By setting $\nabla_{\boldsymbol{A}}\mathcal{L}$ to 0, we can update $\boldsymbol{A}$ via solving the following linear equation

$$\boldsymbol{\Xi}\boldsymbol{A}\boldsymbol{\Xi}^H + \kappa\boldsymbol{A} = \boldsymbol{C} \quad (39)$$

where $\boldsymbol{C} = \eta\mathbb{T}_3(\boldsymbol{V}^k) - \rho\boldsymbol{B}^k + \boldsymbol{W}^H\hat{\boldsymbol{R}}_y\boldsymbol{W} - \lambda\boldsymbol{I} - \boldsymbol{\Upsilon}^k + \boldsymbol{\Lambda}^k$, $\kappa = \eta - \rho$ and $\boldsymbol{\Xi} = \boldsymbol{W}^H\boldsymbol{W}$. Clearly, $\boldsymbol{A}$ can be solved via

$$(\boldsymbol{\Xi}^* \otimes \boldsymbol{\Xi} + \kappa\boldsymbol{I})\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{C}) \quad (40)$$

Nevertheless, this approach has a computational complexity of $\mathcal{O}(N^6M^6)$. To reduce the computational complexity, we propose a more computationally-efficient approach to solve (39). Since $\boldsymbol{\Xi} = \boldsymbol{W}^H\boldsymbol{W}$, $\boldsymbol{\Xi}$ can be diagonalized via the eigen-decomposition (EVD), i.e.,

$$\boldsymbol{\Xi} = \boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_1^H \quad (41)$$

where $\boldsymbol{A}_2$ is a diagonal matrix and $\boldsymbol{A}_1$ is a unitary matrix (meaning $\boldsymbol{A}_1^H = \boldsymbol{A}_1^{-1}$). Substituting (41) into (39) results in

$$\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_1^H\boldsymbol{A}(\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_1^H)^H + \kappa\boldsymbol{A} = \boldsymbol{C} \quad (42)$$

By defining $\boldsymbol{A}_3 \triangleq \boldsymbol{A}_1^H\boldsymbol{A}\boldsymbol{A}_1$ and $\boldsymbol{C}_1 \triangleq \boldsymbol{A}_1^H\boldsymbol{C}\boldsymbol{A}_1$, we can express (42) as

$$\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_2 + \kappa\boldsymbol{A}_3 = \boldsymbol{C}_1 \quad (43)$$

Since $\boldsymbol{A}_2$ is a diagonal matrix, we can explicitly solve (43) in an elementwise manner. Once $\boldsymbol{A}_3$ is obtained, $\boldsymbol{A}$ can be simply reconstructed as

$$\boldsymbol{A} = \boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_1^H \quad (44)$$

Update of $\boldsymbol{V}$: Keeping the terms that only depend on $\boldsymbol{V}$ in (34), we have

$$\hat{\boldsymbol{V}}^{k+1} = \arg\min_{\boldsymbol{V}} \ \frac{\eta}{2}\text{Tr}\left(\mathbb{T}_3^H(\boldsymbol{V})\mathbb{T}_3(\boldsymbol{V})\right) - \text{Re}\left(\text{Tr}\big((\boldsymbol{\Upsilon}^k + \eta\boldsymbol{A}^{k+1})^H\mathbb{T}_3(\boldsymbol{V})\big)\right) \quad (45)$$

Directly taking derivative with respect to $\boldsymbol{V}$ is difficult. Nevertheless, the elements in $\boldsymbol{V}$ can be optimized separately. To this end, we reshape $\boldsymbol{V} \in \mathbb{C}^{(2N-1)\times(2M_v-1)\times(2M_h-1)}$ to a vector $\boldsymbol{v} \in \mathbb{C}^{(2N-1)(2M_v-1)(2M_h-1)}$. Also, we define an index set $\mathbb{I}_{v_l}$ as

$$\mathbb{I}_{v_l} = \{(i,j)|\mathbb{T}_3(\boldsymbol{V})(i,j) \equiv v_l\} \quad (46)$$

for all $l \in \{1,\cdots,(2N-1)(2M_v-1)(2M_h-1)\}$. The cardinality of the set $\mathbb{I}_{v_l}$ is denoted by $|\mathbb{I}_{v_l}|$. Based on this definition, the terms that only depend on v_l in the cost function (45) can be rewritten as

$$\hat{v}_l^{k+1} = \arg\min_{v_l} \frac{\eta|\mathbb{I}_{v_l}|}{2}v_lv_l^* - \left(\sum_{(i,j)\in\mathbb{I}_{v_l}} \text{Re}\left(\Delta^*(i,j)v_l\right)\right) \quad (47)$$

where $\Delta \triangleq \boldsymbol{\Upsilon}^k + \eta\boldsymbol{A}^{k+1}$. It is clear that $\hat{v}_l^{k+1}$ is given by

$$\hat{v}_l^{k+1} = \frac{1}{\eta|\mathbb{I}_{v_l}|}\left(\sum_{(i,j)\in\mathbb{I}_{v_l}} \Delta(i,j)\right) \quad (48)$$

Update of $\boldsymbol{B}$: The optimization in (35) is equivalent to

$$\hat{\boldsymbol{B}}^{k+1} = \arg\min_{\boldsymbol{B}} \ \frac{\rho}{2}\|\boldsymbol{B} - (\boldsymbol{A}^{k+1} - \boldsymbol{\Lambda}^k/\rho)\|_F^2 \quad \text{s.t.} \quad \boldsymbol{B} \succcurlyeq 0 \quad (49)$$

Apparently, the solution to the above optimization is obtained by projecting $\tilde{\boldsymbol{B}} \triangleq \boldsymbol{A}^{k+1} - \boldsymbol{\Lambda}^k/\rho$ onto the positive semi-definite cone, which is equivalent to setting all negative eigenvalues of $\tilde{\boldsymbol{B}}$ to zero.

Algorithm 1 Proposed ADMM Algorithm

Initialization: $\boldsymbol{V}^0, \boldsymbol{B}^0, \boldsymbol{\Upsilon}^0, \boldsymbol{\Lambda}^0, \eta, \rho, k = 1$.
repeat
1. Update $\boldsymbol{A}^k$ via (44);
2. Update each element in $\boldsymbol{V}^k$ via (48);
3. Update $\boldsymbol{B}^k$ via solving (49);
4. Update $\boldsymbol{\Upsilon}^k$ and $\boldsymbol{\Lambda}^k$ via (36) and (37) respectively;
5. $k = k + 1$

until Convergence

For clarity, the proposed ADMM algorithm is summarized in Algorithm 1. Compared with the off-the-shelf CVX solver, the proposed ADMM algorithm is more computationally efficient. Specifically, for the CVX solver, since it is based on the interior point method, the overall computational complexity of solving the optimization problem (29) is at the order of $\mathcal{O}((NM)^{3.5}\log(1/\epsilon))$, where ϵ is the desired recovery precision [28]. For the proposed ADMM algorithm, its computational complexity is dominated by two eigenvalue decompositions (EVD), i.e., the EVD of $\boldsymbol{\Xi} \in \mathbb{C}^{NM\times NM}$ (i.e., (41)) and the EVD of $\tilde{\boldsymbol{B}} \in \mathbb{C}^{NM\times NM}$ when projecting $\tilde{\boldsymbol{B}}$ onto its semi-definite cone. The former EVD can be computed offline since $\boldsymbol{\Xi} = \boldsymbol{W}^H\boldsymbol{W}$ can be fixed over the entire training process. The EVD of $\tilde{\boldsymbol{B}}$ involves a complexity of $\mathcal{O}((NM)^3)$. Nevertheless, notice that $\boldsymbol{B}$ has a low-rank property. Therefore, one can employ the truncated EVD to further reduce the computational complexity. Generally, the truncated EVD has a computational complexity of order $7NM\bar{k}^2$, where $\bar{k}$ is the number of the required eigenvectors which in general can be set to 2 or 3 times of r (i.e., the rank of $\boldsymbol{B}$). Hence, the overall computational complexity of the proposed ADMM algorithm at each iteration is $\mathcal{O}(NM)$, which is far less than the complexity $\mathcal{O}((NM)^{3.5}\log(1/\epsilon))$ required by the CVX solver.

C. Extensions

The proposed method can be extended to scenarios where there exists a direct link between the BS and the user. In such a case, a two-step estimation scheme can be employed to estimate the cascade CCM. Specifically, at the first step, we turn off the IRS and calculate the sample covariance matrix of the received signal, say $\hat{\boldsymbol{R}}^{(1)}$. Clearly, $\hat{\boldsymbol{R}}^{(1)}$ is a compressed form of the direct channel's CCM. Then in the second step, we turn on the IRS and calculate a new sample sample covariance matrix, say $\hat{\boldsymbol{R}}^{(2)}$. Due to the statistical independence between the direct channel and the cascade channel, we have $\hat{\boldsymbol{R}}^{(2)} =$

$\hat{\boldsymbol{R}}^{(1)} + \hat{\boldsymbol{R}}_y$, with $\hat{\boldsymbol{R}}_y$ denoting the sample covariance matrix given by (14). Thus, the CCM of the cascade channel $\boldsymbol{R}_h$ can be estimated from $\hat{\boldsymbol{R}}^{(2)} - \hat{\boldsymbol{R}}^{(1)}$ via our proposed method.

V. Performance Analysis of The CCM Estimator

In this section, we analyze the estimation performance of the CCM estimator (29). Specifically, we are interested in quantifying the amount of training overhead required to achieve a reliable estimate of the true CCM $\boldsymbol{R}_h$. To facilitate our analysis, we consider the noise-free case, i.e., $\sigma^2 = 0$ and $\boldsymbol{R}_y = \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H$. In the following we use $\boldsymbol{R}_h$ and $\mathbb{T}_3(\boldsymbol{V})$ interchangeably since these two essentially have the same meaning.

Suppose $\mathbb{T}_3(\boldsymbol{X}) \in \mathbb{C}^{I_1 I_2 I_3 \times I_1 I_2 I_3}$ is a 3-level Toeplitz matrix parameterized by a tensor $\boldsymbol{X} \in \mathbb{C}^{(2I_1-1)\times(2I_2-1)\times(2I_3-1)}$. For any matrix $\boldsymbol{M} \in \mathbb{C}^{I \times I_1 I_2 I_3}$, $\check{\boldsymbol{M}} \in \mathbb{C}^{I^2 \times (2I_1-1)(2I_2-1)(2I_3-1)}$ is referred to as the transforming matrix of $\boldsymbol{M}$ if it satisfies

$$\check{\boldsymbol{M}}\boldsymbol{x} = \mathrm{vec}(\boldsymbol{M}\mathbb{T}_3(\boldsymbol{X})\boldsymbol{M}^H) \tag{50}$$

where $\boldsymbol{x}$ is a vector by reshaping $\boldsymbol{X}$ into a vector. In addition, define $r_e(\boldsymbol{A}) \triangleq \frac{\mathrm{Tr}(\boldsymbol{A})}{\|\boldsymbol{A}\|_2}$ as the effective rank of matrix $\boldsymbol{A}$. Our result is summarized as follows.

Theorem 1: Let $\boldsymbol{V} \in \mathbb{C}^{(2N-1)\times(2M_v-1)\times(2M_h-1)}$ be the ground truth and $\hat{\boldsymbol{V}} \in \mathbb{C}^{(2N-1)\times(2M_v-1)\times(2M_h-1)}$ be the solution of (29). Given observations $\{\boldsymbol{y}_t\}_{t=1}^T$, and set

$$J \geq u \triangleq \sqrt{(2N-1)(2M_v-1)(2M_h-1)} \tag{51}$$

and

$$\lambda \geq c\|\boldsymbol{W}\|_F^2\|\boldsymbol{R}_y\|_2 \max\{\sqrt{\tilde{\delta}}, \tilde{\delta}\} \tag{52}$$

where c is a constant and $\tilde{\delta}$ is defined as

$$\tilde{\delta} \triangleq \frac{r_e(\boldsymbol{R}_y)\log(TJ)}{T} \tag{53}$$

then with probability at least $1 - 4T^{-1}$, the average per-entry root mean square error (RMSE) of the solution to (29) satisfies

$$\frac{1}{u}\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F \leq \frac{16\lambda\sqrt{r}}{\sigma_{\min}^2(\check{\boldsymbol{W}})}\frac{\sqrt{NM}}{u} \tag{54}$$

where r is the rank of $\boldsymbol{R}_h$, $\check{\boldsymbol{W}}$ is the transforming matrix of $\boldsymbol{W}$, and $\sigma_{\min}(\check{\boldsymbol{W}})$ denotes the smallest singular value of $\check{\boldsymbol{W}}$.

Proof: See Appendix I. ∎

The above theorem is a generalization of Theorem 4 in [23]. Specifically, [23] analyzes the structured covariance estimation performance under a partial observation framework where the aim is to recover the complete CCM from a submatrix of the sample CCM. Theorem 1 generalizes this partial observation model to an arbitrarily linear compression model in which we do not have direct access to the entries of the sample CCM; instead, only the sample covariance matrix $\hat{\boldsymbol{R}}_y$ is available. Also, the CCM in this work has a multi-level Toeplitz structure, which is different from that of [23].

Recalling $M = M_v M_h$, the term $\sqrt{NM}/u$ in (54) tends to be a constant for sufficiently large values of M and N. Therefore, the average per-entry RMSE is upper bounded by $\lambda\sqrt{r}$ times a scale factor. According to (52), we know that

$$\lambda\sqrt{r} \geq \bar{u} \triangleq c\sqrt{r}\|\boldsymbol{W}\|_F^2\|\boldsymbol{R}_y\|_2 \max\{\sqrt{\tilde{\delta}}, \tilde{\delta}\} \tag{55}$$

Therefore, in our optimization problem (29), we set $\lambda = \bar{u}/\sqrt{r}$ such that the average pre-entry RMSE of $\boldsymbol{V}$ has the smallest upper bound, i.e.,

$$\begin{aligned}\frac{1}{u}\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F &\leq \frac{\sqrt{NM}}{u}\frac{16\bar{u}}{\sigma_{\min}^2(\check{\boldsymbol{W}})} \\ &= \frac{\sqrt{NM}}{u}\frac{16c\|\boldsymbol{W}\|_F^2\|\boldsymbol{R}_y\|_2 \max\{\sqrt{r\tilde{\delta}}, \sqrt{r}\tilde{\delta}\}}{\sigma_{\min}^2(\check{\boldsymbol{W}})}\end{aligned} \tag{56}$$

From (56), we can see that the average pre-entry RMSE vanishes as long as $r\tilde{\delta}$ tends to 0. To this objective, it can be verified that the number of time frames T should be in the order of $r_e(\boldsymbol{R}_y) r \log(J)$ or in the order of $r^2\log(J)$ since the effective rank of a matrix is no greater than its true rank, i.e. $r_e(\boldsymbol{R}_y) \leq r$ [29]. To see why $r\tilde{\delta}$ tends to 0 when $T \sim \mathcal{O}(r^2\log(J))$, let $T = \tau r^2 \log(J)$, where τ is a constant. In this case, we have

$$r\tilde{\delta} \leq \frac{\log(\tau r^2 \log(J))}{\tau \log(J)} \tag{57}$$

where the right-hand side of the above inequality decreases to a small value as τ increases. In summary, when the number of time frames T is in the order of $r^2\log(J)$, the average per-entry RMSE can be upper bounded by an arbitrarily small value, which means that we can obtain a reliable estimate of the true CCM $\boldsymbol{R}_h$. Note that our performance guarantee is non-asymptotic and holds for a finite number of measurement vectors. In other words, the result has accounted for the covariance estimation error due to finite samples.

Recall that the proposed downlink training protocol consists of T time frames, and each time frame comprises J time slots. Thus the total amount of training overhead is TJ. Since the number of time slots J has to satisfy (51), which means that J is on the order of $\sqrt{NM}$. Therefore, the total amount of training overhead is at the order of $r^2\sqrt{NM}\log(NM)$. Note that here $r = PL$ is the rank of $\boldsymbol{R}_h$. In mmWave systems, due to sparse scattering characteristics, both P and L are relatively small. Hence r is generally far less than NM. Based on this, we can conclude that when the total number of training symbols is in the order of $\sqrt{NM}\log NM$, we can provide a reliable estimate of the CCM $\boldsymbol{R}_h$. Note that $\sqrt{NM}\log NM$ is much smaller than the dimension of $\boldsymbol{R}_h$ (i.e., N^2M^2).

VI. Two-Timescale Beamforming Based on Estimated CCM

We discuss how to perform two-timescale beamforming based on the estimated CCM. In the following, we first consider single-user scenarios and then discuss the extension to the multi-user scenarios.

A. Single-User Scenarios

The received signal at the user can be expressed as

$$y_t = \boldsymbol{h}^H\boldsymbol{\Psi}\boldsymbol{G}\boldsymbol{f}s_t + n_t \tag{58}$$

where s_t is the transmitted symbol satisfying $\mathbb{E}(|s_t|^2) = 1$, $\boldsymbol{f}$ is the precoding vector, and $n_t \sim \mathbb{CN}(0, \sigma^2)$ denotes the

zero-mean complex Gaussian noise. From [15], we know that the two-timescale beamforming problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{\psi}} \quad & \mathbb{E}\{\max_{\boldsymbol{f}} R\} \\ s.t. \quad & |\psi_m| = 1 \quad \forall m \in \{1, \cdots, M\} \\ & \|\boldsymbol{f}\|^2 \leq P_{\text{max}} \\ & R = \log_2(1 + |\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G} \boldsymbol{f}|^2 / \sigma^2) \end{aligned} \tag{59}$$

where $\boldsymbol{\psi} = \text{diag}(\boldsymbol{\Psi})$ with ψ_m as its mth element and P_{max} is the total transmit power budget. The optimization problem in (59) has two levels. The inner one is the rate-maximization problem with respect to $\boldsymbol{f}$. This is implemented in each channel realization with the given phase shift matrix $\boldsymbol{\Psi}$. The outer one is an expectation maximization problem with respect to the IRS phase shift coefficients, in which the expectation of the achievable rate is taken over all possible channel realizations.

When $\boldsymbol{\Psi}$ is given, the optimal precoding vector is the maximum-ratio transmission (MRT). In this case, it can be readily verified that the above optimization problem is simplified into a problem concerning the optimization of the reflection coefficients $\boldsymbol{\Psi}$:

$$\begin{aligned} \max_{\boldsymbol{\psi}} \quad & \mathbb{E}\left\{\log_2(1 + P_{\text{max}} \|\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G}\|^2 / \sigma^2)\right\} \\ s.t. \quad & |\psi_m| = 1 \quad \forall m \in \{1, \cdots, M\} \end{aligned} \tag{60}$$

Directly solving (60) is intractable. To handle this issue, we resort to maximize its tight upper bound, which is given by [15]

$$\begin{aligned} &\mathbb{E}\{\log_2(1 + P_{\text{max}} \|\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G}\|^2 / \sigma^2)\} \\ &\qquad \leq \log_2(1 + P_{\text{max}} \mathbb{E}\{\|\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G}\|^2\} / \sigma^2) \end{aligned} \tag{61}$$

Note that the upper bound shown in (61) is sufficiently tight and thus is a good approximation of the original objective function, especially when P_{max}/σ^2 is large [14], [15]. Maximizing the upper bound yields the following optimization:

$$\begin{aligned} \max_{\boldsymbol{\psi}} \quad & \mathbb{E}(\|\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G}\|^2) \\ s.t. \quad & |\psi_m| = 1 \quad \forall m \in \{1, \cdots, M\} \end{aligned} \tag{62}$$

where $\mathbb{E}(\|\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G}\|^2)$ is given by

$$\begin{aligned} \mathbb{E}(\|\boldsymbol{h}^H \boldsymbol{\Psi} \boldsymbol{G}\|^2) &= \mathbb{E}(\|\boldsymbol{\psi}^T \boldsymbol{H}\|^2) \\ &= \boldsymbol{\psi}^T \mathbb{E}(\boldsymbol{H} \boldsymbol{H}^H) \boldsymbol{\psi}^* \\ &= \boldsymbol{\psi}^T \bar{\boldsymbol{R}}_h \boldsymbol{\psi}^* \end{aligned} \tag{63}$$

in which $\bar{\boldsymbol{R}}_h$ is the covariance matrix of the cascade channel $\boldsymbol{H}$. Note that $\bar{\boldsymbol{R}}_h = \mathbb{E}(\boldsymbol{H}\boldsymbol{H}^H)$ and $\boldsymbol{R}_h = \mathbb{E}(\text{vec}(\boldsymbol{H})\text{vec}(\boldsymbol{H})^H)$. Therefore, $\bar{\boldsymbol{R}}_h$ can be directly obtained from $\boldsymbol{R}_h$. Thus optimization of passive reflection coefficients can be done simply based on the estimated CCM of the cascade channel. Note that the optimization problem (62) with its cost function given in (63) is a nonconvex quadratically constrained quadratic problem, which can be solved by being relaxed as a semidefinite programming (SDP) problem and then using the Gaussian randomization approximation solution [30].

B. Multi-User Scenarios

Since downlink transmission is considered, the extension of the proposed CCM estimation method to multi-user scenarios is straightforward. Specifically, each user receives signals reflected from the IRS, and estimates its own downlink CCM via our proposed method. This information is fed back to the BS via a dedicated channel for subsequent two-timescale beamforming, as detailed next.

Suppose there are K single-antenna users and denote the channel between the IRS and the kth user by $\boldsymbol{h}_k$. The received signal at the kth user can be written as:

$$y_k = \boldsymbol{h}_k^H \boldsymbol{\Psi} \boldsymbol{G} \boldsymbol{f}_k s_k + \boldsymbol{h}_k^H \boldsymbol{\Psi} \boldsymbol{G} \sum_{j \neq k}^{K} \boldsymbol{f}_j s_j + n_k \tag{64}$$

where $\boldsymbol{f}_k$ ($\boldsymbol{f}_j$) and s_k (s_j)are, respectively, the transmit precoding vector and the symbol of the kth (jth) user, and $n_k \sim \mathbb{CN}(0, \sigma_q^2)$ denotes the zero-mean complex Gaussian noise. The two-timescale beamforming problem for multi-user scenarios can thus be formulated as [15]

$$\begin{aligned} \max_{\boldsymbol{\psi}} \quad & \mathbb{E}\left\{ \max_{\{\boldsymbol{f}_k\}} \quad \sum_{k=1}^{K} \log_2(1 + \text{SINR}_k) \right\} \\ s.t. \quad & |\psi_m| = 1 \quad \forall m \in \{1, \cdots, M\} \\ & \sum_{k=1}^{K} \|\boldsymbol{f}_k\|^2 \leq P_{\text{max}} \end{aligned} \tag{65}$$

where SINR_k is the signal-to-interference-noise ratio at the kth user, given by

$$\text{SINR}_k = \frac{|\boldsymbol{h}_k^H \boldsymbol{\Psi} \boldsymbol{G} \boldsymbol{f}_k|^2}{\sum_{j \neq k}^{K} |\boldsymbol{h}_k^H \boldsymbol{\Psi} \boldsymbol{G} \boldsymbol{f}_j|^2 + \sigma_k^2} \tag{66}$$

When $\boldsymbol{\Psi}$ is given, the optimization problem (65) becomes a traditional MIMO precoding problem that can be solved via many iterative algorithms, e.g., the generalized MMSE method. Nevertheless, for multi-user scenarios, $\{\boldsymbol{f}_k\}$ cannot be expressed analytically in terms of $\boldsymbol{\psi}$. This poses a challenge in obtaining an analytical expression of the ergodic rate as a function of the long term variable $\boldsymbol{\psi}$.

To address the above difficulty, the work [18] proposed to formulate the long-term passive beamforming problem into a problem maximizing the combined-effective-channel-gain (CECG). The CECG characterizes the strength of the effective channel and is defined as

$$\sum_{k=1}^{K} \mathbb{E}\left(\|\boldsymbol{h}_k^H \boldsymbol{\Psi} \boldsymbol{G}\|^2\right) = \boldsymbol{\psi}^T \sum_{k=1}^{K} \mathbb{E}\left(\boldsymbol{H}_k \boldsymbol{H}_k^H\right) \boldsymbol{\psi}^* \tag{67}$$

where $\boldsymbol{H}_k \triangleq \text{diag}(\boldsymbol{h}_k^H)\boldsymbol{G}$ is the cascade channel of the kth user. Define $\bar{\boldsymbol{R}}_h^k = \mathbb{E}(\boldsymbol{H}_k \boldsymbol{H}_k^H)$ and $\tilde{\boldsymbol{R}}_h \triangleq \sum_{k=1}^{K} \bar{\boldsymbol{R}}_h^k$. The long-term passive beamforming problem can be formulated into a form similar to (62):

$$\max_{\boldsymbol{\psi}} \quad \boldsymbol{\psi}^T \tilde{\boldsymbol{R}}_h \boldsymbol{\psi}^* \tag{68}$$

$$s.t. \quad |\psi_m| = 1 \quad \forall m \in \{1, \cdots, M\} \tag{69}$$

From our above discussion, we see that for multi-user scenarios, long-term passive beamforming can also be accomplished

with the knowledge of different users' CCMs $\{\bar{\boldsymbol{R}}_h^k\}$, which can be estimated via the proposed method.

VII. Simulation Results

In this section, we present simulation results to illustrate the effectiveness of the proposed low-rank PSD Toeplitz-structured CCM (LRT-CCM) estimation method. We compare our method with the conventional CCM estimation approach which estimates the channel at each time frame and then reconstructs the CCM with these estimated channel samples. Such an approach is referred to as the conventional CCM estimation method. For this approach, we use the compressed sensing-based method [6] to estimate the channel at each time frame.

In our simulations, the BS employs a uniform linear array (ULA) of $N = 8$ antennas and the IRS is a planar array with $M_v \times M_h = 16 \times 16$ reflecting elements. The three-dimensional coordinates of the BS, the IRS, and the user are set to $(5, 0, 10)$, $(0, 50, 20)$ and $(10, 60, 1.8)$, respectively. The BS-IRS channel and the IRS-user channel are generated according to (3) and (7), respectively. The number of the signal paths is set to $L = P = 3$, and each channel comprises an LOS path and two NLOS paths. The corresponding angles (including AoAs and AoDs) of the LOS paths are determined by the geometry configuration, and the associated complex gains of the LOS paths are generated according to a complex Gaussian distribution $\mathbb{CN}(0, 10^{-0.1\kappa})$, where $\kappa = 61.4 + 29.2\log_{10}(d) + \epsilon$ with d denoting the length of the path and ϵ being a random variable following $\mathbb{N}(0, (8.7\text{dB})^2)$. The angles associated with the NLOS paths are randomly selected from the interval $[-\pi, \pi]$. The channel coefficients of these NLOS paths follow a distribution $\mathbb{CN}(0, \delta^2)$, where δ is determined by the Rician factor (i.e., the ratio of the energy of the LOS path to that of all NLOS paths). We set the Rician factor to 10 dB and the transmitted power to $P_{\max} = 30$dBm. The signal-to-noise ratio (SNR) is defined as

$$\text{SNR} = \mathbb{E}\left(10\log_{10}\left(\varsigma^2/\sigma^2\right)\right) \tag{70}$$

where ς^2 is the received signal power.

We evaluate the CCM estimation performance via the relative efficiency metric (REM), which is widely adopted [31]–[33] and defined as

$$\eta = \frac{\text{tr}(\boldsymbol{U}_1^H \boldsymbol{R}_h \boldsymbol{U}_1)}{\text{tr}(\boldsymbol{U}_2^H \boldsymbol{R}_h \boldsymbol{U}_2)} \tag{71}$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are, respectively, the matrices constructed by the eigenvectors of the estimated CCM $\hat{\boldsymbol{R}}_h$ and the eigenvectors of the true CCM $\boldsymbol{R}_h$. Clearly, a higher value of η indicates a more accurate CCM estimate. The value $1 - \eta$ means the lost of the signal power due to the mismatch between the optimal beamformer and the estimated one. All results are averaged over 200 independent Monte Carlo runs.

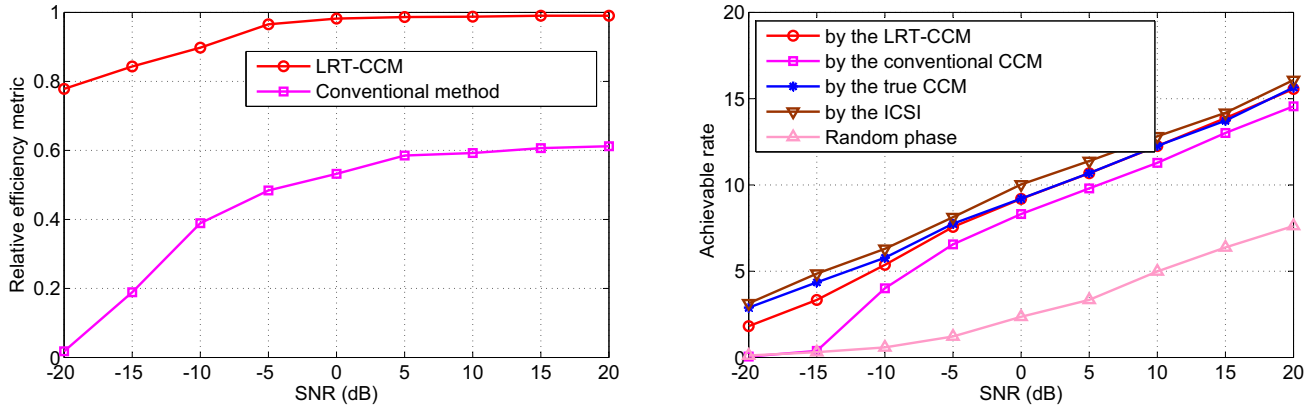
Fig. 2(a) plots the REMs of different methods as a function of the SNR, where the number of time slots J is set to 120 and the number of time frames T is set to 100. We see that our proposed method presents a substantial performance improvement over the conventional CCM estimation method, particularly in the low SNR regime. In fact, our proposed method can still provide a reliable CCM estimate even when the SNR is below -10dB, whereas the conventional CCM estimation method performs poorly in such a low SNR region. In Fig. 2(b), we plot the achievable rate attained by the two-timescale beamforming scheme based on the estimated CCM. To better evaluate the performance, we also include the achievable rates attained by the two-timescale beamforming scheme based on the true CCM, the two-timescale beamforming scheme in which the reflecting coefficients are randomly chosen from a unit circle (referred to as random passive beamforming), and the joint beamforming scheme [34] that utilizes the true I-CSI. Note that the beamforming approach [34] that exploits the I-CSI provides an upper bound on the performance that is achievable by any two-timescale beamforming schemes.

Several points can be made from Fig. 2(b). First, our proposed method incurs only a very mild performance loss as compared with the two-timescale beamforming scheme based on the true CCM. This result indicates that the proposed method can yield a CCM estimate that is good enough for subsequent beamforming. Second, the two-timescale beamforming scheme can achieve performance close to that of the beamforming method that utilizes the I-CSI, which demonstrates the effectiveness of the two-timescale beamforming scheme. Lastly, all methods present a substantial performance advantage over the random passive beamforming scheme.

Next, we examine the impact of the number of time frames on the estimation and beamforming performance. Fig. 3 plots the performance of respective methods as a function of T, where we set SNR to 0dB, and J is set to $J = 60$ and $J = 120$, respectively. It can be observed from Fig. 3 that a small value of T, say $T = 20$ is sufficient to achieve a decent performance for our proposed method. Increasing the number of time frames can lead to better performance for both methods, but the performance improvement is very limited. Since the total number of measurements required for training is TJ, this result suggests that our proposed method can provide a reliable CCM estimate using a training overhead as small as $TJ = 20 \times 60 = 1200$. Fig. 4 illustrates the effect of the number of time slots on the estimation and beamforming performance of respective methods, where we set SNR $= 0$dB, and T is set to $T = 40$ and $T = 100$, respectively. We see that increasing the number of time slots leads to better performance. Nevertheless, a small value of J, say, $J = 60$, is enough to provide a decent performance for our proposed method. Again, this result demonstrates the ability of the proposed method in providing an accurate CCM using a small amount of training overhead.

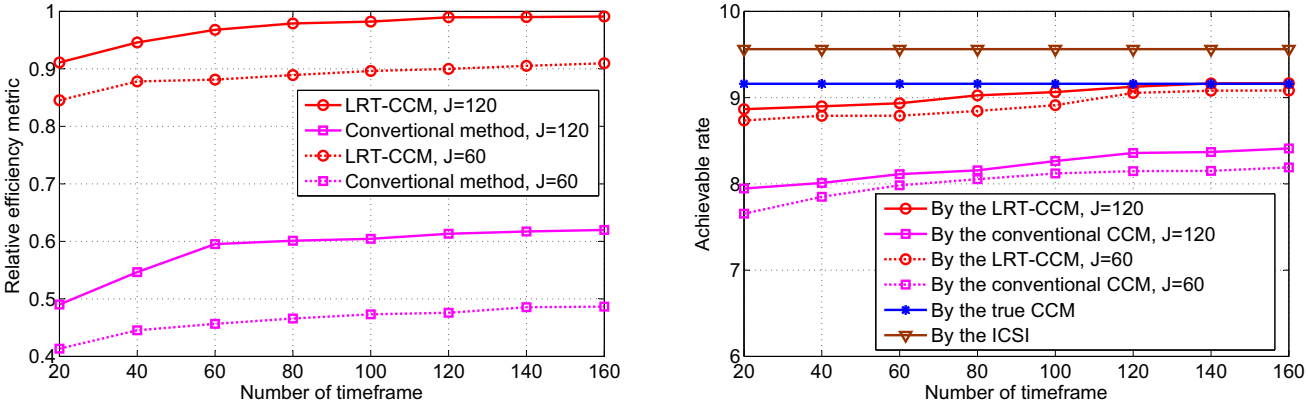
VIII. Conclusions

In this paper, we considered the CCM estimation for IRS-assisted mmWave communication systems. We exploited the low-rank property and PSD 3-level Toeplitz structure of the CCM and formulated the CCM estimation problem as a convex SDP problem, which was further efficiently solved by an ADMM algorithm. In addition, we analyzed the estimation performance of the proposed solution, as well as the training overhead required to obtain a reliable estimate of the



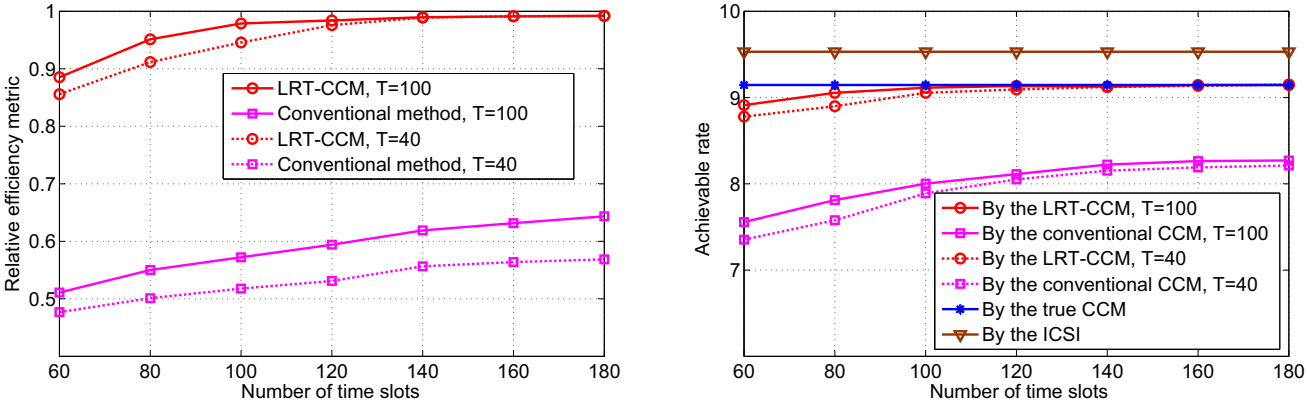
(a) REM (b) Achievable rate

Fig. 2: REMs and achievable rates of respective methods, where we set $T = 100$ and $J = 120$.



(a) REM (b) Achievable rate

Fig. 3: REMs and achievable rates of respective methods, where we set SNR $= 0$dB, $J = 60$ and $J = 120$.



(a) REM (b) Achievable rate

Fig. 4: REMs and achievable rates of respective methods, where we set SNR $= 0$dB, $T = 40$ and $T = 100$.

CCM. Lastly, we discussed how to perform the two-timescale beamforming based on the estimated CCM. Simulation results showed that our proposed method can provide a reliable CCM estimate using a small amount of training overhead.

APPENDIX I
PROOF OF THEOREM 1

Based on the definition, the trace of a PSD matrix is equivalent to its nuclear norm. For simplicity, we consider

the following equivalent form of (29):

$$\hat{\boldsymbol{R}}_h = \arg\min_{\boldsymbol{R}_h} \frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H\right\|_F^2 + \lambda\|\boldsymbol{R}_h\|_* \quad (72)$$

where $\|\cdot\|_*$ denotes the nuclear norm.

In order to prove Theorem 1, we first introduce the following theorem [29].

Theorem 2: For the convex optimization problem

$$\widehat{\boldsymbol{\Theta}}_{\lambda_n} \in \arg\min_{\boldsymbol{\Theta}} \{\mathcal{L}(\boldsymbol{\Theta}) + \lambda_n \mathcal{R}(\boldsymbol{\Theta})\} \quad (73)$$

where $\lambda_n > 0$ is a user-defined regularization parameter and $\mathcal{R}(\cdot)$ is a norm. Suppose that $\mathcal{L}$ is a convex and differentiable function, and consider any optimal solution $\widehat{\boldsymbol{\Theta}}$ to the aforementioned optimization problem with a strictly positive regularization parameter satisfying

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\Theta}^*)) \quad (74)$$

where $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$ and $\boldsymbol{\Theta}^*$ is the unknown true value. Denote $\mathcal{M} \subseteq \bar{\mathcal{M}}$ as the subspace to capture the constraints specified by the norm-based regularizer and $\bar{\mathcal{M}}^\perp$ as the orthogonal complement of space $\bar{\mathcal{M}}$. Then for any pair $\left(\mathcal{M}, \bar{\mathcal{M}}^\perp\right)$ over which $\mathcal{R}$ is decomposable[1], the error $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Theta}}_{\lambda_n} - \boldsymbol{\Theta}^*$ belongs to the set

$$\begin{aligned}&\mathbb{C}\left(\mathcal{M}, \bar{\mathcal{M}}^\perp; \boldsymbol{\Theta}^*\right)\\ &\quad\triangleq \{\boldsymbol{\Delta} | \mathcal{R}(\boldsymbol{\Delta}_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\boldsymbol{\Delta}_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\boldsymbol{\Theta}^*_{\mathcal{M}^\perp})\}\end{aligned} \quad (75)$$

where $\mathcal{M}^\perp$ is the orthogonal complement of the space $\mathcal{M}$. In (75) $\boldsymbol{\Delta}_{\bar{\mathcal{M}}}$ denotes the projection of $\boldsymbol{\Delta}$ onto the subspace $\bar{\mathcal{M}}$, which is defined as

$$\boldsymbol{\Delta}_{\bar{\mathcal{M}}} = \arg\min_{\boldsymbol{\Delta}_1 \in \bar{\mathcal{M}}} \|\boldsymbol{\Delta} - \boldsymbol{\Delta}_1\|_F^2 \quad (76)$$

$\boldsymbol{\Delta}_{\bar{\mathcal{M}}^\perp}$ and $\boldsymbol{\Theta}^*_{\mathcal{M}^\perp}$ can be similarly defined (and hence we omit them here). Specifically, $\mathcal{R}(\boldsymbol{\Theta}^*_{\mathcal{M}^\perp}) = 0$ when $\boldsymbol{\Theta} \in \mathcal{M}$, and under such a circumstance (75) turns to be

$$\mathbb{C}\left(\mathcal{M}, \bar{\mathcal{M}}^\perp; \boldsymbol{\Theta}^*\right) \triangleq \{\boldsymbol{\Delta} | \mathcal{R}(\boldsymbol{\Delta}_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\boldsymbol{\Delta}_{\bar{\mathcal{M}}})\} \quad (77)$$

Proof: See [29]. ■

Theorem 2 implies that with a proper regularization parameter and $\boldsymbol{\Theta} \in \mathcal{M}$, the estimation error of the problem (73) satisfies (77). We know that $\text{rank}(\boldsymbol{R}_h) = r$ and the SVD of $\boldsymbol{R}_h$ is given as $\boldsymbol{R}_h = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$. Let $\text{row}(\boldsymbol{R}_h)$ and $\text{col}(\boldsymbol{R}_h)$ denote the row and column space of $\boldsymbol{R}_h$ respectively, and meanwhile let $\boldsymbol{U}^r$ and $\boldsymbol{V}^r$ be the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. We now can define the subspace $\mathcal{M}$ and $\bar{\mathcal{M}}^\perp$ as

$$\mathcal{M} = \bar{\mathcal{M}} \triangleq \{\boldsymbol{R} | \text{row}(\boldsymbol{R}) = \boldsymbol{V}^r, \text{col}(\boldsymbol{R}) = \boldsymbol{U}^r\} \quad (78)$$

$$\bar{\mathcal{M}}^\perp \triangleq \{\boldsymbol{R} | \text{row}(\boldsymbol{R}) = (\boldsymbol{V}^r)^\perp, \text{col}(\boldsymbol{R}) = (\boldsymbol{U}^r)^\perp\} \quad (79)$$

Utilizing these defined subspaces, we can conclude that $\boldsymbol{R}_h \in \mathcal{M}$, and meanwhile the estimation error of $\boldsymbol{R}_h$, i.e., $\boldsymbol{\Delta} \triangleq \hat{\boldsymbol{R}}_h - \boldsymbol{R}_h$, can be decomposed into two parts, i.e.,

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2 \quad (80)$$

[1]A norm-based regularizer $\mathcal{R}$ is decomposable with respect to $\left(\mathcal{M}, \bar{\mathcal{M}}^\perp\right)$ if

$$\mathcal{R}(\boldsymbol{\Theta} + \boldsymbol{\Gamma}) = \mathcal{R}(\boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Gamma})$$

for all $\boldsymbol{\Theta} \in \mathcal{M}$ and $\boldsymbol{\Gamma} \in \bar{\mathcal{M}}^\perp$. Details can be found in [29].

with $\text{rank}(\boldsymbol{\Delta}_1) = r$, $\boldsymbol{\Delta}_1 \in \mathcal{M}$ and $\boldsymbol{\Delta}_2 \in \bar{\mathcal{M}}^\perp$.

In our problem, $\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{2}\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H\|_F^2$. Therefore, we have

$$\nabla\mathcal{L}(\boldsymbol{R}_h) = \frac{1}{2}\left[\boldsymbol{W}^H(\boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H - \hat{\boldsymbol{R}}_y)\boldsymbol{W}\right]^T \quad (81)$$

In addition, it is clear that the dual norm of the nuclear norm is the spectral norm. Therefore, if the regularization parameter λ satisfies

$$\begin{aligned}\lambda \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{R}_h)) &= \|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H)\boldsymbol{W}\|_2\\ &= \|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\boldsymbol{W}\|_2\end{aligned} \quad (82)$$

then the estimation error, based on Theorem 2, belongs to the set

$$\begin{aligned}\boldsymbol{\Delta} &\triangleq \hat{\boldsymbol{R}}_h - \boldsymbol{R}_h = \mathbb{T}_3(\hat{\boldsymbol{V}}) - \mathbb{T}_3(\boldsymbol{V}) = \mathbb{T}_3(\hat{\boldsymbol{V}} - \boldsymbol{V})\\ &\in \{\boldsymbol{\Delta} | \boldsymbol{\Delta} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2,\ \|\boldsymbol{\Delta}_2\|_* \leq 3\|\boldsymbol{\Delta}_1\|_*\}\end{aligned} \quad (83)$$

From (82), we can see that the choice of λ depends on the value of $\|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\boldsymbol{W}\|_2$. The following lemma provides an upper bound on $\|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\boldsymbol{W}\|_2$.

Lemma 1: Given T observation samples $\{\boldsymbol{y}_t\}_{t=1}^T$, $\hat{\boldsymbol{R}}_y$ is obtained via $\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_t\boldsymbol{y}_t^H$. Then with probability at least $1 - 4T^{-1}$, we have

$$\begin{aligned}&\|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\boldsymbol{W}\|_2\\ &\qquad\leq \delta \triangleq c\|\boldsymbol{W}\|_F^2\|\boldsymbol{R}_y\|_2 \max\{\sqrt{\tilde{\delta}}, \tilde{\delta}\}\end{aligned} \quad (84)$$

where $\tilde{\delta}$ is given in (53).

Proof: See Appendix II. ■

Since $\hat{\boldsymbol{R}}_h$ is the optimal solution to (72), we have

$$\begin{aligned}&\frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\hat{\boldsymbol{R}}_h\boldsymbol{W}^H\right\|_F^2 + \lambda\|\hat{\boldsymbol{R}}_h\|_*\\ &\qquad\leq \frac{1}{2}\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H\right\|_F^2 + \lambda\|\boldsymbol{R}_h\|_*\end{aligned} \quad (85)$$

which can be converted to

$$\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\hat{\boldsymbol{R}}_h\boldsymbol{W}^H\right\|_F^2 - \left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H\right\|_F^2 \leq 2\lambda\Big(\|\boldsymbol{R}_h\|_* - \|\hat{\boldsymbol{R}}_h\|_*\Big) \quad (86)$$

Recalling $\boldsymbol{\Delta} = \hat{\boldsymbol{R}}_h - \boldsymbol{R}_h$, the left side of (86) can be further expressed as

$$\begin{aligned}&\left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\hat{\boldsymbol{R}}_h\boldsymbol{W}^H\right\|_F^2 - \left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H\right\|_F^2\\ &= \left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}(\boldsymbol{R}_h + \boldsymbol{\Delta})\boldsymbol{W}^H\right\|_F^2 - \left\|\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H\right\|_F^2\\ &= 2\langle -\hat{\boldsymbol{R}}_y + \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H, \boldsymbol{W}\Delta\boldsymbol{W}^H\rangle + \left\|\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^H\right\|_F^2\\ &= 2\langle \boldsymbol{W}^H(-\hat{\boldsymbol{R}}_y + \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H)\boldsymbol{W}, \boldsymbol{\Delta}\rangle + \left\|\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^H\right\|_F^2\end{aligned} \quad (87)$$

Substituting (87) into (86), we have

$$\begin{aligned}
&\left\|\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^H\right\|_F^2 \\
&\leq 2\langle \boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H)\boldsymbol{W}, \boldsymbol{\Delta}\rangle + 2\lambda\Big(\|\boldsymbol{R}_h\|_* - \|\hat{\boldsymbol{R}}_h\|_*\Big) \\
&\overset{(a)}{\leq} 2\|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H)\boldsymbol{W}\|\|\boldsymbol{\Delta}\|_* + 2\lambda\Big(\|\hat{\boldsymbol{R}}_h + \boldsymbol{\Delta}\|_* - \|\hat{\boldsymbol{R}}_h\|_*\Big) \\
&\overset{(b)}{\leq} 2\|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{W}\boldsymbol{R}_h\boldsymbol{W}^H)\boldsymbol{W}\|\|\boldsymbol{\Delta}\|_* + 2\lambda\|\boldsymbol{\Delta}\|_* \\
&\overset{(c)}{\leq} 4\lambda\left\|\boldsymbol{\Delta}\right\|_* \qquad (88)
\end{aligned}$$

where (a) follows from Holder's inequality, (b) comes from the triangle inequality, and (c) follows from the assumption $\lambda \geq \delta$. Furthermore, we have

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_* &= \|\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2\|_* \leq \|\boldsymbol{\Delta}_1\|_* + \|\boldsymbol{\Delta}_2\|_* \\
&\overset{(a)}{\leq} \|\boldsymbol{\Delta}_1\|_* + 3\left\|\boldsymbol{\Delta}_1\right\|_* \\
&\overset{(b)}{\leq} 4\sqrt{r}\|\boldsymbol{\Delta}_1\|_F \\
&\leq 4\sqrt{r}\left\|\boldsymbol{\Delta}\right\|_F = 4\sqrt{r}\left\|\mathbb{T}_3(\hat{\boldsymbol{V}} - \boldsymbol{V})\right\|_F \\
&\leq 4\sqrt{rNM}\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F \qquad (89)
\end{aligned}$$

where (a) follows from (83), and (b) is a result of the relationship between the nuclear norm and F-norm of a rank-r matrix. Putting (88) and (89) together results in

$$\left\|\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^H\right\|_F^2 \leq 16\lambda\sqrt{rNM}\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F \qquad (90)$$

Furthermore, we utilize the following lemma to find a lower bound of $\left\|\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^H\right\|_F^2$.

Lemma 2: Consider a 3-level Toeplitz matrix $\mathbb{T}_3(\boldsymbol{X})$ with $\boldsymbol{X} \in \mathbb{C}^{(2I_1-1)\times(2I_2-1)\times(2I_3-1)}$ and the matrix $\boldsymbol{W} \in \mathbb{C}^{I^2\times I_1 I_2 I_3}$. If

$$I \geq \sqrt{(2I_1-1)(2I_2-1)(2I_3-1)} \qquad (91)$$

then with high probability there exists a full-column rank transforming matrix $\check{\boldsymbol{W}}$ of $\boldsymbol{W}$ such that

$$\|\boldsymbol{W}\mathbb{T}_3(\boldsymbol{X})\boldsymbol{W}^H\|_F^2 \geq \sigma_{\min}^2(\check{\boldsymbol{W}})\|\boldsymbol{X}\|_F^2 \qquad (92)$$

where $\sigma_{\min}(\check{\boldsymbol{W}})$ is the smallest singular value of $\check{\boldsymbol{W}}$.

Proof: See Appendix III. ∎

Since we have

$$J \geq u \triangleq \sqrt{(2N-1)(2M_v-1)(2M_h-1)} \qquad (93)$$

and meanwhile $\boldsymbol{\Delta} = \mathbb{T}_3(\hat{\boldsymbol{V}} - \boldsymbol{V})$ is a 3-level Toeplitz matrix, based on Lemma 2, we know that, with high probability the following holds

$$\|\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^H\|_F^2 \geq \sigma_{\min}^2(\check{\boldsymbol{W}})\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F^2 \qquad (94)$$

where $\check{\boldsymbol{W}}$ is the transforming matrix of $\boldsymbol{W}$. Substituting (94) into (90) leads to

$$\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F \leq \frac{16\lambda\sqrt{rNM}}{\sigma_{\min}^2(\check{\boldsymbol{W}})} \qquad (95)$$

As a result, the average per-entry RMSE is given by

$$\frac{1}{u}\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_F \leq \frac{16\lambda\sqrt{r}}{\sigma_{\min}^2(\check{\boldsymbol{W}})}\frac{\sqrt{NM}}{u} \qquad (96)$$

which completes the proof.

Appendix II
Proof of Lemma 1

The received signal $\boldsymbol{y}_t$ follows a complex Gaussian $\mathbb{CN}(0, \boldsymbol{R}_y)$. Therefore, according to Theorem 2.2 in [35] we know that, with probability at least $1 - 4T^{-1}$, the following inequality holds

$$\|\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y\|_2 \leq c\|\boldsymbol{R}_y\|_2\max\{\sqrt{\tilde{\delta}}, \tilde{\delta}\} \qquad (97)$$

where c is a constant. Furthermore, we have

$$\begin{aligned}
&\|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\boldsymbol{W}\|_2 \\
&\overset{(a)}{\leq} \|\boldsymbol{W}^H(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\boldsymbol{W}\|_F \\
&\overset{(b)}{\leq} \|\boldsymbol{W}\|_F^2\|(\hat{\boldsymbol{R}}_y - \boldsymbol{R}_y)\|_2 \\
&\leq c\|\boldsymbol{W}\|_F^2\|\boldsymbol{R}_y\|_2\max\{\sqrt{\tilde{\delta}}, \tilde{\delta}\} \qquad (98)
\end{aligned}$$

where (a) follows from the fact that $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$ and (b) follows from $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_F$.

Appendix III
Proof of Lemma 2

According to the Kronecker product property, we have

$$\text{vec}(\boldsymbol{W}\mathbb{T}_3(\boldsymbol{X})\boldsymbol{W}^H) = (\boldsymbol{W}^* \otimes \boldsymbol{W})\text{vec}(\mathbb{T}_3(\boldsymbol{X})) \qquad (99)$$

Since $\mathbb{T}_3(\boldsymbol{X})$ is a 3-level Toeplitz matrix, we can express (99) as

$$(\boldsymbol{W}^* \otimes \boldsymbol{W})\text{vec}(\mathbb{T}_3(\boldsymbol{X})) = \check{\boldsymbol{W}}\boldsymbol{x} \qquad (100)$$

where $\boldsymbol{x} = \text{vec}(\boldsymbol{X})$ and $\check{\boldsymbol{W}} \in \mathbb{C}^{I^2\times(2I_1-1)(2I_2-1)(2I_3-1)}$ is constructed by combining those columns in $\boldsymbol{W}^* \otimes \boldsymbol{W}$ which correspond to the same element in $\mathbb{T}_3(\boldsymbol{X})$. Furthermore, $\check{\boldsymbol{W}}$ is a full-column rank matrix with high probability due to the fact that $\boldsymbol{W}$ is a random matrix and the condition (91) holds. Since $\check{\boldsymbol{W}}$ is a full-column rank matrix, we have

$$\|\check{\boldsymbol{W}}\boldsymbol{x}\|_2^2 \geq \sigma_{\min}^2(\check{\boldsymbol{W}})\|\boldsymbol{x}\|_2^2 \equiv \sigma_{\min}^2(\check{\boldsymbol{W}})\|\boldsymbol{X}\|_F^2 \qquad (101)$$

Due to that fact that $\|\boldsymbol{A}\|_F^2 = \|\text{vec}(\boldsymbol{A})\|_2^2$ for an arbitrary matrix $\boldsymbol{A}$, we have

$$\|\boldsymbol{W}\mathbb{T}_3(\boldsymbol{X})\boldsymbol{W}^H\|_F^2 = \|\check{\boldsymbol{W}}\boldsymbol{x}\|_2^2 \geq \sigma_{\min}^2(\check{\boldsymbol{W}})\|\boldsymbol{X}\|_F^2 \qquad (102)$$

which completes the proof.

References

[1] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.

[2] 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 Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 960–14 973, 2020.

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

[4] Y. Cao, T. Lv, W. Ni, and Z. Lin, "Sum-rate maximization for multi-reconfigurable intelligent surface-assisted device-to-device communications," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7283–7296, 2021.

[5] 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," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1035–1071, 2022.

[6] P. Wang, J. Fang, H. Duan, and H. Li, "Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems," *IEEE Signal Processing Letters*, vol. 27, pp. 905–909, 2020.

[7] S. Liu, Z. Gao, J. Zhang, M. Di Renzo, and M.-S. Alouini, "Deep denoising neural network assisted compressive channel estimation for mmwave intelligent reflecting surfaces," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9223–9228, 2020.

[8] 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 Communications Letters*, vol. 25, no. 5, pp. 1403–1407, 2021.

[9] Y. Lin, S. Jin, M. Matthaiou, and X. You, "Channel estimation and user localization for irs-assisted mimo-ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 4, pp. 2320–2335, 2021.

[10] G. T. de Araújo, A. L. De Almeida, and R. Boyer, "Channel estimation for intelligent reflecting surface assisted MIMO systems: A tensor modeling approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 789–802, 2021.

[11] X. Zheng, P. Wang, J. Fang, and H. Li, "Compressed channel estimation for IRS-assisted millimeter wave ofdm systems: A low-rank tensor decomposition-based approach," *IEEE Wireless Communications Letters*, vol. 11, no. 6, pp. 1258–1262, 2022.

[12] F. Yang, J.-B. Wang, H. Zhang, C. Chang, and J. Cheng, "Intelligent reflecting surface-assisted mmwave communication exploiting statistical CSI," in *2020 IEEE International Conference on Communications (ICC)*. Dublin, Ireland, 2020, pp. 1–6.

[13] X. Hu, F. Gao, C. Zhong, X. Chen, Y. Zhang, and Z. Zhang, "An angle domain design framework for intelligent reflecting surface systems," in *2020 IEEE Global Communications Conference (GLOBECOM)*. Taipei, Taiwan, 2020, pp. 1–6.

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

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

[16] X. Gan, C. Zhong, C. Huang, and Z. Zhang, "RIS-assisted multi-user MISO communications exploiting statistical CSI," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6781–6792, 2021.

[17] Y. Jia, Y. Cui, and W. Jiang, "Robust optimization of instantaneous beamforming and quasi-static phase shifts in an IRS-assisted multi-cell network," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4394–4409, 2021.

[18] P. Wang, J. Fang, Z. Wu, and H. Li, "Two-timescale beamforming for IRS-assisted millimeter wave systems: A deep unrolling-based stochastic optimization approach," in *2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. Trondheim,Norway, 2022, pp. 191–195.

[19] Y. Cao, T. Lv, and W. Ni, "Two-timescale optimization for intelligent reflecting surface-assisted MIMO transmission in fast-changing channels," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 424–10 437, 2022.

[20] S. Park and R. W. Heath, "Spatial channel covariance estimation for the hybrid MIMO architecture: A compressive sensing-based approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8047–8062, 2018.

[21] S. Park, A. Ali, N. González-Prelcic, and R. W. Heath, "Spatial channel covariance estimation for hybrid architectures based on tensor decompositions," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1084–1097, 2019.

[22] Y. Wang, Y. Zhang, Z. Tian, G. Leus, and G. Zhang, "Super-resolution channel estimation for arbitrary arrays in hybrid millimeter-wave massive MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 947–960, 2019.

[23] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1257–1269, 2015.

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

[25] I. Viering, H. Hofstetter, and W. Utschick, "Spatial long-term variations in urban, rural and indoor environments," in *Proceedings of the 5th Metting COST*. Lisbon, Pirtugal, 2002.

[26] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[27] Z. Yang, L. Xie, and P. Stoica, "Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3685–3701, 2016.

[28] Z. Zhang, Y. Wang, and Z. Tian, "Efficient two-dimensional line spectrum estimation based on decoupled atomic norm minimization," *Signal Processing*, vol. 163, pp. 95–106, 2019.

[29] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[30] A. M.-C. So, J. Zhang, and Y. Ye, "On approximating complex quadratic optimization problems via semidefinite programming relaxations," *Mathematical Programming*, vol. 110, no. 1, pp. 93–110, 2007.

[31] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 303–318, 2016.

[32] S. Park, J. Park, A. Yazdan, and R. W. Heath, "Exploiting spatial channel covariance for hybrid precoding in massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 14, pp. 3818–3832, 2017.

[33] C. K. Anjinappa, A. C. Gürbüz, Y. Yapıcı, and I. Güvenç, "Off-grid aware channel and covariance estimation in mmwave networks," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3908–3921, 2020.

[34] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*. Changchun, China, 2019, pp. 735–740.

[35] F. Bunea and L. Xiao, "On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA," *Bernoulli*, vol. 21, no. 2, pp. 1200–1230, 2015.

Hongwei Wang received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2013 and 2019, respectively. He is currently a Post-Doctoral researcher with the University of Electronic Science and Technology of China. His research interests include statistical signal processing, compressed sensing and sparse theory, millimeter wave communications, and wideband spectrum sensing.

Jun Fang (SM'19) received the B.S. and M.S. degrees from the Xidian University, Xi'an, China in 1998 and 2001, respectively, and the Ph.D. degree from the National University of Singapore, Singapore, in 2006, all in electrical engineering.

During 2006, he was a postdoctoral research associate in the Department of Electrical and Computer Engineering, Duke University. From January 2007 to December 2010, he was a research associate with the Department of Electrical and Computer Engineering, Stevens Institute of Technology. Since 2011, he has been with the University of Electronic of Science and Technology of China, where he is a Professor. His research interests include compressed sensing and sparse theory, massive MIMO/mmWave communications, and statistical inference.

Dr. Fang was the recipient of the IEEE Jack Neubauer Memorial Award in 2013 for the best systems paper published in the *IEEE Transactions on Vehicular Technology*. He served as an Associate Technical Editor for *IEEE Communications Magazine* from 2012 to 2020. He is currently a Member of the IEEE SPS Sensor Array and Multichannel TC, and a Senior Associate Editor for *IEEE Signal Processing Letters*.

PLACE PHOTO HERE

Huiping Duan received her Ph.D. in Electrical Engineering from Nanyang Technological University (NTU) in 2008, and the M.S. and B.S. degrees from Xidian University in 2001 and 1998, respectively. Since 2011, she has been with the University of Electronic Science and Technology of China, where she is currently an Associate Professor. Her research interests include compressed sensing and sparse theory, array signal processing, and statistical inference.

Hongbin Li (M'99-SM'08-F'19) received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, in 1991 and 1994, respectively, and the Ph.D. degree from the University of Florida, Gainesville, FL, in 1999, all in electrical engineering.

From July 1996 to May 1999, he was a Research Assistant in the Department of Electrical and Computer Engineering at the University of Florida. Since July 1999, he has been with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, where he is currently the Charles and Rosanna Batchelor Memorial Chair Professor. He was a Summer Visiting Faculty Member at the Air Force Research Laboratory in the summers of 2003, 2004 and 2009. His general research interests include statistical signal processing, wireless communications, and radars.

Dr. Li received the IEEE Jack Neubauer Memorial Award in 2013 from the IEEE Vehicular Technology Society, Outstanding Paper Award from the IEEE AFICON Conference in 2011, Provost's Award for Research Excellence in 2019, Harvey N. Davis Teaching Award in 2003, and Jess H. Davis Memorial Research Award in 2001 from Stevens Institute of Technology, and Sigma Xi Graduate Research Award from the University of Florida in 1999. He has been a member of the IEEE SPS Signal Processing Theory and Methods Technical Committee (TC) and the IEEE SPS Sensor Array and Multichannel TC, an Associate Editor for *Signal Processing* (Elsevier), *IEEE Transactions on Signal Processing*, *IEEE Signal Processing Letters*, and *IEEE Transactions on Wireless Communications*, as well as a Guest Editor for *IEEE Journal of Selected Topics in Signal Processing* and *EURASIP Journal on Applied Signal Processing*. He has been involved in various conference organization activities, including serving as a General Co-Chair for the 7th IEEE Sensor Array and Multichannel Signal Processing (SAM) Workshop, Hoboken, NJ, June 17-20, 2012. Dr. Li is a member of Tau Beta Pi and Phi Kappa Phi.